Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____

Date

Exploring Early Lupus Subtypes

By

Rebecca Ann Speckman
Doctor of Philosophy

Epidemiology

William M. McClellan
Advisor

Roberd Bostick
Committee Member

Cristina Drenkard
Committee Member

Mitchel Klein
Committee Member

S. Sam Lim
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Exploring early lupus subtypes

By

Rebecca A. Speckman
B.A., Washington University, 1999

Advisor: William M. McClellan, MD, MPH

An abstract of
a dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology.

2010

Abstract

Exploring early lupus subtypes
By Rebecca A. Speckman

Systemic lupus erythematosus (SLE), or lupus, is an autoimmune disease characterized by the presence of autoantibodies to elements of the cell nucleus. SLE can cause permanent damage to any organ system. Diagnosis is made by the recognition of a "constellation of symptoms and signs," as there is no gold standard test for SLE. Patients diagnosed with lupus by a physician or classified as having lupus by the accepted research classification criteria have wide inter-patient variety in organ system involvement; many lupus patients exhibit only a subset of the possible clinical manifestations. It is possible that several disease subtypes, or alternatively several diseases with overlapping manifestations, are encompassed by what is currently called lupus.

In my first study, I found that K-modes cluster analysis did not outperform K-means in the presence/absence data examined. However, because the K-modes and K-means suggested different partitions in a real data set, it may be useful to consider the two methods as complementary in the setting of subtype exploration (class discovery). In my second study, I found that recommended indices for choosing the best presence/absence data partition were not reliable. In my third study, I used K-means, K-modes, and latent class analysis to create clusters of early lupus patients. A set of four clusters created by K-means was selected as candidate early lupus subtypes.

Exploring early lupus subtypes

By

Rebecca A. Speckman
B.A., Washington University, 1999

Advisor: William M. McClellan, MD, MPH

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology.

2010

Contents

List of Tables

List of Figures

CHAPTER 1

Introduction and context

Systemic lupus erythematosus (SLE), an autoimmune rheumatic disease characterized by antibodies to components of the cell nucleus, has many possible manifestations and can cause permanent damage to any organ. Although some of the pathogenesis of SLE has been described, the etiology of SLE is considered to be largely unknown. SLE is defined in the clinical setting by the recognition of a "constellation of symptoms and signs," and in the research setting by the American College of Rheumatology Research Classification Criteria for SLE (Tan *et al.*, 1982; Hochberg, 1997). By either of these definitions, there is wide variation in manifestations between SLE patients. Furthermore, some patients diagnosed with "definite" lupus by specialists familiar with the disease (rheumatologists, dermatologists, or nephrologists) do not meet the research classification criteria (Petri and Magder, 2004).

The wide inter-patient variation in clinical and laboratory manifestations of SLE suggests that there may be several SLE phenotypes, or subgroups of patients with homogenous manifestations. SLE phenotypes could be conceptualized as the overt manifestations of disease subtypes, with a common underlying etiology and additional etiologic factors leading to different disease subtypes. Alternatively, there may be several diseases with different underlying etiologies that have some overlapping clinical and laboratory manifestations. These possibilities are not mutually exclusive.

Formal class discovery methods such as cluster analysis can identify sub-groups of patients with shared features. The hypothesis that such phenotypes have distinct etiologies is analogous to the epidemiologic practice of preliminary case definition for patients with a previously undescribed disease. Typically, in the

investigation of a "new" disease, a collection of features shared by the group of sick patients is used as a preliminary case definition. The preliminary case definition is used to investigate the etiology and outcomes of the disease, and the case definition is revised as information is gained about the disease process or clinical outcomes.

Exploration of disease phenotypes in order to elucidate disease etiology has been advocated as a strategy specifically for SLE, and investigation of subsets has long been a staple of the SLE research literature (Edworthy *et al.*, 1988). However, few studies have used formal class discovery methods such as cluster analysis to explore subsets of patients (To and Petri, 2005; Jurencak *et al.*, 2009).

1.1 Dissertation overview

The goal of my dissertation research was to explore early lupus subtypes.

Formal cluster analysis investigation. The steps in a formal investigation of disease subtypes using cluster analysis are depicted in Figure 1.1, and Figure 1.2 shows how the studies of my dissertation fit into this scheme.

If the purpose of a study is to investigate heretofore undescribed subtypes, the number of clusters of the best partition for a given cluster analysis method will create is unknown. Therefore, when using cluster analysis methods that require a pre-specified number of clusters to create homogeneous sub-groups, the first step in the analysis of real data is to create partitions for a range of numbers of clusters. Next, the "best" partition is selected from the range of partitions created, a process often called "choosing the correct number of clusters." The clusters of this partition serve as candidate clusters, or candidate subtypes. Last, the meaningfulness of the candidate clusters is assessed by examining associations with patient characteristics, etiologic risk factors, and clinical outcomes.

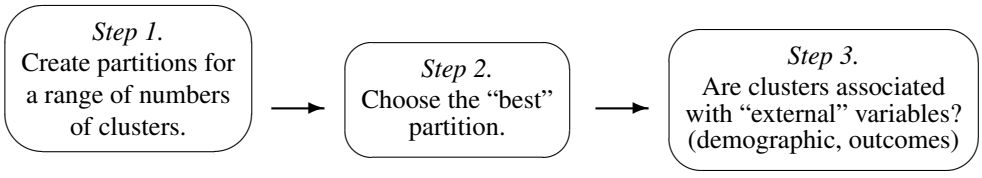

Figure 1.1: Steps in a cluster analysis investigation of subtypes.

Motivation. The primary motivation of my dissertation was that patients diagnosed with SLE or classified as having SLE by the research classification criteria are heterogeneous, which may be hampering research progress on the etiology and outcomes of SLE. Delineation of SLE phenotypes that hypothetically arise from different etiologies has been suggested as a strategy to clarify the classification of SLE, but formal (objective) subtype discovery methods have rarely been employed.

In initial surveys of subtype discovery methods that could be used to create candidate early lupus subtypes, *K*-means cluster analysis appeared to be well-suited for the task. However, several variations of *K*-means that seemed particularly applicable to the lupus setting had not been adequately investigated.

Studies overview. In Studies 1 and 2, I compared the performance of cluster analysis methods that could be used for Steps 1 and 2 of Figure 1.1, where performance was defined as the ability of these methods to create clusters that agreed with known subtypes. As the goal purpose of the dissertation was to identify previously undescribed early lupus subtypes, data from early lupus patients did not have "known" subtypes and thus could not be used for Studies 1 and 2. In Study 3, I used selected cluster analysis methods to create a set of clusters from the Georgia Lupus Registry to serve as candidate early lupus subtypes. Then, I measured association of the candidate subtypes with demographic variables and clinical outcomes.

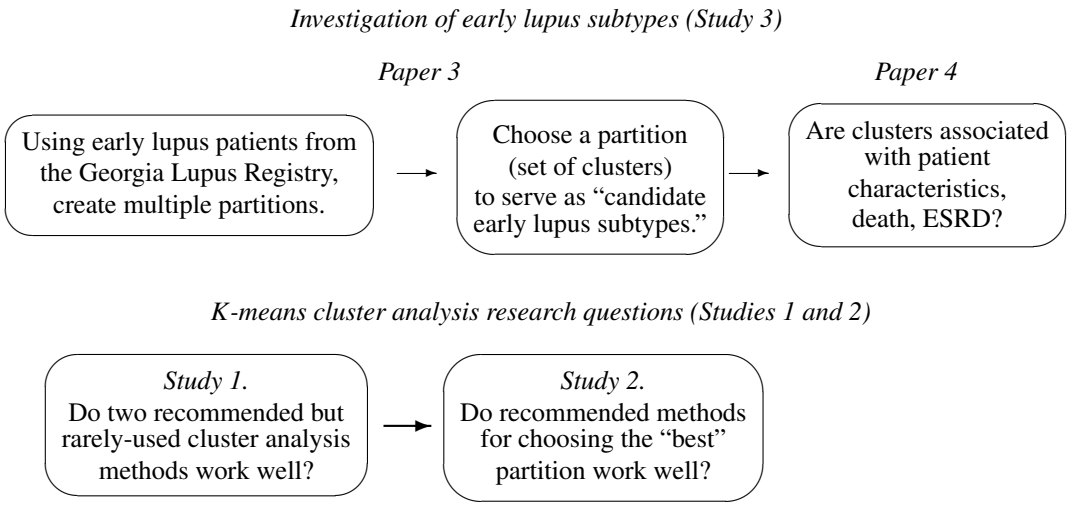

Figure 1.2: Overview of dissertation studies.

Aim 1: Evaluate the performance of several cluster analysis methods that can be used for presence/absence data.

I compared the ability of several cluster analysis methods to recover cluster structure from simulated datasets and a "benchmark" real data set, which will have known cluster structure. The methods evaluated were three types of *K*-centroids cluster analysis: *K*-means, *K*-modes with the simple matching distance, and *K*-modes with the Jaccard distance. *K*-centroids algorithms partition the objects (cases, or patients) in a dataset into a pre-specified (*K*) number of clusters in such a way as to minimize the sum of distances from objects to cluster "centers." *K*-modes has been suggested to be more appropriate for presence-absence data, such as the presence and absence of disease manifestations, but this suggestion has not been evaluated.

Evaluation of cluster analysis methods. The standard way in which a cluster analysis method is evaluated is to use the method to "blindly" create clusters in data with known subtypes, ignoring knowledge of the known subtypes. Then, the clusters of patients created by the cluster analysis method are compared to the known subtypes.

Data. In this aim, I used simulated and real data. The real data was a benchmark data set with information on disease manifestations from patients with overlapping erythematosquamous diseases. I created two types of simulated data. First, I created a large number of data sets using a factorial design, in which data set design factors were completely crossed. Second, I created a small number of data sets with characteristics very similar to the real, benchmark data.

Study 2: Evaluate the performance of relative validity indices for choosing the best partition.

Because *K*-centroids cluster analysis methods create partitions with a user-specified number of clusters, it is necessary to choose a partition from a range of partitions created by the clustering algorithm. Only one previous study had evaluated the performance of relative validity indices in binary (presence/absence) data.

Data. In this study, I used the benchmark data used in Study 1, some simulated data used in Study 1, and additional simulated data. The additional simulated data was created in order to add a new data design factor, number of underlying subtypes.

Study 3: Identify putative subtypes in patients from the Georgia Lupus Registry

In Study 3, which is Papers 3 and 4, I identified candidate early lupus subtypes among patients with early lupus from the Georgia Lupus Registry (GLR).

Study population. The GLR collects clinical and laboratory data from the medical records of patients with possible lupus presenting before January 1, 2005 in Fulton and Dekalb counties. "Early lupus" was defined as patients with first diagnosis of "possible lupus" from January 1, 2002 to December 31, 2004, meeting one of the following case definitions of SLE: (1) four or more American College of Rheumatology (ACR) criteria for SLE classification (Tan *et al.*, 1982; Hochberg, 1997), or (2) three ACR criteria and diagnosis of SLE by a rheumatologist.

Several of the methods evaluated in Studies 1 and 2 were used to create partitions and to identify the best partition(s) of the data, and a single set of clusters was chosen as candidate early lupus subtypes. Association of the candidate subtypes with patient demographics and clinical outcomes (mortality and end-stage renal disease) was measured.

1.2 Subtyping as case definition

Numerous diseases in fields such as rheumatology, neurology, and psychiatry are diagnosed by the recognition of a constellation of symptoms and signs, or a subset of possible manifestations. The epidemiologist wishing to study such diseases is faced with an obstacle early in the process of investigation: what is an appropriate case definition for the disease? Too restrictive of a case definition may lead to the omission of valuable information. Too broad of a case definition may weaken the ability to identify risk factors or treatment effects that are relevant to a subset of patients with the disease (Lasky and Stolley, 1994).

Often, there are many possible underlying pathophysiologic processes for a syndromic disease but no clear etiology. In these situations, restricting or subsetting case definitions to create homogenous groups might yield advances in our understanding of pathophysiology of such diseases.

1.2.1 Case definition

In order to gain understanding about the causes and outcomes of a disease, epidemiologic investigations are conducted. Definition of what constitutes a case is one of the first steps in an epidemiologic investigation. The purpose of case definition is to ensure that cases represent the same entity (Torrence, 1997).

In the setting of an "undefined" or "new" disease, the preliminary case definition is usually a list of the symptoms and signs shared by the patients, or a clinical syndrome. In other words, it might be said that case definition in the new disease setting is the description of a group of people who have homogenous disease characteristics. As new information about the disease is discovered, for example a pathogen that is associated with the collection of symptoms and signs, the case definition and other disease definitions (e.g. diagnostic methods) are modified accordingly.

AIDS is a classic example of an evolving case definition (Lasky and Stolley, 1994; Tyler and Last, 1998). The preliminary case definition published by the Centers for Disease Control was a list of disease manifestations: lymphopenia, opportunistic infections with no apparent reason for immunosuppression, and others (CDC, 1982). The case definition was modified as researchers learned about the biology of the disease, most notably to include infection by the HIV virus (CDC, 1985).

1.2.2 "Lumping and splitting"

J. N. Morris considered the identification of syndromes, including the delineation of multiple syndromes within a larger clinical phenomenon, to be one of the core uses of epidemiology (Morris, 1955). Historic examples of the splitting of a clinical syndrome include the division of hepatitis A and hepatitis B or the recognition of different inflammatory arthritidies. Sometimes, a myriad of manifestations are "lumped" when it is discovered that they result from the same disease, for example tuberculosis (Tyler and Last, 1998).

In diseases with a wide variety of symptoms and signs, restriction of investigation to a more homogeneous subset of patients is analogous to case definition in the "new" disease setting; it is hoped that the restriction might allow researchers to uncover etiologic or outcome information pertaining specifically to that subgroup. Furthermore, it is possible that patients encompassed by the current definition of a heterogeneous disease actually include several smaller, homogenous groups of patients; in other words, there may be subtypes of the disease. These subtypes may differ in etiology or response to treatment. Developing case definitions for disease subtypes could be beneficial to epidemiologic research and also may aid in earlier clinical recognition of disease (Tyler and Last, 1998).

Examples. Experts in many heterogeneous diseases have suggested splitting, or creating subtypes, as an investigative approach that might help elucidate etiology.

The autism spectrum disorders (ASD) encompass patients with a wide variety of symptoms, and are diagnosed solely by the recognition of constellations of behavioral findings. The triad of behavioral features necessary for diagnosis of core autism is social impairment, communication impairment, and repetitive behaviors, all occurring before three years of age. Heterogeneity is also found in the other neurodevelopmental and clinical features associated with autism. Although there has been much research on the neuropathology and etiology of the ASDs, the etiopathology of these disorders remains largely a mystery. Happe *et al.* (2006) suggest that the lack of progress in uncovering the etiology of autism is the result of the assumption that the triad of impairments must be explained together (Happe *et al.*, 2006). Geschwind and Levitt use the term 'autisms' to refer to diseases in the autism spectrum to emphasize that "there are distinct clinical entities whose phenotypic overlap and etiologies remains to be defined" (Geschwind and Levitt, 2007). As an example of splitting used to explore etiology, a recent linkage scan used stratified cases by phenotype to identify genetic regions predisposing to subsets of the ASD's (Schellenberg *et al.*, 2006).

The role of epidemiologists. While the tools typically used in class discovery are not "traditional" epidemiologic methods, class discovery for the purposes of better case definition may in some cases be an important step in epidemiologic investigation. Tyler and Last (1998) write that "epidemiologists use a wide range of scientific information... In the end, it is the reasoning of the epidemiologist that ties these facts together." Perhaps it is the role of the epidemiologist to think critically and work with clinicians and statisticians to choose an appropriate case definition of a heterogeneous disease for a given study, or even to conduct a study with the primary goal of refining a case definition (Tyler and Last, 1998).

1.3 SLE case definition

1.3.1 Current concepts of SLE

According to the Arthritis Foundation of Americas Primer on the Rheumatic Diseases, SLE is "a prototypic autoimmune disease with a diverse array of clinical manifestations, which is characterized by the production of antibodies to components of the cell nucleus" (Klippel, 2001). SLE fits into several categories of disease: it is a connective tissue disease (CTD), an autoimmune disease (AID), and it falls into the rheumatic disease family. It is also sometimes called a diffuse connective tissue disease (DCTD) and an autoimmune rheumatic disease (ARD). Most SLE occurs sporadically, but some cases have a family history of SLE or other autoimmune diseases.

1.3.2 History of SLE disease concept

Hochberg (1993) breaks the history of lupus erythematosus (LE) as a disease concept into three periods: the classical period, in which the cutaneous manifestations were described, the neoclassical period, during which systemic manifestations were described, and the modern period, during which key scientific advances such as the discovery of the LE cell were made.

The word "lupus," the Latin for "wolf," was first used in the 13th century to describe ulcerative skin lesions with a "wolf's bite" appearance (Lahita, 1987). In 1933, Biett described an erythematous facial rash that would today be called lupus erythematosus, using the term *érythème centrige*. Biett's pupil Casenave used the term lupus erythematosus in 1851 (in the French form, *lupus érythémateux*) to refer to the same disease, and the term was in common use by dermatologists by 1870 (Holubar, 1980; Lahita, 1987).

Kaposi, von Hebras son-in-law, was the first to describe systemic manifestations of lupus erthythematosus (Lahita, 1987), in 1871. He described two different forms of LE: discoid lupus and an "aggregated form." Kaposi was also the first to describe non-dermatological symptoms: "...various grave and even dangerous constitutional symptoms may be intimately associated... and that death may result." He named fever, pleuritis, pneumonia, and anemia as related symptoms. Between 1895 and 1903, Osler reported complications of *erythema exudativum multiforme* (which probably included a wide variety of disorders) including non-deforming arthritis, endocarditis, pericarditis, central nervous system symptoms, and acute nephritis. Importantly, Osler noted that the disease could be present with no dermatologic manifestations. Jadassohn also reported the systemic nature of lupus erythematosus around the same time. Libman and Sacks reported a non-bacterial vegetative endocarditis in 1923. "Libman-Sacks" endocarditis was later recognized as lupus erythematosus – two patients had a malar rash, and all had some typical SLE visceral manifestations. The "wire-loop" kidney lesions now considered to be typical of SLE were first described as being nearly unique

to lupus erythematosus in 1935 by Baehr, Klemperer, and Schifrin (Lahita, 1987).

Hargreaves, Richmond, and Morton described the "LE cell" in 1948, prompting further investigation of serum abnormalities (Lahita, 1987). Friou discovered the presence of anti-nuclear antibodies (ANA) in patients with LE, and the identification of specific autoantibodies such as anti-DNA followed (Hochberg, 1993).

By the time of publication of Dubois' seminal 1966 textbook on lupus, the concept of lupus as a widely heterogeneous disease was firmly in place (Dubois, 1966). Dubois described lupus erythematosus as "a syndrome whose manifestations range from a localized skin lesion to a destructive systemic disorder without any cutaneous changes."

1.3.3 Classification Criteria for SLE

1971 ARA Criteria

In response to a lack of uniformity in the definition of SLE in research, the American Rheumatism Association (ARA) released the Preliminary Criteria for the Classification of SLE in 1971 (Cohen *et al.*, 1971). The panel charged with creating the SLE classification criteria devised a format similar to the previously devised rheumatoid arthritis (RA) criteria: a person would be classified as having SLE if they had at any time exhibited 4 or more of a list of 14 clinical symptoms or signs. The stated purpose of this classification criteria was to create a uniform classification system to allow the comparison of research on the natural history, clinical management, and epidemiology of SLE. The panel emphasized that the criteria were intended for classification use, not for diagnostic use.

The ARA panel solicited SLE cases and controls from 52 rheumatologists. Each rheumatologist selected 5 unequivocal SLE patients (in their own opinion), 5 "possible" SLE patients, 5 rheumatoid arthritis (RA) patients, and 5 controls (no SLE or RA) from their own patient population. Each rheumatologist then responded "present" or "absent" for a list of 74 clinical signs and symptoms. (The 74 items included any sign or symptom felt by at least one member of the ARA panel to be an SLE manifestation).

The criteria were published with a reported "greater than 90%" sensitivity although no external validation study had been performed. In an analysis of patients with physician-diagnosed SLE in the Stanford databank of rheumatic disease, Fries and Siegel (1973) found that 57.2% of patients had four or more ARA criteria at the first visit. When cumulative findings were considered, 73.3% of SLE patients had four or more ARA criteria.

The 1982 ACR criteria

The ARA re-evaluated the 1971 SLE criteria to address the concerns about sensitivity and to incorporate serological measures that had not been in widespread use when the 1971 criteria were developed (Tan *et al.*, 1982). (These criteria are now referred to as the ACR criteria after the new name of the ARA, the American College of Rheumatology.) The panel asked 18 rheumatologists to choose their next 10 SLE patients and 10 non-SLE controls (age-, race-, and sex-matched patients with non-traumatic, non-degenerative connective tissue disorders).

The sensitivity and specificity of 30 combinations of items was tested. The panel reduced the number of criteria from 14 to 11, with four for mucocutaneous manifestations. Each of the non-mucocutaneous criteria represented one organ system, with several manifestations that would qualify as a "positive" for that organ system. The "immunologic disorder" criterion was modified in 1997 to include anti-phospholipid antibodies (Hochberg, 1997). One-third of the sample was held aside from the original analyses for validation purposes. The panel found that the sensitivity and specificity of the new criteria were 96%. The panel considered thresholds other than 4 items but found that they "could determine no advantage to a different threshold level."

Critiques of ACR criteria

The creators of the ACR Classification Criteria for SLE did not believe that the criteria were infallible. In fact, one of the purposes of creating the criteria was to "provide a focal definition for criticism" that would lead to modifications of the definition of SLE (Fries, 1987). According to Fries, "the bottom line is that disease criteria organize clinical knowledge, and that improved knowledge improves disease criteria."

There are several problems with the method for development of the ACR criteria, their function as criteria, and their use in research (Edworthy *et al.*, 1988; Petri and Magder, 2004; Smith and Shmerling, 1999). Studies addressing these problems can provide important new knowledge with which to address the definition of SLE and related disorders.

Use of expert opinion as gold standard. In creation of the SLE classification criteria, the American College of Rheumatology used consensus expert opinion as the gold standard. There are several problems with this method.

First, the use of expert opinion as a gold standard implies that there is a shared construct of SLE among the experts, meaning that experts all define SLE in the same way. However, the ACR classification criteria development panel excluded some manifestations from consideration as criteria because experts did not agree on their inclusion, showing that the experts did not all define SLE in the same way. In an investigation of

Table 1.1: The ACR Criteria for the Classification of SLE.

Criterion	Definition
1. Malar "butterfly" rash	Fixed erythema, flat or raised, over the malar eminences, tending to spare the nasolabial folds
2. Discoid rash	Erythematous raised patches with adherent keratotic scaling and follicular plugging; atrophic scarring may occur in older lesions
3. Photosensitivity	Skin rash as a result of unusual reaction to sunlight, by patient history or physician observation
4. Oral ulcers	Oral or nasopharyngeal ulceration, usually painless, observed by a physician
5. Arthritis	Nonerosive arthritis involving two or more peripheral joints, characterized by tenderness, swelling, or effusion
6. Serositis	a. Pleuritis (convincing history of pleuritic pain or rub heard by physician or evidence of pleural effusion), *OR* b. Pericarditis (documented by ECG, rub, or evidence of pericardial effusion)
7. Renal disorder	a. Persistent proteinuria ($>$ 0.5 grams/day or $>$ 3+ if quantitation not performed), *OR* b. Cellular casts (may be red cell, hemoglobin, granular, tubular, or mixed)
8. Neurologic disorder	a. Seizures (in the absence of offending drugs or known metabolic derangements; e.g. uremia, ketoacidosis, or electrolyte imbalance) *OR* b. Psychosis (in the absence of offending drugs or known metabolic derangements; e.g. uremia, ketoacidosis, or electrolyte imbalance)
9. Hematologic disorder	a. Hemolytic anemia (with reticulocytosis), *OR* b. Leukopenia ($<$4000/ml total on two or more occasions), *OR* c. Lymphopenia ($<$1500/ml on two or more occasions), *OR* d. Thrombocytopenia ($<100,000$/ml in the absence of offending drugs)
10. Immunologic disorder	a. Anti-DNA (antibody to native DNA in abnormal titer), *OR* b. Anti-Sm (presence of antibody to Sm nuclear antigen), *OR* c. Positive finding of antiphospholipid antibodies based on 1) an abnormal serum level of IgG or IgM anticardiolipin antibodies, 2) a positive test results for lupus anticoagulant using a standard ethod, or 3) a false-positive serologic test for syphilis known to be positive for at least 6 months and confirmed by *Treponema pallidum* immobilization (TPI) or fluorescent treponemal antibody (FTA) absorption test
11. Antinuclear antibody (ANA)	Abnormal titer of ANA by immunofluorescence or an equivalent assay at any point in time and in the absence of drugs known to be associated with "drug-induced lupus" syndrome

The proposed classification is based on 11 criteria. For the purpose of identifying patients in clinical studies, a person shall be said to have SLE if any four or more of the 11 criteria are present, serially or simultaneously, during any interval of observation. (Tan *et al.*, 1982; Hochberg, 1997)

physicians' concepts of another rheumatic disease, psoriatic arthritis (PsA), Symmons *et al.* (2006) found that agreement of the existence of a disease entity by experts did not ensure that the experts shared a disease construct. Although all rheumatologists surveyed agreed that PsA is a "true disease entity," there was wide variation in how they diagnosed standardized patients.

The second problem with use of expert opinion as the gold standard is the assumption that if experts do have a shared construct, it is "right."

The third problem is that even if experts have a shared construct and that construct does describe a group of patients that share a disease, a case definition based on the shared construct is not necessarily a good first step in the description of a disease. Even if the experts agree on diagnosis on a case-by-case basis, the shared construct may encompass a group of patients with heterogenous manifestations.

Dermatologic criteria. According to a position paper on the ACR criteria by a group of dermatologists, the ACR criteria over-emphasize dermatologic criteria (Albrecht *et al.*, 2004). Four of the ACR criteria are individual dermatologic manifestations, while manifestations from other organ systems are grouped into single criteria. Paradoxically, a patient with dermatologic but no systemic (i.e. not other than dermatologic) manifestations can meet the criteria for SLE. Because the cases and controls used for development of the ACR criteria were provided by rheumatologists, the controls may have under-represented dermatologic conditions that appear similar to the ACR criteria dermatologic items. As a result, the specificity of the dermatologic criteria items is much lower in practice than was reported by the ACR criteria committee (Albrecht *et al.*, 2004).

Limited scope of manifestations. The ACR criteria do not include some organ systems that can be affected by SLE, such as the gastrointestinal system. In organ systems that are included in the criteria, the manifestations that are part of the criteria are sometimes a poor representation of all manifestations possible for the organ system. For example, the ACR criteria includes two neurologic manifestations, seizures and psychosis, while an ad hoc ACR committee on neuropsychiatric lupus nomenclature described 19 possible neuropsychiatric lupus syndromes (Liang *et al.*, 1999).

Performance of ACR criteria in early or possible SLE.

Cross-sectional design. The ACR criteria were designed using symptoms and signs experienced during each patient's entire follow-up time. This method fails to capture the nature of symptom accrual or the unfolding nature of disease evolution.

ACR criteria accrual. Most large-scale reports of the accrual of ACR criteria elements have been for patients who eventually were classified with SLE by the ACR criteria. We cannot make inferences to patients with less than 4 ACR criteria (at a given time) because some such patients never have 4 criteria and are thus

not included in these studies. However, these studies can help to generate hypotheses about types of onset.

Among the Hispanic SLE patients from Texas in the LUMINA study, 53.5% had "acute onset," defined as 4 ACR criteria in 1 month or less. In contrast, only 5.3% of Hispanic SLE patients from Puerto Rico had an acute onset (Vila *et al.*, 2004). In the entire LUMINA cohort (also including African American and Caucasian patients), 55.4% had 1 ACR criterion as the initial manifestation of SLE. 20.0% presented had 2 criteria at onset, 9.3% had 3, and 15.3% had 4 (Alarcon *et al.*, 2004). The median time to the development of 4 ACR criteria was 9.1 months.

In a study of active duty military participants who eventually developed 4 or more ACR criteria, the average time from first manifestation to the date of meeting 4 criteria was 1.6 years, with medians of 0.17 and 0.67 years for African-American males and females, respectively, and 0.88 and 0.50 for European-American males and females (Arbuckle et al 2003a).

"Incomplete" lupus. Some studies have looked prospectively at patients with less than 4 ACR criteria. Various names have been used for such patients, including "incomplete lupus," "latent lupus," "incipient lupus," and "borderline SLE" (Swaak *et al.*, 2001; Vila *et al.*, 2000; Ganczarczyk *et al.*, 1989; Al Attia, 2006).

In one study of 28 patients who had less than 4 ACR criteria, 16 developed 4 or more ACR criteria in follow-up ranging from 10 to 20 years (Stahl-Hallengren *et al.*, 2004). Among these 16 patients, the median time from first visit to fulfillment of the ACR criteria was 5.3 years, with a range of 1 to 10 years. Early damage as measured by the SLICC/ACR damage index, malar rash, and positive aCL predicted development of complete SLE. (Six patients with malar rash at the first visit and six patients with positive aCL eventually met the ACR criteria.) Patients who remained as incomplete SLE were less likely to develop organ damage.

In a cross-sectional analysis comparing 38 patients with incomplete lupus erythematosus (less than 4 ACR criteria) and 42 with complete SLE, patients with incomplete LE were more likely to have had discoid rash (Greer and Panush, 1989). ILE patients had symptoms for an average of 38 months before first rheumatologist evaluation compared to 9 months for SLE patients. In 57 months of follow-up, two of the ILE patients developed SLE, 10 were diagnosed with discoid lupus, and 2 with subacute cutaneous lupus. 20 had persistant ILE, meaning that they did not develop another rheumatic disease and were not diagnosed with a nonrheumatic disease explaining their symptoms.

(Lom-Orta *et al.*, 1980) found that of 31 patients with physician-diagnosed SLE but less than 4 criteria, 21 accumulated at least 4 criteria in and average of 41 months of follow-up time.

The differences between these studies may reflect different concepts of incomplete SLE, with some physicians or researchers less likely to include patients with only non-specific symptoms.

ACR criteria and early lupus. Taken together, studies of incomplete lupus and ACR criteria accrual suggest that the ACR criteria are not well-suited for the classification of early lupus, particularly slowly-evolving lupus, or in patients who may have one of several autoimmune diseases. These studies also indicate that presenting symptoms may help physicians predict the likelihood of severe consequences of SLE.

Alternative criteria

Clough *et al.* (1984) designed a weighted diagnostic criteria based on the 1971 ACR Criteria. If the sum of a patient's weights met or exceeded 2.0 points, the patient was diagnosed with SLE. High weight was given to specific items such as cellular casts and discoid skin lesions (each 1.5), low weight to nonspecific items such as psychosis or arthritis, and negative weight (-1.8) was given to a negative ANA test.

In 2002, Costenbader *et al.* (2002) created the Boston Weighted Criteria, which were based on the Clough weighted criteria but intended for research classification purposes. The primary purpose of the Boston Weighted Criteria was to increase the sensitivity for clinical studies. They made several modifications, including dropping items with very low specificity (alopecia, Raynaud's syndrome), and new serologic items were added (anti-phospholipid, anti-beta2-glycoprotein antibodies). In contrast to Clough's criteria, Costenbader *et al.* (2002) included an item that was in itself sufficient for SLE classification, renal pathology of World Health Organization glomerulonephritis (WHO GN) classes 3 through 6. Also in contrast to Clough's criteria, the four serologic tests included were each given a weight of 0.5. A score of 2.0 or greater indicated SLE classification.

Sanchez *et al.* (2003) found that the Boston Weighted Criteria had a sensitivity of 90% and specificity of 60% in their SLE population compared to 86% and 72%, respectively, for the ACR criteria.

Clinical diagnosis vs. research classification

It is often stated in the literature that classification criteria are not intended for diagnostic purposes (Tan *et al.*, 1982). However, the line between research classification criteria and clinical diagnosis is often blurred. The criteria are probably sometimes used for diagnosis, and some studies have equated the time that a patient meets 4 ACR criteria as time of diagnosis (Panush *et al.*, 1993; Arbuckle *et al.*, 2003).

Classification criteria cannot take the place of clinical diagnosis for several reasons. Criteria intended to classify patients for research purposes cannot practically include every factor that physicians might consider in making a diagnosis. Furthermore, diagnosis is a subjective process because no two patients are alike, and this "artful" element of diagnosis will undoubtedly remain a feature of the diagnosis of complex diseases. However, in order for research to be applicable to patients diagnosed with the disease, the classification criteria should agree with diagnosis to the extent that is feasible. Fries writes that ideally, classification criteria would be identical to diagnostic criteria (Fries *et al.*, 1994).

Some authors see little to worry about when classification criteria and diagnosis do not agree well. One member of the ACR Diagnostic and Therapeutic Criteria Committee appears to have used disagreement between experienced rheumatologists' diagnoses of vasculitis and classification by the ACR Vasculitis Criteria as evidence that it is acceptable for classification criteria to perform poorly against individual diagnoses (Hunder, 1998).

The difference between diagnosis and classification criteria is not simply a philosophical or semantic construction. In the absence of an accepted disease definition or "diagnostic criteria" set, a "classification criteria" set becomes the *de facto* SLE definition for non-expert physicians. Fries (1987) writes that one purpose of classification criteria is for teaching of the cardinal disease features. Because diagnosis by an expert is the diagnostic gold standard for SLE, the allowance of a separation between diagnosis and the classification criteria can have detrimental consequences – the diagnosis or referral of a patient with SLE might be delayed, or conversely, a patient without SLE might be diagnosed as having SLE. In one study, close to a third (29%) of patients referred to a rheumatology center for SLE had a positive ANA test but no autoimmune disease diagnosed by the center. Half (51%) of these patients had been treated with corticosteriods (Narain *et al.*, 2004).

Perspectives from other rheumatic diseases

Psoriatic arthritis is another rheumatic disease embroiled in classification debates and faced with the history of definition by expert opinion. Symmons *et al.* (2006) suggest that a prospective study of the effect of a history of psoriasis in patients with any kind of inflammatory arthritis would be an independent way to examine the relationship between psoriasis and arthritis, or in other words would be a good way to avoid the subjectivity of using expert opinion as the gold standard. Similarly, Nived and Sturfelt (2005) and Ward (2005) suggest that we need to develop classification and diagnostic criteria on unselected patients in order to avoid bias from type of medical care center.

1.3.4 Subtypes in SLE

The goal of creating the ACR criteria was to develop a "unified" classification system, but the result of the effort is actually the antithesis of a unified definition. ACR-classified SLE and physician-diagnosed SLE has a very wide variety of manifestations, presentations, and severity. Given that the etiology and outcomes of SLE are largely unknown, one possible way to make progress towards a revised disease definition is to identify homogenous groups of patients.

Dubois (1966) firmly espoused the idea that SLE consists of clinical subsets. Other authors have stated beliefs that there are SLE subsets based on research findings or clinical experience (Calvo-Alen *et al.*, 1995; Petri and Magder, 2004; To and Petri, 2005).

Formal investigation of subtypes. To and Petri (2005) used K-means cluster analysis to identify three autoantibody clusters in the Hopkins Lupus Cohort. They described a cluster with a high percentage of members with anti-Sm and anti-RNP autoantibodies relative to the other clusters, a cluster with a high or relatively high percentags of anti-dsDNA, anti-Ro, and anti-La autoantibodies, and a cluster with high percentages of anti-dsDNA and anticardiolipin autoantibodies and relatively high lupus anticoagulant.

Jurencak *et al.* (2009) described three antibody-based clusters of pediatric-onset SLE patients; a cluster with a high proportion of dsDNA antibodies, a cluster with high proportions of anti-dsDNA, antichromatin, antiribosamal P, anti-U1RNP, anti-Sm, anti-Ro, and anti-La antibodies, and a cluster with a high proportion of anti-dsDNA, anti-RNP, and anti-Sm antibodies. Patients in the cluster with low percentages of all autoantibodies except for anti-DNA was more likely to be Caucasion than other clusters.

1.4 Georgia Lupus Registry

The Georgia Lupus Registry (GLR) is an ongoing population-based registry led by the Division of Rheumatology at Emory University. The GLR is one of two lupus registries sponsored by the Centers for Disease Control and Prevention Arthritis Program, and also receives funding from the Georgia Department of Community Health.

The primary aim of the GLR is to ascertain the prevalence in 2002 and incidence in 2002-2004 of SLE in the two counties (Fulton, Dekalb) of Atlanta, Georgia.

Case ascertainment

Because the registry is acting as a public health agent under the auspices of the state, the registry has a waiver of consent and is able to request and review medical records without patient consent. Sources for potential registry cases include hospitals, physician offices (rheumatology, nephrology, dermatology), laboratories, pathology reports (skin, renal), public media, and patient advocacy groups. Potential cases are identified by screening administrative data for keywords or ICD-9 diagnostic codes for (a) lupus erythematosus (discoid), (b) systemic lupus erythematosus, (c) other specified connective tissue disease, (d) unspecified connective tissue disease, (e) hemorrhagic disorder due to circulating anticoagulant, which encompasses antiphospholipid antibody syndrome, and (f) lupus nephritis. Other methods of case identification include laboratory results, pathology reports, and self-referral by patients.

Case definition: "potential lupus"

Data is abstracted from medical records for patients with a physician diagnosis of potential or definite lupus, who lived in Dekalb or Fulton counties in 2002, 2003, or 2004. Keywords that constitute a diagnosis of potential or definite lupus (when used with the word lupus) include: apparent(ly), appears to, comparable

with, compatible with, consistent with, favor(s), most likely, presumed, probable, suspect, suspicious, typical of, cannot be ruled out, possible, questionable, suggest, and worrisome.

Data collection

Demographic, clinical, and laboratory data are abstracted from medical records (physician notes, hospital admission or discharge notes, laboratory or pathology reports) by trained abstractors.

Demographic information. Race, sex, Spanish/Hispanic ethnicity, marital status, smoking status, and primary paper at diagnosis and currently are abstracted when available.

Lupus identification. The method used to identify an SLE patient (for example, the triggering ICD-9 code), the date of diagnosis, and the final stated diagnosis are recorded.

Final diagnosis categories are drug-induced lupus, discoid lupus (without systemic features), primary antiphospholipid syndrome, other specified connective tissue disease, SLE and other connective tissue disease, unspecified diffuse connective tissue disease, chronic or subacute cutaneous/non-discoid lupus limited to the skin, and no stated diagnosis or miscoded subject (none of the possible final diagnoses are stated by a physician). "Other connective tissue disease" is further specified as systemic sclerosis or scleroderma, primary Sjögren Syndrome, polymyositis/dermatomyositis, rheumatoid arthritis, mixed connective tissue disease, systmic vasculities (includes a large group of diseases), relapsing polychondritis, or other. The type of physician that stated the final diagnosis is recorded if the diagnosis is made by one or more rheumatologists, one or more dermatologists, one or more nephrologists, or some combination of these three specialties. Physician type can also be recorded as a pathologist or "no rheumatologist, dermatologist, or nephrologist."

Date of diagnosis is defined as the earliest date a physician first stated diagnosis of SLE or related connective tissue disease (listed in previous paragraph), using the diagnostic keywords described previously.

From medical records or progress notes

Lupus manifestation variables: overview. Two types of lupus manifestations are abstracted, elements of the American College of Rheumatology (ACR) Classification Criteria for SLE (Tan *et al.*, 1982) and manifestations that are not in the ACR criteria. If an ACR criterion can be met by one or more of a selection of manifestations, each manifestation is treated as a separate variable in registry data collection. The first date of occurrence is recorded for manifestations that are elements of the ACR criteria, along with the manner in which the first date of occurrence was ascertained (physician listed specific date in record, date estimated from historical document in medical record, patient report only, unknown/not documented).

Clinical variables. *Mucocutaneous.* ACR criteria elements abstracted are malar rash, discoid rash, photosensitivity, and mucosal ulcers. Non-ACR variables are subacute cutaneous LE (SCLE), bullous skin

lesions, panniculitis, alopecia, cutaneous vasculitis, Raynaud's phenomenon, livedo reticularis, urticaria, chilblain lupus, lupus tumidus, and LE mucosus. *Pulmonary/cardiovascular.* ACR elements include serositis, pleuritis, and pericarditis. Non-ACR variables are interstitial lung disease, pneumonitis, and myocardial infarction. *Gastrointestinal.* Peritonitis is abstracted. *Renal.* Dialysis, renal transplant, lupus nephritis. *Neurologic.* ACR elements are seizures and psychosis, both "not attributable to another disease process." Non-ACR variables are mononeuritis multiplex, cerebrovascular disease, transverse myelitis, aseptic meningitis, chorea, and cranial and/or peripheral neuropathy. *Muskuloskeletal.* One variable, arthritis, is an element of an ACR criteria item. Non-ACR variables are Jaccoud's arthropathy, avascular necrosis, myositis, and variables related to possible myositis (proximal weakness, elevated muscle enzymes, muscle biopsy positive for inflammation, electromyogram, MRI suggestive of muscle inflammation.).

Laboratory variables. These laboratory variables are recorded as positive if keywords are present in the medical record or progress note. (Data abstracted directly from laboratory reports is described in a later section.) *Hematologic.* ACR criteria elements are hemolytic anemia and lupus anti-coagulant, the non-ACR variable is antiphospholipid syndrome. *Immunologic.* False-positive syphilis test is an ACR element, low complements is not.

SLE damage. Some of the variables collected indicate (possible) damage caused by SLE: myocardial infarction, cerebrovascular disease, Jaccoud's arthropathy, avascular necrosis (can be caused by disease processes or iatrogenically), end-stage renal disease.

Laboratory and pathology variables

Hematologic and urine laboratory manifestations are recorded as positive if they occur two or more times, with six or more months between occurrences. *Hematologic.* All hematologic variables collected are elements of the ACR criteria: leukopenia, lymphopenia, and thrombocytopenia. *Urine.* Several variables describe proteinuria, and ACR criteria element: 24-hour urine protein (>500mg or $\geq$ 3 grams), dipstick protein on urinalysis, and spot protein to creatinine ratio on urinalysis ($>$ 0.5, $\geq$ 3). The occurrence of cellular casts on urinalysis is recorded. *Immunologic.* Antinuclear antibody (ANA), a distinct ACR criterion, is abstracted. Elements of the "immunologic" ACR criterion are anti-DNA or anti-ds DNA, anti-Sm (anti-Smith), anticardiolipin antibodies, and false positive syphilis test. Non-ACR immunologic variables are anti-RNP, anti-Ro/SSA, anti-la/SSB, or anti-beta2 glycoprotein antibodies, rheumatoid factor, low C3, and low C4 (complement).

Pathology. Information on renal biopsies is recorded.

CHAPTER 2

Cluster analysis: Literature review

Class discovery is the process of "discovering," or uncovering, previously undescribed classes in a heterogenous population. In other words, class discovery is the process of identifying homogenous classes. This type of method has been used in many research area, including bacteriology, where it is known as *numerical taxonomy*, genetics, paleontology, and psychology.

The terms *classification* and *cluster analysis* are sometimes used to denote what we have described as class discovery. Following the convention of most medical literature, we will reserve the term *classification* to refer to the process of assigning individuals to previously described classes, i.e. to classes that have already been "discovered." We will use *cluster analysis* to refer to a specific family of class discovery methods.

Class discovery methods. A variety of methods have been used for the purposes of class discovery, including cluster analysis, finite mixture models (which includes latent class analysis), factor analysis, principal components analysis, and neural networks. Of these methods, cluster analysis and finite mixture models are the most frequently used. (In fact, within each of these families, there are numerous examples of researchers in different fields seemingly inventing the same method independently.)

Cluster analysis uses ad-hoc algorithms to place objects in a dataset into clusters, according to some definition of "distance" or similarity between objects, so that the objects are placed into homogenous clusters. In finite mixture modeling, probabilistic models of underlying population mixtures are fit to the data. In my research, I explored aspects of K-centroids, a popular type of cluster analysis.

2.1 K-centroids cluster analysis

2.1.1 Cluster analysis overview

Cluster analysis refers to a family of procedures that place objects in a data set into clusters in such a way as to maximize intra-cluster similarity and between-cluster dissimilarity. While the term is sometimes used as an umbrella term for all methods that are used for class discovery, the term is used in this manuscript to refer to ad-hoc numerical algorithms that place objects into clusters based on some measure of similarity. Cluster analysis does not entail the estimation of statistical models underlying the observed data; objects in the dataset are placed into clusters based on their similarity to each other.

Hierarchical cluster analysis. In hierarchical cluster analysis (HCA), objects are placed into a hierarchy based on object-to-object similarity. A hierarchical clustering is often represented with a dendrogram, or clustering tree. HCA can be performed by successively dividing the data (divisive HCA), but is more commonly performed by starting with every object as a cluster and then successively grouping together clusters or objects and clusters (agglomerative HCA). Because HCA algorithms place objects into a hierarchy, a decision about where to "cut" the hierarchy must be made if a partition of the data is desired (i.e., a set of mutually exclusive clusters).

Ward's method for hierarchical cluster analysis has a close relationship to the most common partition-based cluster analysis method, K-means. In Ward's method, each data point is initially a unique cluster, so that there are N clusters. In each step of the agglomeration, the two clusters are joined that yield the lowest "loss" criterion (Ward, 1963). Ward suggested using the within-cluster sum of squared error (SSW, described in a later section) as the loss criterion, and his suggested method with this criterion is often referred to as "Ward's method."

Partition-based cluster analysis. Partition-based cluster analysis methods, sometimes known as optimization methods, divide the data into groups in such a way as to optimize a given criterion for the "goodness" of a partition.

The K-means cluster analysis method is perhaps the most widely used partition-based clustering method (Steinley, 2006a). The most common form of the K-means algorithm minimizes the total within-cluster sum of squared error, or the average distances from each object in the data set to the mean of the cluster to which it belongs.

The remaining parts of this section will focus exclusively on partition-based clustering methods. The K-means cluster analysis method and its generalized form, K-centroids, will be described.

Notation and terminology. The $N \times M$ matrix $\mathbf{X}$ is data points for N individuals measured on M attributes,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & \\ \vdots & & & \\ x_{N1} & & \dots & x_{NM} \end{bmatrix}_{N \times M},$$

where the element x_{im} is the measurement for the i-th individual on the m-th attribute. An individual might also be called an object, and an attribute might also be called a dimension or variable. The data point $\mathbf{x_i} = [x_{i1} \dots x_{im}]$ is the data for the i-th individual.

In partition-based cluster analysis, N data points are divided into K clusters, C_1, C_2, ..., C_K, based on their observed values for M attributes. In the disease definition setting, we might think of this as dividing N patients into K putative subtypes, based on the presence or absence of M symptoms. A *partition*, denoted by $\mathscr{P}$, is a particular set of mutually exhaustive clusters. The term *clustering* is used to refer to a set of clusters of data that are the result of the application of a clustering method to the data. (A partition of the data is a clustering of the data, as is a nested tree of similarity given by a hierarchical clustering method.)

The term *clusters* refers to groups that result from the application of a clustering method to the data, and *subtypes* to refer to naturally-occurring groups in the data.

2.2 The K-centroids method

K-centroids is the generalized form of the K-means cluster analysis method. The K-means method will be presented first, then generalized forms including K-medians and K-modes.

The K-means algorithm is an iterative procedure for finding a partition of the data that minimizes the sum of within-cluster distances from individuals to cluster means (Bock, 2007).

2.2.1 The minimum within-cluster sum of squares criterion.

In "standard" K-means cluster analysis, the sum of within-cluster squared error is minimized. In other words, individuals are placed into clusters in such a way as to minimize the sum of the squared Euclidean distances from individuals to their respective cluster means, where the kth cluster mean is defined as a vector of the cluster-specific sample means for each attribute,

$$\bar{\mathbf{x}}_\mathbf{k} = \begin{bmatrix} \bar{x}_1^{(k)} & \bar{x}_2^{(k)} & \dots & \bar{x}_M^{(k)} \end{bmatrix}_{1 \times M},$$

and the squared Euclidean distance from data point $\mathbf{x_i}$ to cluster mean $\bar{\mathbf{x}}^{(\mathbf{k})}$ is

$$d^2(\mathbf{x_i}, \bar{\mathbf{x}}_\mathbf{k}) = \sum_{m=1}^{M} (x_{im} - \bar{x}_m^{(k)})^2.$$

The sum of within-cluster squared error for the k-th cluster (SSW_k) is

$$SSW_k = \sum_{i \in C_k} \sum_{m=1}^{M} (x_{im} - \bar{x}_m^{(k)})^2. \tag{2.1}$$

The SSW, or pooled SSW (the sum of all SSW_k), is

$$SSW = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{m=1}^{M} (x_{im} - \bar{x}_m^{(k)})^2 = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x_i} - \bar{\mathbf{x}}^{(\mathbf{k})}\|^2. \tag{2.2}$$

The SSW is an overall measure of cluster compactness.

Matrix formulation

K-means clustering is often described in terms of $\mathbf{W}$, the pooled within-cluster cross-products and sum-of-squares matrix (e.g. Everitt, 1980; Xu and Wunsch, 2009). $\mathbf{W}$ is defined as follows.

The vector of attribute-wise distances $\mathbf{d_{ik}}$ from data point $\mathbf{x_i}$ to cluster mean $\bar{\mathbf{x}}^{(\mathbf{k})}$ is

$$\begin{aligned}
\mathbf{d_{ik}} &= \mathbf{x_i} - \bar{\mathbf{x}}^{(\mathbf{k})} \\
&= \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{iM} \end{bmatrix} - \begin{bmatrix} \bar{x}_1^{(k)} & \bar{x}_2^{(k)} & \dots & \bar{x}_M^{(k)} \end{bmatrix} \\
&= \begin{bmatrix} (x_{i1} - \bar{x}_1^{(k)}) & (x_{i2} - \bar{x}_2^{(k)}) & \dots & (x_{iM} - \bar{x}_M^{(k)}) \end{bmatrix}
\end{aligned}$$

The matrix of squares and cross-products of deviations for $\mathbf{x_i}$ compared to $\bar{\mathbf{x}}^{(\mathbf{k})}$ is obtained by multiplying $\mathbf{d'_{ik}}$ by $\mathbf{d_{ik}}$,

$$\begin{aligned}
\mathbf{d'_{ik}}\mathbf{d_{ik}} &= \begin{bmatrix} (x_{i1} - \bar{x}_1^{(k)}) \\ (x_{i2} - \bar{x}_2^{(k)}) \\ \dots \\ (x_{iM} - \bar{x}_M^{(k)}) \end{bmatrix} \begin{bmatrix} (x_{i1} - \bar{x}_1^{(k)}) & (x_{i2} - \bar{x}_2^{(k)}) & \dots & (x_{iM} - \bar{x}_M^{(k)}) \end{bmatrix} \\
&= \begin{bmatrix} (x_{i1} - \bar{x}_1^{(k)})^2 & (x_{i1} - \bar{x}_1^{(k)})(x_{i2} - \bar{x}_2^{(k)}) & \dots & (x_{i1} - \bar{x}_1^{(k)})(x_{iM} - \bar{x}_M^{(k)}) \\ (x_{i2} - \bar{x}_2^{(k)})(x_{i1} - \bar{x}_1^{(k)}) & (x_{i2} - \bar{x}_2^{(k)})^2 & \dots & (x_{i2} - \bar{x}_2^{(k)})(x_{iM} - \bar{x}_M^{(k)}) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{iM} - \bar{x}_M^{(k)})(x_{i1} - \bar{x}_1^{(k)}) & \dots & \dots & (x_{iM} - \bar{x}_M^{(k)})^2 \end{bmatrix}_{M \times M}
\end{aligned}$$

For C_k (the k-th cluster), $\mathbf{W_k}$ is the sum of $\mathbf{d'_{ik}d_{ik}}$ over all data points in C_k,

$$\mathbf{W_k} = \begin{bmatrix} \sum_{i\in C_k}(x_{i1}-\bar{x}_1^{(k)})^2 & \sum_{i\in C_k}(x_{i1}-\bar{x}_1^{(k)})(x_{i2}-\bar{x}_2^{(k)}) & \dots & \sum_{i\in C_k}(x_{i1}-\bar{x}_1^{(k)})(x_{iM}-\bar{x}_M^{(k)}) \\ \sum_{i\in C_k}(x_{i2}-\bar{x}_2^{(k)})(x_{i1}-\bar{x}_1^{(k)}) & \sum_{i\in C_k}(x_{i2}-\bar{x}_2^{(k)})^2 & \dots & \sum_{i\in C_k}(x_{i2}-\bar{x}_2^{(k)})(x_{iM}-\bar{x}_M^{(k)}) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i\in C_k}(x_{iM}-\bar{x}_M^{(k)})(x_{i1}-\bar{x}_1^{(k)}) & \dots & \dots & \sum_{i\in C_k}(x_{iM}-\bar{x}_M^{(k)})^2 \end{bmatrix}_{M\times M}$$

The (m,m)-th diagonal element of $\mathbf{W}$ is the sum of squared error for the m-th attribute, in the k-th cluster. The trace of a square matrix is the sum of its diagonal elements. Thus, trace$(\mathbf{W_k})$ is equivalent to the within-clusters sum-of-squared-error for cluster C_k (SSW_k, Equation (2.1)),

$$SSW_k = \text{tr}(\mathbf{W_k}).$$

The pooled $\mathbf{W}$, usually simply called $\mathbf{W}$, is the sum of the cluster-specific within-cluster cross-products and sum-of-squares matrices,

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{W_k}. \tag{2.3}$$

The total sum of within-cluster squared error (SSW), Equation (2.2), is equal to the trace of the pooled $\mathbf{W}$,

$$SSW = \text{tr}(\mathbf{W}). \tag{2.4}$$

The sample covariance matrix, $\mathbf{S}$, (for dataset $\mathbf{X}$) is obtained by multiplying every element of $\mathbf{W}$ by $\left(\frac{1}{n-1}\right)$,

$$\mathbf{S} = \left(\frac{1}{n-1}\right)\mathbf{W}.$$

The matrix $\mathbf{T}$ is the total scatter matrix of the data, and describes deviations of the data from the grand mean. The deviation of data point $\mathbf{x_i}$ from the grand mean vector $\bar{\mathbf{x}}$ is $\mathbf{d_i}$, where

$$\begin{aligned} \mathbf{d_i} &= \mathbf{x_i} - \bar{\mathbf{x}} \\ &= \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{iM} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_M \end{bmatrix} \\ &= \begin{bmatrix} (x_{i1}-\bar{x}_1) & (x_{i2}-\bar{x}_2) & \dots & (x_{iM}-\bar{x}_M) \end{bmatrix}. \end{aligned}$$

Multiplying $\mathbf{d}_\mathbf{i}'$ by $\mathbf{d}_\mathbf{i}$ gives a matrix of squares and cross-products of errors for the i-th data point,

$$\begin{aligned}
\mathbf{d}_\mathbf{i}'\mathbf{d}_\mathbf{i} &= \begin{bmatrix} (x_{i1}-\bar{x}_1) \\ (x_{i2}-\bar{x}_2) \\ \cdots \\ (x_{iM}-\bar{x}_M) \end{bmatrix} \begin{bmatrix} (x_{i1}-\bar{x}_1) & (x_{i2}-\bar{x}_2) & \ldots & (x_{iM}-\bar{x}_M) \end{bmatrix} \\
&= \begin{bmatrix} (x_{i1}-\bar{x}_1)^2 & (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \ldots & (x_{i1}-\bar{x}_1)(x_{iM}-\bar{x}_M) \\ (x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & (x_{i2}-\bar{x}_2)^2 & \ldots & (x_{i2}-\bar{x}_2)(x_{iM}-\bar{x}_M) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{iM}-\bar{x}_M)(x_{i1}-\bar{x}_1) & \ldots & \ldots & (x_{iM}-\bar{x}_M)^2 \end{bmatrix}_{M\times M},
\end{aligned}$$

and the matrix $\mathbf{T}$ is the sum of $\mathbf{d}_\mathbf{i}\mathbf{d}_\mathbf{i}'$ over all datapoints,

$$\mathbf{T} = \begin{bmatrix} \sum_i (x_{i1}-\bar{x}_1)^2 & \sum_i (x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) & \ldots & \sum_i (x_{i1}-\bar{x}_1)(x_{iM}-\bar{x}_M) \\ \sum_i (x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & \sum_i (x_{i2}-\bar{x}_2)^2 & \ldots & \sum_i (x_{i2}-\bar{x}_2)(x_{iM}-\bar{x}_M) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i (x_{iM}-\bar{x}_M)(x_{i1}-\bar{x}_1) & \ldots & \ldots & \sum_i (x_{iM}-\bar{x}_M)^2 \end{bmatrix}_{M\times M}.$$

The total sum of squared error (SST) is the trace of $\mathbf{T}$,

$$\begin{aligned}
SST &= \mathrm{tr}(\mathbf{T}) \\
&= \sum_{m=1}^{M} \sum_i (x_{im}-\bar{x}_m)^2
\end{aligned} \tag{2.5}$$

where $\bar{x}_m^{(k)}$ is the mean of the m-th attribute in the k-th cluster, and $\bar{x}_m$ is the grand mean of the m-th attribute.

The matrix $\mathbf{B}$ is the between-clusters squares of sums and cross-products matrix of the data, and describes deviations of the cluster means from the grand mean. The deviation of the k-th cluster mean vector $\bar{\mathbf{x}}^{(\mathbf{k})}$ from the grand mean vector $\bar{\mathbf{x}}$ is $\mathbf{d}_\mathbf{k}$,

$$\begin{aligned}
\mathbf{d}_\mathbf{k} &= \bar{\mathbf{x}}^{(\mathbf{k})} - \bar{\mathbf{x}} \\
&= \begin{bmatrix} \bar{x}_1^{(k)} & \bar{x}_2^{(k)} & \ldots & \bar{x}_M^{(k)} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_M \end{bmatrix} \\
&= \begin{bmatrix} (\bar{x}_1^{(k)}-\bar{x}_1) & (\bar{x}_2^{(k)}-\bar{x}_2) & \ldots & (\bar{x}_M^{(k)}-\bar{x}_M) \end{bmatrix},
\end{aligned}$$

and the squares and cross-products matrix for the deviation of the k-th cluster mean $\bar{\mathbf{x}}^{(\mathbf{k})}$ from the grand mean $\bar{\mathbf{x}}$ is obtained by multiplying n_k by $\mathbf{d}_\mathbf{k}'\mathbf{d}_\mathbf{k}$,

$$
\begin{aligned}
n_k \mathbf{d}'_{\mathbf{ik}} \mathbf{d}_{\mathbf{ik}} &= n_k \begin{bmatrix} (\bar{x}_1^{(k)} - \bar{x}_1) \\ (\bar{x}_2^{(k)} - \bar{x}_2) \\ \cdots \\ (\bar{x}_M^{(k)} - \bar{x}_M) \end{bmatrix} \begin{bmatrix} (\bar{x}_1^{(k)} - \bar{x}_1) & (\bar{x}_2^{(k)} - \bar{x}_2) & \ldots & (\bar{x}_M^{(k)} - \bar{x}_M) \end{bmatrix} \\
&= n_k \begin{bmatrix} (\bar{x}_1^{(k)} - \bar{x}_1)^2 & (\bar{x}_1^{(k)} - \bar{x}_1)(\bar{x}_2^{(k)} - \bar{x}_2) & \ldots & (\bar{x}_1^{(k)} - \bar{x}_1)(\bar{x}_M^{(k)} - \bar{x}_M) \\ (\bar{x}_2^{(k)} - \bar{x}_2)(\bar{x}_1^{(k)} - \bar{x}_1) & (\bar{x}_2^{(k)} - \bar{x}_2)^2 & \ldots & (\bar{x}_2^{(k)} - \bar{x}_2)(\bar{x}_M^{(k)} - \bar{x}_M) \\ \vdots & \vdots & \ddots & \vdots \\ (\bar{x}_M^{(k)} - \bar{x}_M)(\bar{x}_1^{(k)} - \bar{x}_1) & \ldots & \ldots & (\bar{x}_M^{(k)} - \bar{x}_M)^2 \end{bmatrix}_{M \times M}.
\end{aligned}
$$

$\mathbf{B}$ is the sum of $\mathbf{d}'_{\mathbf{ik}} \mathbf{d}_{\mathbf{ik}}$ over all clusters,

$$
\mathbf{B} = \begin{bmatrix} \sum_k n_k (\bar{x}_1^{(k)} - \bar{x}_1)^2 & \sum_k n_k (\bar{x}_1^{(k)} - \bar{x}_1)(\bar{x}_2^{(k)} - \bar{x}_2) & \ldots & \sum_k n_k (\bar{x}_1^{(k)} - \bar{x}_1)(\bar{x}_M^{(k)} - \bar{x}_M) \\ \sum_k n_k (\bar{x}_2^{(k)} - \bar{x}_2)(\bar{x}_1^{(k)} - \bar{x}_1) & \sum_k n_k (\bar{x}_2^{(k)} - \bar{x}_2)^2 & \ldots & \sum_k n_k (\bar{x}_2^{(k)} - \bar{x}_2)(\bar{x}_M^{(k)} - \bar{x}_M) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_k n_k (\bar{x}_M^{(k)} - \bar{x}_M)(\bar{x}_1^{(k)} - \bar{x}_1) & \ldots & \ldots & \sum_k n_k (\bar{x}_M^{(k)} - \bar{x}_M)^2 \end{bmatrix}_{M \times M}.
$$

The between-clusters sum of squared error (SSB) is the trace of $\mathbf{B}$,

$$
\begin{aligned}
SSB &= \mathrm{tr}(\mathbf{B}) \\
&= \sum_{k=1}^{K} n_k \sum_{m=1}^{M} (\bar{x}_m^{(k)} - \bar{x}_m),
\end{aligned} \tag{2.6}
$$

where $\bar{x}_m^{(k)}$ is the mean of the m-th attribute in the k-th cluster, and $\bar{x}_m$ is the grand mean of the m-th attribute. SSB describes the separation of clusters.

$\mathbf{T}$ is the sum of $\mathbf{B}$ and $\mathbf{W}$,

$$
\mathbf{T} = \mathbf{B} + \mathbf{W},
$$

therefore

$$
\mathrm{tr}(\mathbf{T}) = \mathrm{tr}(\mathbf{B}) + \mathrm{tr}(\mathbf{W}),
$$

or

$$
SST = SSB + SSW.
$$

From this equality, it follows that minimization of the trace of $\mathbf{W}$ is equivalent to maximizing the trace of $\mathbf{B}$. In other words, minimizing the sum of within-cluster squared error also maximizes the sum of between-clusters error.

Example. To show intermediate calculations, consider a dataset that has two dimensions ($P = 2$), and the k-th cluster (C^k) has 3 data points ($N_k = 3$).

$$\mathbf{X}^{(\mathbf{k})} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} \text{ is the data matrix for cluster } C_k$$

And the mean of cluster C_k is:

$$\bar{\mathbf{x}}^{(\mathbf{k})} = \begin{bmatrix} \bar{x}_1^{(k)} & \bar{x}_2^{(k)} \end{bmatrix}$$

For the moment, we will omit the k in the notation: (maybe I'd better not.)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix}$$

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 & \bar{x}_2 \end{bmatrix}$$

Now, the distance matrix $\mathbf{D}$ is obtained by:

$$\mathbf{D}_{3\times 2} = \mathbf{X}_{3\times 2} - \begin{bmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{x}} \\ \bar{\mathbf{x}} \end{bmatrix}_{3\times 2} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 \\ \bar{x}_1 & \bar{x}_2 \\ \bar{x}_1 & \bar{x}_2 \end{bmatrix}$$

$$= \begin{bmatrix} (x_{11} - \bar{x}_1) & (x_{12} - \bar{x}_2) \\ (x_{21} - \bar{x}_1) & (x_{22} - \bar{x}_2) \\ (x_{31} - \bar{x}_1) & (x_{32} - \bar{x}_2) \end{bmatrix}$$

And $\mathbf{D}'\mathbf{D}$ is...

$$= \begin{bmatrix} (x_{11} - \bar{x}_1) & (x_{21} - \bar{x}_1) & (x_{31} - \bar{x}_1) \\ (x_{12} - \bar{x}_2) & (x_{22} - \bar{x}_2) & (x_{32} - \bar{x}_2) \end{bmatrix} * \begin{bmatrix} (x_{11} - \bar{x}_1) & (x_{12} - \bar{x}_2) \\ (x_{21} - \bar{x}_1) & (x_{22} - \bar{x}_2) \\ (x_{31} - \bar{x}_1) & (x_{32} - \bar{x}_2) \end{bmatrix}$$

The (i, j)-th element of $\mathbf{D}'\mathbf{D}$ obtained by multiplying the i-th row of $\mathbf{D}'$ by the j-th column of $\mathbf{D}$. The first element of $\mathbf{D}'\mathbf{D}$, $(i = 1, j = 1)$, is...

$$\mathbf{D}'\mathbf{D}(1,1) = \begin{bmatrix} (x_{11}-\bar{x}_1) & (x_{21}-\bar{x}_1) & (x_{31}-\bar{x}_1) \end{bmatrix} * \begin{bmatrix} (x_{11}-\bar{x}_1) \\ (x_{21}-\bar{x}_1) \\ (x_{31}-\bar{x}_1) \end{bmatrix}$$

$$= \begin{bmatrix} [(x_{11}-\bar{x}_1)(x_{11}-\bar{x}_1)] + [(x_{21}-\bar{x}_1)(x_{21}-\bar{x}_1)] + (x_{31}-\bar{x}_1)(x_{31}-\bar{x}_1)] \end{bmatrix}$$

$$= \begin{bmatrix} [(x_{11}-\bar{x}_1)^2 + (x_{21}-\bar{x}_1)^2 + (x_{31}-\bar{x}_1)^2] \end{bmatrix}$$

This is simply the sum of squared deviations for the first variable:

$$\mathbf{D}'\mathbf{D}(1,1) = \begin{bmatrix} \sum_{i=1}^{3}(x_{i1}-\bar{x}_1)^2 \end{bmatrix}$$

The second element of $\mathbf{D}'\mathbf{D}$, $(i = 1, j = 2)$, is obtained by multiplying the first row of $\mathbf{D}'$ by the second column of $\mathbf{D}$:

$$\mathbf{D}'\mathbf{D}(1,2) = \begin{bmatrix} (x_{11}-\bar{x}_1) & (x_{21}-\bar{x}_1) & (x_{31}-\bar{x}_1) \end{bmatrix} * \begin{bmatrix} (x_{12}-\bar{x}_2) \\ (x_{22}-\bar{x}_2) \\ (x_{32}-\bar{x}_2) \end{bmatrix}$$

$$= \begin{bmatrix} [(x_{11}-\bar{x}_1)(x_{12}-\bar{x}_2)] + [(x_{21}-\bar{x}_1)(x_{22}-\bar{x}_2)] + (x_{31}-\bar{x}_1)(x_{32}-\bar{x}_2)] \end{bmatrix}$$

Thus, the diagonal of $\mathbf{D}'\mathbf{D}$ contains the sum of squared error (within) for each variable, and the non-diagonal elements (where $i \neq j$) contain sums of cross-products for the i-th and j-th variable.

$$\mathbf{W}^{(\mathbf{k})} = \begin{bmatrix} \sum_{i=1}^{3}(x_{i1}-\bar{x}_1)^2 & \sum_{i=1}^{3}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2) \\ \sum_{i=1}^{3}(x_{i2}-\bar{x}_2)(x_{i1}-\bar{x}_1) & \sum_{i=1}^{3}(x_{i2}-\bar{x}_2)^2 \end{bmatrix}$$

2.2.2 K-means algorithms

Multiple formulations of the K-means algorithm have been suggested, and these can be divided into two main types: *individual-reassignment* algorithms, and *batch* algorithms.

Individual-reassignment algorithm

In individual reassignment algorithms for K-means, each data point is considered for reassignment individually, and cluster means are updated after each individual is considered. Algorithms of this type were described by MacQueen (1967), McRae (1971), and Hartigan (1975). The basic algorithm is:

1. Create an initial partition of the data and compute SSW (Equation (2.2)) for the partition.
2. For $i = 1$ to N:
   (a) For the i-th individual, calculate the change in the SSW that would occur if the individual is moved from its current cluster to each cluster to which it does not currently belong.
   (b) Move the individual to the cluster for which a move decreases SSW the most (if there is such a move).
   (c) Update SSW to be the SSW resulting from the move in 2(b).
3. Repeat steps 2(a) through 2(c) for every individual in the dataset: this is called one "pass" through the data.
4. Repeat step 3 (a pass through the whole data set) until one pass has been completed with no individual reassignments (i.e. convergence).

The individual-reassignment K-means algorithm has also been called the incremental K-means algorithm (Kogan, 2007, chap. 2).

Individual reassignment formula. In Step 2(a), data point $\mathbf{x_i}$ in cluster C_r is considered for membership in each cluster C_t; where $t = 1, 2, \ldots K$, $(t \neq r)$. The step calls for the move that most decreases the clustering criterion (SSW), if such a move exists.

By algebraic manipulation, an "updating formula" can be obtained for the change in SSW with a move of data point $\mathbf{x_i}$ from cluster C_r to cluster C_j that uses only the current cluster means ($\bar{\mathbf{x}}_\mathbf{r}$ and $\bar{\mathbf{x}}_\mathbf{t}$), the current number of data points in each cluster (n_r and n_t), and the data point $\mathbf{x_i}$ being evaluated for reassignment. This increases the computational efficiency of the algorithm.

The following expression is evaluated for each possible move for data point $\mathbf{x_i}$ (Späth, 1985):

$$\frac{n_r}{n_r - 1} \|\bar{\mathbf{x}}_\mathbf{r} - \mathbf{x_i}\|^2 > \frac{n_t}{n_t + 1} \|\bar{\mathbf{x}}_\mathbf{t} - \mathbf{x_i}\|^2 \tag{2.7}$$

SSW_r is the within-cluster sum of squares of C_r, the cluster to which data point $\mathbf{x_i}$ currently belongs. SSW_t is the within-cluster sum of squares of C_t, the cluster to which data point $\mathbf{x_i}$ does not currently belong. The

left side of Expression (2.7) is the amount that SSW_r will decrease if data point $\mathbf{x_i}$ is removed from cluster C_r. The right side of the expression is the amount that SSW_t will increase if data point $\mathbf{x_i}$ is moved to cluster C_t.

If Expression (2.7) is true for cluster C_t, this means that the total error sum of squares will decrease if data point $\mathbf{x_i}$ moves from its current cluster (C_r) to cluster C_t. If expression (2.7) is true for at least one C_t $(t \neq r)$, then data point $\mathbf{x_i}$ is moved to the cluster for which the right side of the expression is smallest, i.e. for which the decrease in the SSW will be the greatest.

Batch algorithm

In batch K-means clustering, distances are computed for all data points (individuals) to all cluster means at one time, then all data points are simultaneously moved to the cluster with the nearest mean. This is repeated until no individuals change cluster membership (Kogan, 2007, chap. 2). The basic algorithm is

1. Choose initial cluster means, often called "seeds." (For example, randomly select K data points.)
2. Calculate the squared Euclidean distance from all data points to each cluster mean.
3. Simultaneously assign each individual to the cluster with the nearest mean. (In other words, all individuals move at one time.)
4. Repeat steps (2) and (3) until no individuals are reassigned (i.e. convergence).

The batch algorithm was described by Anderberg (1973), Bock (1970, 1974), and Späth (1980, 1985). Späth refers to the batch K-means algorithm as "H-means" and the individual-reassignment K-means algorithm as "K-means" (Späth, 1980, 1985; Brusco and Steinley, 2007).

Combined algorithm

In Späth's HK-means algorithm, the batch algorithm is performed until convergence, then the single-reassignment algorithm is performed until convergence (Späth, 1985):

1. Perform batch K-means algorithm until convergence.
2. Starting with the partition created by batch K-means, perform individual-reassignment K-means algorithm until convergence.

This algorithm is often presented as a variant of K-means (e.g. Steinley, 2006a). Brusco and Steinley found that the combined algorithm outperformed both the individual-reassignment algorithm and the batch algorithm (Brusco and Steinley, 2007).

Common usage and algorithm performance

Some authors have proposed versions of the K-means algorithm that included "forcing passes" meant to overcome local optima, but these algorithms are seldom used (Friedman and Rubin, 1967). Brusco and Steinley found that the combined algorithm outperformed both the individual-reassignment algorithm and the batch algorithm (Brusco and Steinley, 2007).

2.2.3 Extensions to other metric spaces

In his seminal paper on K-means clustering, MacQueen described the extension of K-means to metric spaces other than Euclidean (MacQueen, 1967) by using different centroids and distances. The generalized optimization criteria would thus be "low average error in some sense." Hartigan also suggested extending K-means to distances other than the Euclidean distance (Hartigan, 1975, chap. 4).

Generalized optimization criterion

Here, the generalized version of the K-means optimization criterion, *SSW* (Equation (2.2)), to other distance measure/centroid combinations is presented. The cluster centroid $\mathbf{c^{(k)}}$ is the attribute-wise centroid of cluster C_k:

$$\mathbf{c^{(k)}} = \begin{bmatrix} c_1 & c_2 & \dots & c_M \end{bmatrix} \tag{2.8}$$

The optimization function $L(\mathbf{X}, \mathscr{C})$ is:

$$L(\mathbf{X}, \mathscr{C}) = \sum_{k=1}^{K} \sum_{i \in C_k} d(\mathbf{x_i}, \mathbf{c^{(k)}}), \tag{2.9}$$

where $d(\mathbf{x_i}, \mathbf{c^{(k)}})$ is the distance from data point $\mathbf{x_i}$ to cluster centroid $\mathbf{c^{(k)}}$, and $\mathscr{C}$ is the collection of cluster centroids:

$$\mathscr{C} = \{\mathbf{c^{(1)}}, \mathbf{c^{(2)}}, \dots, \mathbf{c^{(K)}}\} \tag{2.10}$$

Minkowski distance

The Minkowski distance between data points $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$d_p(\mathbf{x}, \mathbf{y}) \quad = \quad \left(\sum_{m=1}^{M} |x_m - y_m|^p \right)^{\frac{1}{p}} \tag{2.11}$$

The Euclidean distance, also called the L_2-norm, is the Minkowski distance of order $p = 2$. The Minkowski distance of order $p = 1$ is known as the *cityblock* or *Manhattan* distance, and is the sum of the absolute difference between x_m and y_m over all m (all variables).

2.2.4 K-medians

The *K-medians* clustering method uses the attribute-wise cluster median and cityblock distance, rather than the cluster mean and squared Euclidean distance of K-means. The related clustering criterion, $L(\mathbf{X}, \mathscr{C})$, is the sum of absolute differences from data points to cluster medians,

$$L(\mathbf{X}, \mathscr{C}) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{m=1}^{M} |x_i - c_m^{(k)}|. \tag{2.12}$$

Algorithm. The K-medians algorithm is simply the K-means algorithm, using cityblock distance and median in place of squared Euclidean distance and mean. Thus, any of the K-means algorithm variations presented (individual-reassignment, batch, or combined) can be used with K-medians.

Median definition. If the number of observations (N) is odd, the sample median is the observation in the middle of an ordered list of observations. If the number of observations is even, the sample median is the mean of the two middle observations (Weiss, 2002). A cluster median is the vector of the cluster-wise medians for each variable (attribute).

For a group of observations with more than one variable, the cluster median is a vector of the group-wise medians for each variable (attribute).

Updating formulae. As in K-means, there is an updating formula for individual reassignment decisions which can be used in place of calculating the full clustering criterion for every possible move of a data point in the single reassignment phase (Späth, 1985).

PAM

Kaufman and Rousseeuw created a K-means like algorithm called PAM (Partitioning Around Medoids) (Kaufman and Rousseeuw, 1987). The authors' definition of medoid is a "representative data point." This is not equivalent to a centroid, which can be a vector of values not found in the data.

2.2.5 K-modes

K-modes is an extension of K-means to categorical data (Huang, 1998). The cluster mode is used as the centroid, and the simple matching distance is used as the distance measure. The simple matching distance, $d_0(\mathbf{x}, \mathbf{y})$, is the total number of attributes for which an individual and a cluster mode have different values, and has a maximum possible value of M (the number of attributes),

$$d_0(\mathbf{x}, \mathbf{y}) = \sum_{m=1}^{M} \delta(x_m, y_m) \tag{2.13}$$

where

$$\delta(x_m, y_m) = \begin{cases} 0 & (x_m = y_m) \\ 1 & (x_m \neq y_m) \end{cases} . \tag{2.14}$$

Huang's K-modes method is an individual-reassignment algorithm, but like other extensions of K-means, K-modes can theoretically be formulated with any of the K-means algorithms.

Mode definition. The sample mode is the value that occurs most frequently in the data set (Weiss, 2002). A cluster mode is the vector of the cluster-wise modes for each variable (attribute).

2.2.6 Jaccard distance K-modes

K-modes can be extended by using measures other than the simple matching distance. A presence/absence-based dissimilarity coefficient is a measure of dissimilarity between two data points based on four quantities (Table 2.1): a_{11} denotes the number of attributes present in both data points, a_{10} and a_{01} denote the number of attributes present in one data point but not the other, and a_{00} denotes the number of attributes absent in both data points. Many dissimilarity coefficients have been proposed; the reader is referred to Gower and LeGendre (1986) or Baulieu (1989) for reviews. Here, two dissimilarity measures are described, the Hamming distance or simple matching distance and the Jaccard distance.

Table 2.1: Quantities used to define presence/absence-based dissimilarity coefficients.

		y		
		1	0	
x	1	a_{11}	a_{10}	$a_{11}+a_{10}$
	0	a_{01}	a_{00}	$a_{01}+a_{00}$
		$a_{11}+a_{01}$	$a_{10}+a_{00}$	M

Hamming distance and simple matching distance. For categorical data, where each attribute can have two or more categories, the simple matching coefficient is the number of attributes with a different value in **x** and **y**, Equation 2.13. For binary data, the simple matching distance is $(a_{10}+a_{01})$ in Table 2.1, and is equivalent to the Hamming distance, or the number of non-matching bits. Both the simple matching distance and the Hamming distance are sometimes defined as the number of non-matching attributes divided by the total number of attributes,

$$d_0(\mathbf{x}, \mathbf{y}) = \frac{a_{10}+a_{01}}{a_{11}+a_{10}+a_{01}+a_{00}} \tag{2.15}$$

(e.g. Anderberg, 1973; Leisch, 2006).

The simple matching distance is a "symmetric" distance coefficient, meaning that a shared presence of

attributes and a shared absence of attributes contribute equally to the calculation (Leisch, 2006).

Jaccard distance. The numerator of the Jaccard distance is the number of mismatched attributes, and the denominator is the sum of the number of mismatches and the number of attributes present in both data points (e.g. Anderberg, 1973):

$$d_J(\mathbf{x},\mathbf{y}) = \frac{a_{10}+a_{01}}{a_{11}+a_{10}+a_{01}}. \tag{2.16}$$

The maximum value of the Jaccard distance is 1.

The Jaccard distance is an asymmetric distance measure, meaning that it gives more weight to shared presence of attributes than to shared absence of attributes (Leisch, 2006).

The Jaccard similarity coefficient ($1-d_J(\mathbf{x},\mathbf{y})$) has been used a fair amount as a similarity measure for hierarchical cluster analysis methods (e.g. Goodfellow and Pirouz, 1982; Iruela *et al.*, 2002). However, to our knowledge, the Jaccard distance (or the similarity coefficient) has rarely been used for cluster analysis in the clinical literature, as noted by Deutsch et al (Deutsch *et al.*, 2006). To implement K-modes with the Jaccard distance, Equation (2.16) is used as the distance measure and the cluster mode as the centroid in the generalized K-centroids optimization function (Equation (2.9)). Leisch provided a matrix formulation and `R` package for use of the Jaccard distance with K-centroids, but the author(s) has not found any application of Jaccard distance K-modes by research groups other than that of Leisch (Leisch, 2006).

2.2.7 K-centroids and binary data

Binary data considerations

Binary data, also called presence-absence data, have several qualities that should be considered when performing cluster analysis.

Equivalence of distance measures. For binary data, the squared Euclidean distance (L_2-norm), the cityblock distance (L_1-norm), and the simple matching distance are identical. As an example, consider two data points $\mathbf{x}$ and $\mathbf{y}$, observed for five attributes:

$$\mathbf{x} = [\ 1 \quad 1 \quad 0 \quad 0 \quad 0\ ]$$

$$\mathbf{y} = [\ 1 \quad 0 \quad 1 \quad 0 \quad 1\ ]$$

The squared Euclidean distance between **x** and **y** is:

$$\begin{aligned} d_2(\mathbf{x},\mathbf{y}) &= (1-1)^2+(1-0)^2+(0-1)^2+(0-0)^2+(0-1)^2 \\ &= 0+1+1+0+1=3 \end{aligned}$$

The cityblock distance is:

$$\begin{aligned} d_1(\mathbf{x},\mathbf{y}) &= |1-1|+|1-0|+|0-1|+|0-0|+|0-1| \\ &= 0+1+1+0+1=3 \end{aligned}$$

The simple matching distance, $d_0(\mathbf{x},\mathbf{y})$, is also 3, for three mismatched (non-agreeing) attribute values.

Equivalence of centroids. Another property of binary data is that the median is equivalent to the mode. Consider seven observations on one binary variable, ordered from lowest value to highest value:

$$0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1$$

When ordered from lowest value to highest value, the median value will always be the most common value, which is the mode.

2.2.8 Symmetric and asymmetric binary variables

Symmetric binary variables are variables for which both values have equal conceptual meaning or importance; for example, gender (Leisch, 2006). One would probably consider that two male individuals have the same similarity to each other as do two females (for that variable). *Asymmetric* binary variables are variables for which one value has more conceptual importance; for example, disease attributes. One might consider two individuals who share a disease attribute (i.e. who both have a symptom) to be more similar than two individuals who do not share the attribute.

The simple matching distance is considered to be a symmetric distance coefficient, meaning that a shared presence of attributes and a shared absence of attributes contribute equally to the calculation, whereas the Jaccard distance is an asymmetric distance measure, meaning that it gives more weight to shared presence of attributes than to shared absence of attributes (Leisch, 2006). Conventional wisdom holds that the distance measure used for clustering should be "appropriate" for the type of binary data being considered. For ex-

ample, Späth suggested that K-medians for binary data (which will be called K-modes in this manuscript) is "appropriate for use whenever zero and one have the same significance for the aim of the cluster dissection or analysis" (Späth, 1985).

2.2.9 General K-centroids considerations

It has long been recognized that the K-means algorithms reviewed here converge to local optima that are not the global optima (MacQueen, 1967); (see Selim and Ismail (Selim and Ismail, 1984) and Steinley (Steinley, 2003) for studies). Many authors suggest repeating the K-means algorithm a large number of times (using the same K), with a new random selection of initial seeds for each repetition, or a random partition of the data to form initial clusters (e.g. MacQueen, 1967); Steinley suggests a minimum of 5,000 repetitions as a rough guideline (Steinley, 2003). The partition generated that has the lowest optimization criterion value is chosen as the best solution, or the "optimal partition."

K-centroids clustering methods require pre-specification of the number of clusters. However, when the purpose of using cluster analysis is to uncover previously undescribed subtypes (i.e. class discovery), the number of subtypes is not known. A commonly-used strategy in practical applications of K-means is to perform the clustering algorithm for a range of K, then to use some index of partition validity to choose the best partition (e.g. Jones *et al.*, 2002).

2.3 Choosing the best partition

In K-centroids cluster analysis, the number of clusters K is pre-specified. Unfortunately, the measure of clustering quality that is optimized by K-centroids (the *SSW* for K-means and the *SAD* for K-modes) cannot be directly compared between partitions with different K. This is because as the number of clusters in the partition increases, the value of the sum of deviations decreases or stays the same (i.e. has a monotonic property, Späth, 1985). Choosing the best partition from a set of partitions generated by K-centroids for different values of K is often described as "choosing the best K," or the "right number of clusters," etc. ().

In this manuscript, I consider the "best partition" to be the partition that best agrees with the underlying subtype structure of the data, even if K of this partition is different from the "true" number of classes. (Measurement of "cluster recovery," or "agreement," will be discussed in a later section.) This parallels the situation of class discovery in real data, when the underlying structure of the data is unknown.

When performing these clustering methods on real data, a commonly used strategy is to perform the clustering algorithm for a range of K, typically from $K = 2$ to double the number of what is "expected," then to use some index of partition validity to choose the "best" partition.

2.3.1 Cluster validation overview

The term *cluster validation* is used to describe the process of evaluating the "goodness" of clusterings of data produced by cluster analysis or other class discovery methods (Dubes and Jain, 1979; Halkidi *et al.*, 2001). Cluster validation can be measured using three types of criteria: (1) criteria that compare a clustering of the data to known labels of the data (i.e. known subtypes), (2) criteria that measure the "goodness" of a clustering without a comparison to known labels, and (3) criteria that compare two or more clusterings of the data to each other.

Criteria that compare a clustering of the data to known labels are usually called *external criteria* (Halkidi *et al.*, 2001; Handl *et al.*, 2005; Xu and Wunsch, 2009). Monte Carlo analyses of cluster analysis techniques use external criteria to compare clusterings of the data given by clustering methods to the "known" clusterings (i.e. labels) of simulated data. External criteria can also be used to compare a clustering of data based on attributes of one type to a classification of the data based on other attributes, or to describe the agreement between two different classification systems. For example, a clustering of patients using genetic attributes can be compared to labels given by traditional histologic classification. In the context of class discovery, external criteria can be used to describe the agreement between two different partitions of the data.

The term *internal criteria* is sometimes used to describe any criterion that measures the goodness of a clustering without comparison to external labels Some authors use the term in a more narrow sense, to describe methods that assign a statistical probability to a clustering of the data, and the term *relative criteria* to describe criteria meant to be used for comparison of several partitions of the data (e.g. Jain *et al.*, 1999; Halkidi *et al.*, 2001; Gan *et al.*, 2007).

2.3.2 Relative validity indices for choosing the best partition

The following relative validity indices (criteria) can be used to choose the best partition from among a set of partitions generated by the K-centroids algorithm. These indices fall into two categories: quantities designed specifically for use in choosing the best partition from a set of partitions, and clustering criteria that have been adapted as relative validity indices.

Notation, terminology. $\mathscr{U}_K$ denotes a partition of the data with K clusters, and I_K denotes an index measure for partition $\mathscr{U}_K$. *SSW* is the within-cluster sum of squared error, which is equivalent to the trace of $\mathbf{W}$, and *SSB* is the sum of between-cluster error. For the present discussion, it is assumed that there is one partition of the data for each K under consideration.

Usage of an index

Some relative validity criteria are designed to "control for" the number of clusters in a partition, so that the absolute minimum or absolute maximum indicates the best number of clusters. However, it is not always clear that a proposed index does work best in the supposed manner. In the case of clustering criteria used as relative validity indices, there is often a history of various recommended usages, such as elbow sharpess. Therefore, performance studies of these types of indices often assess several usages of an index: the absolute maximum or minimum, the absolute difference to the left or right, or the second difference (also called the "elbow sharpness"). The earlier works on which this manuscript is based considered the maximum or minimum of the following quantities as possible best usages for the indices. Where an index value for the Kth partition is I_K,

i) The index value itself for the kth partition, I_k

ii) The difference to the left: $I_k - I_{k-1}$

iii) The difference to the right: $I_k - I_{k+1}$

iv) The "second difference," or the elbow sharpness: $(I_K - I_{K+1}) - (I_K - I_{K-1})$

(e.g. Milligan and Cooper, 1985).

For each index listed below, usage of the index suggested by the authors that proposed the index (when available), the usage reported as best by Milligan and Cooper (1985), and the usage reported as best by Dimitriadou et al. (2002) are presented.

Indices based on compactness and separation

Indices based on *SSW* and/or *SSB*

1. *SSW* (Equation 2.2), or trace($\mathbf{W}$) (Equation 2.4).

   SSW is the optimization criterion for the traditional K-means clustering algorithm. Edwards and Cavalli-Sforza (1965) suggested using the minimum *SSW* as an index of the best level of a dendrogram created by divisive hierarchical clustering using the *SSW*, but did not provide an example of its use for this purpose. Friedman and Rubin (1967) explored using the maximum decrease in *SSW* as an indicator of the number of clusters. Milligan and Cooper (1985) used the maximum difference of the index, Dimitriadou et al. (2002) used the maximum second difference of the index.

2. (Ball and Hall, 1967)

(a) Milligan and Cooper (1985) described the Ball and Hall index as "the average distance of the items to their respective cluster centroids," and used the maximum difference of the index.

(b) $\frac{SSW}{K}$

Dimitriadou et al. (2002) presented this as the Ball and Hall index, and used the maximum second difference of the index.

3. $D\log\left(\sqrt{\frac{SSW}{DN^2}}\right)+\log(K)$, where D is the number of attributes (Xu, 1997).

Xu (1997) designed the index to control for the number of clusters, so that the minimum value indicates the best value for K. Dimitriadou et al. (2002) used the maximum second difference of the index.

4. $\frac{SSB/(K-1)}{SSW/(N-K)}$, (Calinski and Harabasz, 1974).

Milligan and Cooper (1985) used the maximum value of the index, Dimitriadou et al. (2002) used the minimum second difference of the index.

5. $log(\frac{SSB}{SSW})$, (Hartigan, 1975).

Milligan and Cooper (1985) used the maximum difference of the index, Dimitriadou et al. (2002) used the maximum second difference of the index.

6. (Ratkowsky and Lance, 1978)

(a) mean$(\sqrt{\frac{SSB_m}{SST_m}})/\sqrt{K}$, where SSB_m is the SSB for the m-th attribute, SST_m is the SST for the m-th attribute (Ratkowsky and Lance, 1978; see Hill, 1980).

Ratkowsky and Lance intended the maximum value of the index to indicate the best number of clusters; Milligan and Cooper (1985) used the maximum value of the index. (Note: Like the Ball and Hall index, this seems counter-intuitive.)

(b) mean$(\sqrt{\frac{SSB_m}{SST_m}})$, where SSB_m is the SSB for the m-th attribute, SST_m is the SST for the m-th attribute (Ratkowsky and Lance, 1978; see Hill, 1980).

This is the form of the index presented by Dimitriadou et al. (2002), who used the maximum difference to the partition at the right side.

Davies-Bouldin index. The Davies-Bouldin index measures the mean "similarity" between each cluster and its "most similar" cluster in a partition (Davies and Bouldin, 1979). The similarity, R_{ij}, between clusters C_i and C_j is a ratio of the sum of the dispersion (compactness) of each cluster to the separation of the clusters:

$$R_{ij}=\frac{S_i+S_j}{M_{ij}}, \tag{2.17}$$

where S_i is the dispersion of cluster C_i, S_j is the dispersion of cluster C_j, and M_{ij} is the distance between the "characteristic vectors" of the clusters.

The Davies-Bouldin index is defined as

7.
$$\bar{R} = \frac{1}{K}\sum_{i=1}^{K} R_i, \tag{2.18}$$

where R_i is the similarity of cluster C_i with its most similar cluster.

The form of Davies-Bouldin index used by Milligan and Cooper (1985) and Dimitriadou et al. (2002) uses *SSW* as the cluster dispersion measure and squared Euclidean distance of cluster means as the cluster separation measure,

$$R_{ij} = \frac{SSW_i + SSW_j}{d_2(\bar{\mathbf{x}}^{(\mathbf{i})}, \bar{\mathbf{x}}^{(\mathbf{j})})}. \tag{2.19}$$

Davies and Bouldin designed the index so that the minimum value indicates the best number of clusters. Milligan and Cooper (1985) and Dimitriadou et al. (2002) used the minimum value of the index.

We also considered a version of the Davies-Boulding index using *SAD* as the cluster dispersion measure and simple matching distance of cluster means as the cluster separation measure.

Indices related to $|\mathbf{W}|$.

Friedman and Rubin (1967) proposed using the ratio of the determinants of $\mathbf{T}$ and $\mathbf{W}$, $\frac{|\mathbf{T}|}{|\mathbf{W}|}$, as an optimization criterion for partition-based cluster analysis. Although the following indices were originally presented for use in conjunction with partitions resulting from optimizing $\frac{|\mathbf{T}|}{|\mathbf{W}|}$, some studies of relative indices have considered their use in other situations (Milligan and Cooper, 1985; Dimitriadou *et al.*, 2002). (Similarly, Friedman and Rubin explored use of the *SSW* in determining the number of clusters when $\frac{|\mathbf{T}|}{|\mathbf{W}|}$ had been optimized to form partitions.)

8. $\frac{|\mathbf{T}|}{|\mathbf{W}|}$, (Friedman and Rubin, 1967).

   This measure was proposed by Friedman and Rubin as an optimization criterion for clustering, not as an index for choosing the number of clusters (see next criterion); however, it has been presented as an index for number of clusters. Milligan and Cooper (1985) used the maximum difference of the index, Dimitriadou et al. (2002) used the minimum second difference of the index.

9. $N \log \frac{|\mathbf{T}|}{|\mathbf{W}|}$, (Scott and Symons, 1971).

   Friedman and Rubin proposed using the maximum increase in natural logarithm of $\frac{|\mathbf{T}|}{|\mathbf{W}|}$ as an indicator of the number of clusters (Friedman and Rubin, 1967), and $N \log \frac{|\mathbf{T}|}{|\mathbf{W}|}$ was proposed by Scott and Symons.

We evaluate the latter index, noting that the indices are equivalent in practice because N is constant. Milligan and Cooper (1985) used the maximum difference of the index, Dimitriadou et al. (2002) used the maximum difference to the partition at the left of the index.

10. $K^2|\mathbf{W}|$, (Marriott, 1971).

Marriott proposed using the minimum value of this index to indicate the number of clusters. Milligan and Cooper (1985) used the maximum difference of the index, Dimitriadou et al. (2002) used the maximum second difference of the index.

2.4 Measuring agreement between two partitions

According to Rand, there are two important aspects of cluster analysis methods: (1) How "easy" a method is (this has to do with how much time it will take to perform an analysis, which is not a trivial issue); and (2) How well does a method perform? (Rand, 1971). In this section, I will review Hubert and Arabie's adjusted Rand index, a measure that is often used to evaluate the performance of a cluster analysis method and in other settings.

A popular approach for evaluating the performance of a cluster analysis method (or to compare the performance of several cluster analysis methods) is to create a collection of artificial data sets, which each have a "known" partition. Then, for each artificial data set, a "measure of agreement" is used to evaluate how well the partition derived from applying the cluster analysis method in question agrees with the "known" partition of the data.

Such measures of agreement could also be used in the analysis of real data, when there is not a "known" partition of the data. (In other words, when there are no so-called "true" or "natural" clusters, or such clusters if they exist are unknown.) For example, the partitions might arise from using different cluster analysis methods on the objects. Or, the two partitions might arise from applying one cluster analysis method to different variables—for instance, the first partition could be formed by performing cluster analysis on laboratory-based variables, and the second partition could be formed by performing cluster analysis on only (physical manifestations). Another example is partitions that arise when two people, or "raters," are asked to divide patients into groups.

Overview. Many methods for measuring agreement between two partitions of a data set have been proposed, and I will not attempt an exhaustive review. I will focus on providing background for a popular measure of agreement between two partitions, Hubert and Arabie's adjusted Rand index (ARI) (Hubert and Arabie, 1985).

2.4.1 Notation for two partitions

Suppose that there is data set (i.e. group, or population) of N patients, and two different partitions for the data, where each partition is a collection of clusters (also called classes). We would like to evaluate how well the two partitions "agree" with each other. I will use notation employed by Hubert and Arabie (Hubert and Arabie, 1985), which is similar to the notation often used in the literature on this subject.

Figure 2.1 is a diagram of set of patients and two partitions of the group. The set has six patients, represented by the numbers 1 through 6 in the vertical box. Partition U has two clusters, with patients 1 through 4 in the first cluster and patients 5 and 6 in the second cluster. Partition V has two clusters, with patients 1 through 3 in the first cluster and patients 4 through 6 in the second cluster.

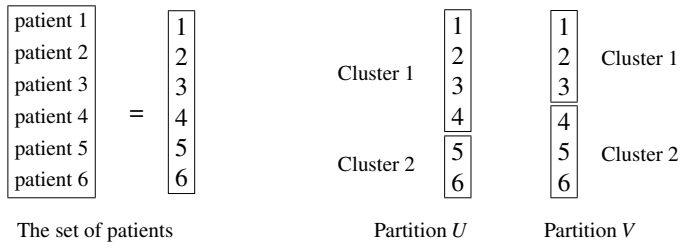

Figure 2.1: Partitions U and V.

Contingency table. One way to compare two partitions is with a contingency table:

Table 2.2: Contingency table for Partitions $\mathscr{U}$ and $\mathscr{V}$

	$\mathscr{V}$				
$\mathscr{U}$	V_1	V_2	$\ldots$	$V_{K'}$	Total
U_1	n_{11}	n_{12}	$\ldots$	n_{11}	$n_{1\cdot}$
U_2	n_{21}	n_{22}	$\ldots$	n_{11}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
U_K	n_{K1}	n_{K2}	$\ldots$	$n_{KK'}$	$n_{K\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot K'}$	n

Partition $\mathscr{U}$ divides the objects into K clusters, $U_1, U_2, \ldots, U_K$. Partition $\mathscr{V}$ divides the objects into K' clusters, $V_1, V_2, \ldots, V_{K'}$.

2.4.2 Measures based on cross-classification

Goodman and Kruskal's λ

Goodman and Kruskal's λ is based on "optimal prediction." The asymmetric version of λ measures the "decrease in probability of error in guessing" an individual's class membership in Partition $\mathscr{V}$ when their

membership in Partition $\mathscr{U}$ is known (Goodman and Kruskal, 1954).

Goodman and Kruskal formulated the problem in terms of proportions, and we will follow their notation. Table 2.3 shows the cross tabulation of Partitions $\mathscr{U}$ and $\mathscr{V}$ with proportions rather than counts.

Table 2.3: Contingency table for Partitions $\mathscr{U}$ and $\mathscr{V}$, with proportions.

	$\mathscr{V}$				
$\mathscr{U}$	V_1	V_2	$\cdots$	$V_{K'}$	Total
U_1	p_{11}	p_{12}	$\cdots$	p_{11}	$p_{1\cdot}$
U_2	p_{21}	p_{22}	$\cdots$	p_{11}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
U_K	p_{K1}	p_{K2}	$\cdots$	$p_{KK'}$	$p_{K\cdot}$
Total	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	$p_{\cdot K'}$	1

Asymmetric λ

In the asymmetric versions of Goodman and Kruskal's λ, λ_b and λ_a, the ability to predict one classification (i.e., partition) based on the other classification is measured.

Consider two cases in which we will guess an individual's V-class: (1) we have no information about the individual's U-class, (2) we know the individual's U-class. In case (1), our best guess for the individual's V-class is the most common class in Partition $\mathscr{V}$, or the B-class with the largest marginal proportion $p_{\cdot b}$. The probability of making a correct guess is $p_{\cdot m}$, where $p_{\cdot m}$ is the largest $p_{\cdot b}$, and the probability of error (an incorrect guess) is $(1 - p_{\cdot m})$. In case 2, where we know an individual's A-class (A_a), the best guess for their B class is the B-class with the highest conditional probability, p_{ab}. Then, the probability of making the correct assignment of B-class given that an individual is in class A_a is p_{am}, where p_{am} is the maximum p_{ab}, and the probability of making an error is $(1 - p_{am})$. The overall probability of making an error in case 2 is $(1 - \sum_a p_{am})$.

Goodman and Kruskal's λ_b is the "relative decrease in probability of error in guessing B_b as between A_a unknown and A_a known":

$$\lambda_b = \frac{\text{(Prob. of error in case 1)} - \text{(Prob. of error in case 2)}}{\text{(Prob. of error in case 1)}} \tag{2.20}$$

$$= \frac{\sum_a p_{am} - p_{\cdot m}}{1 - p_{\cdot m}} \tag{2.21}$$

Properties of λ_b (verbatim from Goodman and Kruskal 1954):

1. λ_b is indeterminate iff the population lies in one column, i.e. there is only one B class.

2. Otherwise the value of λ_b lies between 0 and 1 inclusive.

3. λ_b is 0 iff knowledge of the A classification is of no help in predicting the B classification, i.e., if there exists a b_0 such that $p_{ab_0} = p_{am}$ for all a.

4. λ_b is 1 iff knowledge of an individual's A class completely specifies his B class, i.e., if each row of the cross-classification table contains at most one nonzero p_{ab}.

5. In the case of statistical independence λ_b, when determinate, is zero. The converse need not hold: λ_b may be zero without statistical independence holding.

6. λ_b is unchanged by permutation of rows or columns.

λ_a denotes the parallel measure, or the reduction in error for guessing an individual's A-class membership that can be gained through knowledge of the B classification.

$$\lambda_a = \frac{\sum_b p_{bm} - p_{\cdot m}}{1 - p_{\cdot m}}$$

Symmetric λ

Goodman and Kruskal defined the symmetric λ for the situation in which we are interested in jointly describing the ability of two classifications to predict each other, formulated as "half of the time, we are guessing an individual's A class; half of the time, we are guessing an individual's B class." The probability of case 1 error in this situation is $1 - \frac{1}{2}(p_{\cdot m} + p_{m\cdot})$, and the probability of case 2 error in this situation is $1 - \frac{1}{2}(\sum_a p_{am} + \sum_b p_{bm})$. The symmetric coefficient λ is defined as

$$\lambda = \frac{\frac{1}{2}\left[\sum_a p_{am} + \sum_b p_{bm} - p_{\cdot m} - p_{m\cdot}\right]}{1 - \frac{1}{2}(p_{\cdot m} + p_{m\cdot})}$$

Properties of λ (verbatim from Goodman and Kruskal 1954):

1. λ is determinate except when the entire population lies in a single cell of the table.

2. Otherwise the value of λ lies between 0 and 1 inclusive.

3. λ is 1 iff all the population is concentrated in cells no two of which are in the same row or column.

4. λ is 0 in the case of statistical independence, but the converse need not hold.

5. λ is unchanged by permutation of rows or columns.

6. λ lies between λ_a and λ_b inclusive.

2.4.3 Measures based on matching pairs

Many popular measures of agreement are based on pairs of patients (objects). When there are two partitions for the objects, a given pair of objects is one of the four types in Table 2.4. Often, the lower-case letters a, b, c, and d are used to describe the total number of pairs of types (i), (ii), (iii), and (iv) respectively.

Table 2.4: Four types of object pairs.

(i)	The pair are in the same cluster as each other in Partition $\mathscr{U}$ and the same cluster in Partition $\mathscr{V}$.
(ii)	The pair are in different clusters in Partition $\mathscr{U}$ but the same cluster in Partition $\mathscr{V}$.
(iii)	The pair are in the same cluster in Partition $\mathscr{U}$, but different clusters in Partition $\mathscr{V}$.
(iv)	The pair are in different clusters in Partition $\mathscr{U}$ and different clusters in Partition $\mathscr{V}$.

Figure 2.2 demonstrates the four types of pairs using the example data set depicted in Figure 2.1. Patient 1 and Patient 2 are in the same cluster as each other (U_1) in Partition $\mathscr{U}$, and they are in the same cluster as each other (V_1) in Partition $\mathscr{V}$, so they are type (i). Patients 4 and 5 are in different clusters from each other in Partition $\mathscr{U}$, but they are in the same cluster as each other (V_2) in Partition $\mathscr{V}$, so they are type (ii). Patients 3 and 4 are in the same cluster as each other (U_1) in Partition $\mathscr{U}$, but in different clusters from each other in Partition $\mathscr{V}$, so they are type (iii). Patients 3 and 5 are in different clusters from each other in both Partition $\mathscr{U}$ and Partition $\mathscr{V}$, so they are type (iv). (These are only four of a possible 15 patient pairs.)

Patient 1 and Patient 2 are in the same cluster as each other (Cluster 1) in Partition U, and they are in the same cluster as each other (Cluster 1) in Partition V, so they are type (i). Patients 3 and 5 are in different clusters from each other in both Partition U and Partition V, so they are type (ii). Patients 4 and 5 are in different clusters from each other in Partition U, but they are in the same cluster as each other (Cluster 2) in Partition V, so they are type (iii). Finally, Patients 3 and 4 are in the same cluster as each other (Cluster 1) in Partition U, but in different clusters from each other in Partition V, so they are type (iv). (These are only four of a possible 15 patient pairs.)

Unordered object pairs. Measures of agreement based on object pairs often assume that we do not care about the order of objects within a pair, and in such case, the object pairs are sometimes referred to as *unordered object pairs* for the sake of clarity. For any N objects, there are $\binom{N}{2}$ (i.e. N choose 2) possible unordered object pairs, and this quantity can be easily conceptualized:

$$\binom{N}{2} = (N-1) + (N-2) + \ldots + 1.$$

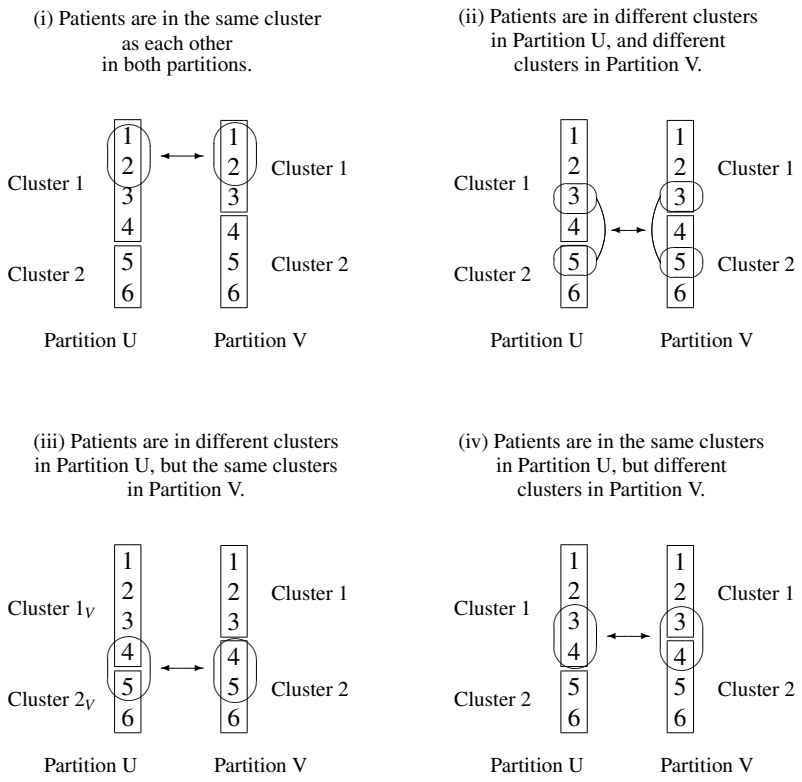

Figure 2.2: Object pair types (i), (ii), (iii) and (iv).

For our example, $N = 6$, so:

$$\binom{6}{2} = 5+4+3+2+1 = 15.$$

In our example, there are 15 unordered object pairs.

Example. Because our example uses a very small data set, we can identify the pairs of each object type by inspecting the pictorial representation of the two partitions. Table 2.5 shows which patient (object) pairs fall into each of the four types. The pair of Patients 1 and 2 is denoted as [1&2], etc.

Table 2.5: Patient (object) pairs by type.

type	*patient pairs*	*total number of type*
i	[1&2], [1&3], [2&3], [5&6]	$a = 4$
ii	[1&5], [1&6], [2&5], [2&6], [3&5], [3&6]	$d = 6$
ii	[1&4], [2&4], [3&4]	$b = 3$
iii	[4&5], [4&6]	$c = 2$

To check our work (i.e., that we have considered every pair), we note that $a+b+c+d = 15$, which we calculated as the total number of unordered object pairs in the data set.

Contingency table formulations

The quantities a, b, c, and d can be computed using formula based on elements of the contingency table in Table 2.2 (Hubert and Arabie, 1985, p. 196). These formulae are shown in Table 2.6.

In this notation, n_{ij} denotes the number of objects (not object pairs) in class U_i of Partition $\mathscr{U}$ and the class V_j of partition $\mathscr{V}$. For example, n_{12} in the contingency table represents the number of objects (not object pairs!) that are in class 1 (U_1) of Partition U and class 2 (V_2) of Partition V. The marginal totals are denoted by $n_{i\cdot}$ for row totals and $n_{\cdot j}$ for column totals. For example, $n_{2\cdot}$ is the total number of objects in the 2nd row, and $n_{\cdot 3}$ is the total number of objects in the 3rd column.

Table 2.6: Formulae for the number of object pairs of the four types.

Pair type	Number of pairs	Formula
(i)	a	$\frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{C} n_{ij}(n_{ij}-1)$
(ii)	d	$\frac{1}{2}(n^2+\sum_{i=1}^{K}\sum_{j=1}^{C} n_{ij}^2-(\sum_{i=1}^{K} n_{i\cdot}^2+\sum_{j=1}^{K'} n_{\cdot j}^2))$
(iii)	c	$\frac{1}{2}(\sum_{j=1}^{K'} n_{\cdot j}^2-\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2)$
(iv)	b	$\frac{1}{2}(\sum_{i=1}^{K} n_{i\cdot}^2-\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2)$

Although these formulae look complicated, they can be broken into just a few mathematical expressions. The symbol $\sum_{i=1}^{R}\sum_{j=1}^{C}$ simply means to sum the following quantity over each cell of the table. So, to find the number of pairs of type (i), the count (number of objects) in each cell (n_{ij}) is multiplied by one less than that count ($n_{ij}-1$), then all of these quantities are added; the total is divided by 2. The formulae for types (ii), (iii), and (iv) use the expression $\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2$. To find this expression, each cell count is squared, then the sum of all of the squared cell counts is taken. The expression $\sum_{j=1}^{C} n_{\cdot}^2{}_j$ means the sum of the square of each *marginal* count for each column (i.e. the total count for each column), and the expression $\sum_{i=1}^{R} n_{i\cdot}^2$ is the sum of the square of each row's marginal count.

In Table 2.7, the previously described example is used to demonstrate this notation. The far right column of the table, under "Sums," gives the marginal totals for Cluster 1 and Cluster 2 of Partition U, which are 3 and 3. (This is the number of objects in each cluster.) The bottom row of the table gives the marginal totals

for Cluster 1 and Cluster 2 of Partition V, which are 4 and 2. The upper left cell gives the number of patients that are in Cluster U_1 and Cluster V_1, which is 3 (Patients 1, 2, and 3 are in both clusters.) Patient 4 is in Cluster U_2 and Cluster V_1, so the cell in the table for these two clusters is 1.

Table 2.7: Contingency table for Partitions $\mathscr{U}$ and $\mathscr{V}$

	Partition $\mathscr{V}$		
Partition $\mathscr{U}$	V_1	V_2	Total
U_1	3	1	4
U_2	0	2	2
Total	3	3	6

Now, we can use the formulae from Table 2.6 to calculate the number of object pairs of each type. K is the number of classes for Partition U, which in our example is 2. So, i can be 1 or 2. K' is the number of classes for Partition V, which is also 2 in our example; so, j also can be either 1 or 2.

$$\begin{aligned} a &= \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1) = \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}(n_{ij}-1) \\ &= \tfrac{1}{2}\Big[3(3-1)+1(1-1)+0(0-1)+2(2-1)\Big] \\ &= \tfrac{1}{2}(8) = 4 \end{aligned}$$

$$\begin{aligned} d &= \frac{1}{2}\left(n^2+\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \left(\sum_{i=1}^{K} n_{i\cdot}^2 + \sum_{j=1}^{K'} n_{\cdot j}^2\right)\right) \\ &= \tfrac{1}{2}\Big(6^2+\left(3^2+1^2+0^2+2^2\right)-\left([4^2+2^2]+[3^2+3^2]\right)\Big) \\ &= \tfrac{1}{2}\Big(36+14-(20+18)\Big) \\ &= \tfrac{1}{2}(12) = 6 \end{aligned}$$

The summations to calculate b and c can be taken from previous work.

$$\begin{aligned} c &= \frac{1}{2}\left(\sum_{j=1}^{K'} n_{\cdot j}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right) \\ &= \frac{1}{2}(18-14) = 2 \end{aligned}$$

$$\begin{aligned} b &= \frac{1}{2}(\sum_{i=1}^{K} n_{i\cdot}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2) \\ &= \frac{1}{2}(20-14) = 3 \end{aligned}$$

Computational formulae. The formulae that have been shown for a, b, c, and d can be rewritten into formulae that are computationally easier, and these computational formulae are often seen in the literature. (I have omitted them here because they are less intuitive than what I have shown.)

2.4.4 The Rand index and similar indices

The Rand index is simply the total number of pairs in agreement, in other words the quantity $a+d$, divided by the total number of pairs (Rand, 1971). The Rand index can take on values from 0 to 1, with 1 being total agreement between the two partitions.

$$\text{Rand index} = \frac{a+d}{a+b+c+d}$$

Recall the formula for $(a+d)$ that uses elements of the two-way contingency table:

$$(a+d) = \binom{n}{2} + 2\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_{i}\binom{n_{i\cdot}}{2} + \sum_{j}\binom{n_{\cdot j}}{2}\right]$$

H is commonly used in formulae in place of $\binom{n}{2}$, (which also equals $(a+b+c+d)$. So, we have the following (algebraic manipulations are shown in Appendix [A]):

$$\begin{aligned} \text{Rand index} &= \frac{a+d}{a+b+c+d} = \frac{(a+d)}{\binom{n}{2}} = \frac{(a+d)}{H} \\ &= 1 + \frac{2\sum_{i,j}\binom{n_{ij}}{2}}{H} - \frac{\left[\sum_{i}\binom{n_{i\cdot}}{2} + \sum_{j}\binom{n_{\cdot j}}{2}\right]}{H} \end{aligned}$$

This measure was actually proposed by numerous authors (Hubert and Arabie, 1985), with the earliest publication in 1958 by Sokal and Michener, according to Klastorin (1985). Nevertheless, the measure is usually attributed to Rand's 1971 publication. In the opinion of the author, it is possible that two qualities of the paper contributed to this– Rand's paper was published in a very prominent journal (*Journal of the American Statistical Association*), and the paper is well-written and easily comprehensible. (Furthermore,

Rand did not cite earlier works describing this measure, and perhaps it can be assumed that he had not seen them.)

Similar indices have also been suggested, as reviewed by Steinley (2004) and Hubert and Arabie (1985). Several authors independently suggested $\frac{b+c}{a+b+c+d}$.Other matching-pair based indices were proposed by Fowlkes and Mallows (1983) and Brennan and Light (Brennan and Light, 1974).

The major limitation of the Rand index is that it does not control for chance (Hubert and Arabie, 1985). (This limitation also applies to other measures of agreement based solely on the quantities a, b, c, and d). A test statistic that "controls for chance" will have a constant expected value under a null probabilistic model, in other words under the condition of chance. In contrast, the expected value of the Rand index approaches 1 (Rand, 1971) and the variance approaches 0 as the number of classes in the partitions increase (Fowlkes and Mallows, 1983). Consequently, the raw value of the Rand index has questionable meaning.

The Rand index has great intuitive appeal as a measure of agreement, and several authors desired to find a form of the Rand index adjusted for chance. In the creation of such a test statistic, two questions arise: First, what form should the agreement measure have, and second, what is an appropriate null model for comparing two partitions?

2.4.5 Hubert and Arabie's adjusted Rand index

Hubert and Arabie proposed applying a general form for an index corrected for chance originally proposed by Cohen (Cohen, 1960) to the Rand index (Hubert and Arabie, 1985). This general form is

$$\text{adjusted index} = \frac{\text{observed value of index} - \text{expected value of index}}{\text{maximum value of index} - \text{expected value of index}}$$

The observed value of the Rand index is calculated from the observed data, as demonstrated earlier. The maximum value of the Rand index is 1, which occurs when all object pairs are in "agreement" between the two partitions (i.e., $(a+d) = 1$). The "expected value of index" means the expected value of the index under a certain null probability distribution.

Fixed margins assumption. Hubert and Arabie, and Brennan and Light before them, suggested that it would be appropriate to assume that the margins of the contingency table comparing the two partitions are fixed under the null distribution (Brennan and Light, 1974; Hubert and Arabie, 1985). (This is the same assumption made in many nonparametric situations, for example with Fisher's exact test.) The "fixed margins" assumption means that the null hypothesis covers all possible situations with the same values for the marginal totals (as well as the same number of categories in each margin) as are observed.

The expected Rand index. Under the fixed margins assumption, we have several important results. First, because the margins are "fixed," we can consider the following quantities to be constants: H, $\sum_i \binom{n_{i.}}{2}$, and $\sum_j \binom{n_{.j}}{2}$. (H is $\binom{n}{2}$, the total number of unordered object pairs.).

The remaining element of the Rand index, $\sum_{i,j} \binom{n_{ij}}{2}$, must be treated as a random variable. Under the fixed margins assumption, it can be shown that

$$E\Big[\sum_{i,j}\binom{n_{ij}}{2}\Big] = \frac{\Big[\sum_i \binom{n_{i.}}{2} \times \sum_j \binom{n_{.j}}{2}\Big]}{\binom{n}{2}} = \frac{\Big[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\Big]}{H} \tag{2.22}$$

(Mantel, 1967; Klastorin, 1985; Fowlkes and Mallows, 1983).

From earlier, we have that

$$\begin{aligned} \text{Rand index} &= 1 + \frac{2}{H}\sum_{i,j}\binom{n_{ij}}{2} - \frac{1}{H}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] \\ &= 1 + \frac{2\sum_{i,j}\binom{n_{ij}}{2}}{H} - \frac{\Big[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\Big]}{H} \end{aligned}$$

To find the expected value of the Rand index, first we will recall the general rules of expectation. If X is a random variable and k is a constant, then $E(k) = k$ and $E(kX) = kE(X)$. Applying these rules of expectation and Equation 2.22, the expected Rand index is:

$$E(\text{Rand index}) = 1 + \frac{2\Big[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\Big]}{H^2} - \frac{\Big[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\Big]}{H}$$

The corrected Rand index. We substitute these expressions into the general formula for an index corrected for chance to get Hubert and Arabie's adjusted Rand index. (Recall that the maximum value of the Rand index is 1.)

$$
\begin{aligned}
ARI_{HA} &= \frac{\text{observed Rand index} - \text{expected Rand index}}{\text{maximum Rand index} - \text{expected value of Rand index}} \\
&= \frac{\left[1 + \frac{2\sum_{i,j}\binom{n_{ij}}{2}}{H} - \frac{\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right]}{H}\right] - \left[1 + \frac{2\left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]}{H^2} - \frac{\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right]}{H}\right]}{1 - \left[1 + \frac{2\left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]}{H^2} - \frac{\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right]}{H}\right]}
\end{aligned}
$$

Algebraic manipulations give a less unwieldy form,

$$
ARI_{\mathrm{HA}} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\frac{2\left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]}{H^2}\right]}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\frac{2\left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]}{H^2}\right]}.
$$

Finally, further algebraic manipulations yield a frequently-seen form of the ARI_{HA} that uses H, a, b, c, and d,

$$
ARI_{\mathrm{HA}} = \frac{H(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{H^2 - [(a+b)(a+c)+(c+d)(b+d)]}.
$$

Properties of ARI_{HA}. The maximum value of ARI_{HA} is 1, which occurs when there is complete agreement between the two clusters. The expected value of ARI_{HA} under chance agreement alone is 0. There is no minimum bound for the ARI_{HA}, but Hubert and Arabie believed that this is not a limitation in practice (Hubert and Arabie, 1985).

Steinley (2004) compared the performance of the ARI_{HA} against another possible measure of cluster agreement, classification rates. A classification rate is simply the number of objects correctly allocated with each cluster. In practice, "labels" must be arbitrarily chosen in order to calculate a classification rate, and the labels that are chosen influence the classification rate. In Monte Carlo studies in which the first partition was fixed and the second partition was generated randomly, i.e. under the condition of chance, correct classification rates were distributed approximately uniformly from 30% to 90%. In other words, the classificate rate method suggested approximately 30% to 90% "correct classification" for partitions generated by chance. In contrast, the distribution of the ARI_{HA} was highly skewed towards 0 under chance, indicating that ARI_{HA}

scores above 0 are a good measure of departure from the assumption of chance agreement (Steinley, 2004).

Steinley's study described some other good properties of the ARI_{HA}. The ARI_{HA} was "fairly invariant" in response to different levels of cluster density (the distribution of objects across clusters), the number of objects, and the number of clusters.

CHAPTER 3

Paper 1: K-centroids variations for the identification of disease subtypes with presence/absence attributes

Abstract

Many diseases, for example systemic lupus erythematosus, encompass patients with a wide variety of manifestations. For such diseases, the diversity in manifestations might be the result of several unnamed diseases with overlapping manifestations or of disease subtypes. Therefore, the identification of subsets of patients with homogenous manifestations may be an important step in elucidating the etiology and pathophysiology of disease, hopefully with gain of insight into treatment of the disease. In this paper, we compared the performance of three variations of the K-means clustering algorithm (K-means with the squared Euclidean distance, K-modes with the simple matching distance, and K-modes with the Jaccard distance) on a benchmark data set of erythemato-squamous diseases with overlapping manifestations. The three clustering methods identified different partitions of the benchmark data for several levels of K (number of clusters), and cross-tabulations of partitions suggested that the algorithms have different tendencies in the splitting or combining of true classes. In Monte Carlo simulations with data representing possible disease subtype scenarios, the performances of K-means and K-modes with the simple matching distance were not appreciably affected by differences in sample size or distribution of patients across subtypes.

3.1 Introduction

Many diseases in fields such as rheumatology, neurology, and psychiatry are diagnosed by the recognition of a constellation of symptoms and signs and exhibit wide variation between patients in disease manifestations. Systemic lupus erythematosus (SLE) is a prototypical example of a clinically heterogeneous, "constellation"-diagnosed disease. SLE is an autoimmune disease characterized by the presence of autoantibodies to elements of the cell nucleus, and can cause permanent damage to any organ system. The American College of Rheumatology research classification criteria for SLE classify a patient as having SLE if the patient has four or more of 11 presence/absence clinical and laboratory attributes (e.g. mucosal ulcers, photosensitivity) (Tan *et al.*, 1982). Patients diagnosed with SLE by a physician or classified as having lupus by the accepted research classification criteria have wide inter-patient variety in organ system involvement. For SLE and similar diseases, it is possible that several disease subtypes, or alternatively several diseases with overlapping manifestations, are encompassed by what is currently considered to be one syndrome or disease entity.

Case definition is an important step in epidemiologic investigation (Tyler and Last, 1998). In the investigation of a "new" disease, for example when a groups of individuals present with similar symptoms and signs that are not attributable to a described disease, the shared disease manifestations often become the working case definition. This case definition is used in epidemiologic and laboratory investigations, then is modified as further information is gained. For example, the preliminary case definition for what we now call acquired immunodeficiency syndrome (AIDS) was a list of unusual symptoms and signs shared by a few patients, and was later modified to include presence of the HIV virus (Lasky and Stolley, 1994). The identification of clusters of SLE patients with similar disease manifestations could similarly be used to form working "case definitions" of subtypes. Just as is hoped for when using working case definitions of previously undescribed diseases, investigation of homogeneous patient sub-groups may yield insight into the etiology or treatment of a protean disease.

Refinement of case definitions through exploration of subtypes is a potentially fertile area for collaboration between statisticians, epidemiologists, and clinicians. Morris, in his *Uses of Epidemiology*, stated that the identification of subtypes of a disease is a "use of epidemiology" (Morris, 1955, 1975); however, formal techniques for subtype identification lie outside the realm of statistical methods frequently taught to or used by epidemiologists. Communication with clinicians is important for hypothesis generation, interpretation of subtypes suggested by statistical methods, and insight into the clinical complexities of these diseases.

Our objective was to evaluate the performance of several K-means cluster analysis variations, identifying data set characteristics that influenced the performance of the algorithms. In Study 1, we compared the performance of K-means, K-modes with the simple matching distance, and K-modes with the Jaccard distance

on a benchmark empiric data set of erythemato-squamous diseases with known disease classes. In Study 2, we compared the performance of the clustering methods on simulated data with qualities similar to the benchmark empiric data. In Study 3, we compared the performance of K-means and K-modes with the simple matching distance on simulated data representing patients with disease subtypes or overlapping diseases, with a variety of subtype structure scenarios. Because SLE and other such diseases are often diagnosed or classified using the presence or absence of symptoms and signs, all studies used presence/absence (binary) data.

3.2 Methods

For each study, we used several types of K-centroids cluster analysis to create mutually exclusive clusters in data with known subtype structure, or disease classes. The clusters created by cluster analysis were compared to the known classes to determine ability of the clustering methods to identify the true subtype structure.

K-centroids cluster analysis

K-centroids cluster analysis is a family of methods that place objects in a data set into clusters in such a way as to minimize the sum of distances from objects to cluster centroids. In K-means cluster analysis, the sum of squared Euclidean distances from objects to cluster means is minimized; in K-modes cluster analysis, the sum of distances of objects to cluster modes is minimized. We used K-modes with two distances: (a) the simple matching distance (the number of mismatched attributes, or attributes with different values in the object and cluster mode), and (b) the Jaccard distance (the number of mismatched attributes divided by sum of the number of non-matching attributes and the number of attributes present in both the object and the cluster mode).

We used the combined K-means algorithm described by Späth (1985) and recommended by Brusco and Steinley (2007) to implement K-centroids cluster analysis. Briefly, K data points were chosen randomly from the data to serve as initial cluster centroids. Batch reassignments were performed to move objects to the nearest cluster centroid until no objects are moved, then individual reassignments were performed. The algorithm was repeated 10,000 times, with a new random selection of cluster seeds for each replication, for each clustering method and level of K. Out of the 10,000 partitions created, the partition with the lowest value of the clustering criterion, $L(\mathbf{X}, \mathscr{C})$, (e.g. for K-means, the SSW) was chosen as the optimal partition for that level of K by that clustering method.

3.2.1 Study 1: Erythemato-squamous diseases empiric data

In our first study, we used an empiric data set of patients with erythemato-squamous diseases from the UCI Machine Learning Repository (Asuncion and Newman, 2007). This data is considered to be "benchmark"

data because the disease classes are known. We applied K-means, standard K-modes, and K-modes with the Jaccard distance to the clinical and histologic attributes in the data. Then, we compared partitions generated by the K-centroids methods to the known disease classes.

The data consist of clinical and histologic attributes, listed in Table 3.1, for patients with six diseases: psoriasis, seborrheic dermatitis, lichen planus, pityriasis rosea, chronic dermatitis, and pityriasis rubra pilaris (Guvenir *et al.*, 1998). These diseases are difficult to distinguish clinically because of overlapping symptoms; diagnosis is made by recognition of a constellation of symptoms and signs, and in some cases there is a histologic attribute that is pathognomic for the disease. In this data, diagnoses were made by specialist clinicians.

K-means, K-modes with the simple matching distance (standard K-modes), and K-modes with the Jaccard distance were performed on the empiric data for $K = 3, \ldots, 9$.

3.2.2 Study 2: Simulated erythemato-squamous data

Simulation studies are a widely-used tool in the evaluation of class discovery methods (Milligan, 1981b). In our first simulation study, we generated data from three scenarios closely related to the empiric erythemato-squamous data. All simulations used the total number of patients ($N = 366$) in the empiric data and the observed number of patients in each disease. Attributes in these simulations were conditionally independent, meaning that intra-class correlations were specified to be 0. Some scenarios were designed to explore the possibility that two subclasses of Disease 1, psoriasis, are represented in the empiric data set used in Study 1.

For each simulated data set from Scenario A, we generated the 366 artificial data points (patients) from 6 diseases, using the observed disease-specific conditional probabilities for the presence of attributes. For data sets from Scenarios B and C, patients with Diseases 2 through 6 were generated from the observed conditional probabilities of attributes for Diseases 2 through 6 in the empiric data. For Scenario B data sets, the Disease 1 class was designed to have an equal number of patients (56) from two "subtypes," roughly representing the two clusters that contained all Disease 1 patients in the K-means seven-cluster partition. Patients of subtype 1.a. were generated from the conditional probabilities in cluster 1 from this partition, and patients of subtype 1.b. were generated from the conditional probabilities of cluster 2 from the same partition. In Scenario C, Disease 1 patients were again split into two subtypes, with the conditional probabilities for each subtype taken from the two clusters that the K-modes algorithm created from Disease 1 patients (for $K = 5, 6$, and 7).

We explored two class labeling scenarios for simulations from Scenarios B and C. First, we labeled patients from Disease 1 subtypes 1.a. and 1.b. as belonging to a single disease class (Disease 1), as they would have been in the empiric data. (In other words, we labeled as if we had no knowledge of the Disease 1 subtypes). In our second labeling scenario, we labeled patients from each Disease 1 subtype as belonging to

a distinct disease class.

K-means, standard K-modes, and Jaccard distance K-modes were performed on each simulated data set for $K = 3, \ldots, 9$.

3.2.3 Study 3: Factorial simulations

In Study 3, we further evaluated K-means and standard K-modes by conducting simulations with a factorial study design. The data set design factors were number of individuals (i.e. patients, objects) in the dataset, underlying subtype structure (subtype structures varied in overlap), relative distribution of individuals across subtypes, the probability of being present for "high probability" attributes, and conditional (within-subtype) dependence of the "high probability" attributes. These design factors were fully crossed, creating 108 cells, and eight datasets were created for each cell (for a total of 864 datasets).

Each clustering algorithm (K-means, K-modes) was performed for $K = 2, 3, \ldots 9$ on each artificial data set, yielding a set of eight partitions of the data set by each algorithm. Recovery of true subtype structure for each of the eight partitions in a set was measured by ARI. The best ARI of these eight was used as an "observation."

3.2.4 Data set design factors

We conceptualized a "subtype" as a group of patients with a high probability of having each of a certain subset of possible disease manifestations; each subtype was defined as a vector of "high" and "low" probabilities for a set of presence/absence (binary) attributes.

Subtype structures. We used presence/absence attributes, representing disease manifestations, to define three subtype structures with different levels of overlap between subtypes, depicted in the Appendix. The three structures were: no overlap in high-probability variables between subtypes, overlap of one variable, and overlap of two variables. Each subtype structure had three subtypes.

Number of individuals. We used two levels for number of individuals N (i.e. patients) in a dataset, 200 and 1,000.

Distribution of individuals across subtypes. We used two levels for the distribution of individuals across subtypes: (1) even distribution across subtype (same proportion in each subtype), and (2) one large subtype with roughly one-half of the individuals, with the remaining subtypes having equal proportions (50%-25%-25%).

The relative density of the clusters directly could represent either "true" relative density, in other words a relative density that would occur in a random sample of the underlying population, or a relative density that would occur from biased sampling of different subtypes. In many studies of SLE, a high proportion of patients have renal disease. It may be that if there is a renal-SLE subtype, that subtype is either a large

proportion of patients with SLE, or that SLE patients with renal manifestations are more likely to receive medical attention and thus be included in studies of SLE.

High probability. We used three levels of probability of presence for the "high probability," attributes: 0.9, 0.8, and 0.7. For all simulated data, the probability of a "low-probability" attribute being present was 0.1.

Intra-subtype dependence. We defined intra-subtype (conditional) dependence as dependence of the high probability attributes within that subtype. We used three levels of intra-subtype dependence: $\rho = 0$, 0.4, and 0.8.

3.2.5 Evaluation of cluster analysis method performance

Measuring performance

We used Hubert and Arabie's adjusted Rand index (ARI) to measure recovery of true disease classes (Hubert and Arabie, 1985; Steinley, 2004). The ARI range is from -1 to 1, with 1 indicating perfect agreement between two partitions and 0 indicating chance agreement.

Study 3 analyses

Study 3 was designed so that the effect of data set design factors on clustering method performance could be ascertained. The analytical data set consisted of 1,728 "observations," each representing the best partition of a simulated data set by one of the clustering methods ($1728 = 864$ simulated datasets $\times$ 2 clustering methods). The independent variables were clustering algorithm and data set design factors.

We used a mixed linear model to determine the effect of clustering method and data set design factors on best ARI. Clustering method and data set design factors were considered fixed effects, and data set was entered into the model as a random effect to account for dependence of repeated observations (for K-means and K-modes) on the same data set. We evaluated two-way interaction terms between design factors and clustering algorithm to determine if the clustering algorithms performed differently under certain data design conditions.

Computational methods

All data generation and performance of clustering algorithms were conducted in MATLAB version 2008b (MathWorks, 2008). K-means and K-modes were performed using the MATLAB statistics toolbox `kmeans.m` function (K-modes by choosing the 'Hamming' option), and K-modes with Jaccard distance was performed with a routine written by the author that utilizes Leisch's matrix formulation (Leisch, 2006). Statistical analyses were conducted with SAS/STAT software version 9.1.2 (SAS, 2009) Artificial data were generated using MATLAB routines written by the author. Emrich and Piedmonte's method was used to generate dependent

multivariate binary data (Emrich and Piedmonte, 1991).

3.3 Results

3.3.1 Study 1 results

Figure 3.1 shows the ARI (maximum value, 1.000) of the best partition found by each K-centroids method, by number of clusters (K). The highest ARI achieved was 0.966, for the K-means six-cluster ($K = 6$) partition, followed by 0.919 for the Jaccard distance K-modes six-cluster partition. The best six-cluster partition found by standard K-modes had worse cluster recovery than the other six-cluster partitions, with an ARI of 0.659. Similarly, for the five-cluster partitions, the K-means and Jaccard distance K-modes partitions had much better cluster recovery (ARI $= 0.866$ and 0.876, respectively) than the standard K-modes partition (ARI $= 0.583$). In contrast, all three seven-cluster partitions had similar cluster recovery, with ARI of 0.801, 0.782, and 0.799 for K-means, Jaccard distance K-modes, and standard K-modes.

Cross-tabulations of partitions with disease classification showed that for the partitions for $K = 2, 3$, and 4, which had similar ARI for each clustering method, the methods "combined" diseases into one cluster in the same manner. For example, all three $K = 4$ partitions (for the three clustering algorithms) had one cluster with the majority of Disease 1 patients, one cluster with the majority of Disease 3 patients, one cluster with the majority of Disease 5 patients, and one cluster with nearly all patients with Diseases 2, 4, and 6. However, two partitions with similar ARI's did not always have similar combinations or splits of disease classes. For example, all three partitions for $K = 8$ had similar ARI's, but K-means split disease classes in a different manner from both K-modes clustering methods: the K-means partition splits patient with Disease 1 into three clusters (with the remaining clusters composed of mostly one disease each), whereas both K-modes algorithms split patients from Disease 1 into 2 clusters and patients from Disease 5 into two clusters (with the remaining clusters composed of mostly one disease each).

The most drastic differences in ARI for given levels of K were observed for $K = 5$ and $K = 6$. In Table 3.2, cross-tabulations of the five-cluster and six-cluster partitions for the three clustering methods with true disease classification are displayed to illustrate how disease classes were differently split or combined in these partitions. The standard K-modes partition for $K = 5$ splits Disease 1 into two clusters and combines Diseases 2, 4, and 6 into one cluster, while the K-means partition does not split Disease 1, and combines Diseases 2 and 4 (not with Disease 6). In the Jaccard K-modes partitions for $K = 5$, patients with Disease 6 were spread among four of the clusters. The clusters from both the K-means and Jaccard distance K-modes six-cluster partitions mostly coincided with the true disease classification of the patients. In contrast, in the six-cluster standard K-modes partition, Disease 1 was split into two clusters (clusters C_1 and C_2), and Diseases 2 and 4 were combined into one cluster (Cluster C_3).

Clusters 1 and 2 from the K-modes partitions for $K = 5$, $K = 6$, and $K = 7$ have exactly the same members, respectively. All patients with Disease 1 (psoriasis) are in one of these two clusters, and Cluster 2 includes one patient with Disease 6. Most attributes have similar conditional probabilities in Cluster 1 and Cluster 2 (i.e. within 0.00 to 0.10 of each other). Five attributes have conditional probabilities that differ by a somewhat larger amount (all differ by more than 0.20): itching, Koebner phenomenon, hyperkeratosis, disappearance of the granular layer, and vacuolization and damage of the basal layer. Table 3 shows the conditional attribute probabilities for these two clusters, and the attribute probabilities if these clusters were grouped into one cluster.

3.3.2 Study 2 results

In Figure 3.2, mean ARI is plotted against K for the three types of simulations. For simulations from Scenario A, the highest mean ARI was achieved by the six-cluster partitions for all three algorithms, with K-means and Jaccard distance K-modes (mean ARI of 0.962 and 0.910, respectively) outperforming standard K-modes (mean ARI of 0.798). The largest difference between the mean ARI of the two partitions was found for the $K = 6$.

For simulations from Scenario B, the mean ARI of all algorithms was similar for $K = 4$ partitions. The mean ARI of the K-modes $K = 5$ and $K = 6$ partitions (0.601 and 0.682) was drastically lower than that of the K-means (0.866 and 0.831) and Jaccard distance K-modes partitions (0.873 and 0.883). Inspection of cross-tabulations of disease classes with clusters revealed that for $K = 5$, the algorithms had made similar "choices" on each simulated data set as with the empiric data: all standard K-modes partitions split Disease 1 into two clusters, while K-means and Jaccard distance K-modes did not. For $K = 6$, K-modes split Disease 1 for all simulated data sets, while K-means and Jaccard distance K-modes split Disease 1 for five (half) or less of the simulated data sets. As K increased beyond the number of designed disease classes, members of disease classes were often "split" differently by the different algorithms. This was true for all disease classes, including those generated from a single probability distribution (Diseases 2 through 6). For example, the nine-cluster K-means partitions for eight of the data sets split Disease 3 into two clusters , and none of these partitions split Disease 5 into two clusters. In contrast, eight of the nine-cluster Jaccard K-modes partitions split Disease 5 into two clusters, and none of these split Disease 3 into two clusters.

The algorithms had similar performance simulations from Scenario C, with the main difference being that the mean ARI for four-cluster K-modes partitions (0.621) was much lower than for K-means and Jaccard distance K-modes (0.782, 0.776). Cross-tabulations showed that for four of the ten simulated data sets, the four-cluster K-modes partition split Disease 1 into 2 clusters, keeping Diseases 2, 4, 5, and 6 in one cluster.

Figure 3.3 shows the true subtype recovery for Scenario B and C simulations when patients from Disease 1

subtypes 1.a. and 1.b. were labeled as two different diseases. With the alternate disease class labels, there was no drastic divergence in algorithm performance for certain numbers of clusters, and K-means outperformed the other algorithms.

We note that the impression given of the relative performance of the clustering algorithms differs drastically according to how patients of the two Disease 1 subtypes are labeled, even though the data (and hence the partitions generated by the algorithms) have not changed.

3.3.3 Study 3 Results

Table 3.4 gives the mean adjusted Rand index (ARI) by each data set design factor level, stratified by clustering algorithm. On average, K-modes had slightly better cluster recovery than K-means, with an average ARI of 0.757 compared to 0.742. Both algorithms performed worse on average as the level of high-probability attributes decreased, as intra-subtype dependence increased, and as the overlap between subtypes increased. K-means performed slightly worse for an uneven distribution of patients, with a mean ARI of 0.749 for data sets with evenly distributed patients and 0.736 for data sets with one subtype containing half of the patients. K-modes performed almost identically for both levels of patient distribution, with mean ARIs of 0.757 and 0.756 respectively. Both algorithms performed nearly as well for smaller data sets (200 patients) as they did for larger data sets (1,000 patients).

Table 3.5 gives the regression coefficients from the mixed linear model. The interaction effect estimate for K-modes with an uneven distribution of patients across subtypes (0.012, $p < 0.001$) and the main effect estimate for uneven distribution ($^{-}0.013$) reflect the observation from Table 4 that K-means performed slightly worse for an uneven distribution of patients compared to an even distribution, while K-modes did not. However, the meaningfulness of this observation is questionable due to the low magnitude of effect.

The regression coefficients of the interaction terms for K-modes with data set design factor levels that cause poorer ARI on average (high-probability of 0.8 and 0.7, intra-subtype dependence, and overlapping attributes) were all positive, reflecting that K-modes performed better than K-means under these conditions. However, the magnitude of even the largest interaction effect estimates was very small compared to the effect estimates for the factors themselves. The regression coefficient for high probability level of 0.7 was $^{-}0.281$, indicating a decrease in ARI of 0.281 relative to data sets with high-probability level 0.9, for all observations (K-means and K-modes). The interaction term for K-modes was highly significant ($p < 0.001$), but the magnitude of the interaction effect estimate (0.009) is very small compared to the main effect of $H = 0.7$. The interaction terms for K-modes with $\rho = 0.4$ and $\rho = 0.8$ were also both highly significant ($p < 0.001$), but with a magnitude (0.014 and 0.009 respectively) that was very small compared to the main effect of these levels on ARI ($^{-}0.136$ for $\rho = 0.4$ and $^{-}0.206$ for $\rho = 0.8$). The interaction terms for K-modes with attribute

overlap were positive (0.004 for both 1 and 2 attributes overlapping), but again, very small compared to the main effect (and in this case, not statistically significant).

3.4 Discussion

In our first study, we compared the performance of three variations of the K-means clustering method, K-means, K-modes with the simple matching distance (standard K-modes), and K-modes with the Jaccard distance, on a benchmark empiric data set of erythemato-squamous diseases. K-means and K-modes with Jaccard distance performed better than standard K-modes at recovering the true disease classes in the data, as measured by Hubert and Arabie's adjusted Rand index (ARI). However, simulations suggested that the worse performance of K-modes on the benchmark data set may largely reflect tendencies of the methods to group or split disease classes differently in conjunction with the presence of unlabeled classes.

In our third study, we compared the performance of K-means and standard K-modes in a factorial simulation study. The algorithms did not perform appreciably differently for the conditions studied. Both algorithms performed worse in the presence of conditional dependence, overlap in high-probability attributes between subtypes, and the probability of presence for high-probability attributes. Number of patients and distribution of patients across subtypes did not meaningfully affect algorithm performance.

The use of generalized versions of K-means has long been suggested by authorities on cluster analysis, and in fact was described in what is often considered the seminal work on K-means (MacQueen, 1967). However, review of the applied cluster analysis literature reveals a vastly greater number of K-means applications than other variations of K-centroids. In our study of an empiric benchmark data set, three K-centroids variations sometimes created very similar partitions and sometimes created different partitions of presence/absence data for a given number of clusters. These different partitions of the data might be useful for hypothesis generation; therefore, we suggest that K-centroids variations should be thought of as complementary methods that may produce multiple partitions of interest, rather than as a set of methods from which one must be selected.

We demonstrated the use of the Jaccard distance with K-modes, which has rarely been reported. We wish to make several comments on the use of K-modes with the Jaccard distance. Conventional wisdom has long held that dissimilarity distances which give more weight to the shared presence of attributes than to the shared absence of attributes, such as the Jaccard distance, should be used in situations with so-called "asymmetric" data (Späth, 1985; Leisch, 2006). Although K-modes with the Jaccard distance performed better than standard K-modes on the benchmark data set for several levels of K, we feel it is possible that these differences largely reflect tendencies of the algorithms to split or combine disease (or subtype) classes with different priorities, rather than a true difference in performance. Furthermore, our simulations suggested

that unlabeled sub-classes might account for much of the observed disparity in performance. We conclude that K-modes with Jaccard distance is not necessarily more suitable than standard K-modes for so-called "asymmetric" data.

We also note that use of the Jaccard distance with a single-reassignment K-centroids algorithm is orders of magnitude slower than either K-means or standard K-modes, due to the absence of a simple "updating" formula for the Jaccard distance. This makes use of the Jaccard distance in large-scale simulation studies impractical at the present time, and also limits the use of K-modes with Jaccard distance in computationally intensive cluster validation techniques such as the gap statistic (Tibshirani *et al.*, 2001).

Our study results illustrate several drawbacks and benefits of using simulated data and benchmark empiric data. The primary benefits of using simulated data are that the structure of the data is completely specified, and a variety of combinations of structural characteristics can be explored (Milligan, 1996). The user decides what probability distribution defines each class or subtype in the data, how many classes there are, and so on. Importantly, data characteristics that are "unknowable" for empiric data with unknown class structure, such as intra-subtype dependence (conditional dependence), can be specified in simulated data. A major drawback of simulated data is that it is difficult to imagine or anticipate all of the complexities that might be encountered in a given set of real data. Therefore, while simulation studies are useful in evaluating the robustness of statistical methods and in identifying particular data characteristics that adversely affect the performance of cluster analysis methods, it is unwise to make generalizations about the performance of methods from a single simulation study (Milligan, 1996). Had we not applied the clustering algorithms to the benchmark data set in addition to the factorial simulation study, we might have erroneously concluded that K-means and K-modes tend to perform similarly.

A benefit of using a benchmark data set with known classes, such as the erythemato-squamous disease empiric data used in our Study 1, is that the data may have complex characteristics resembling the characteristics of data to which we will apply the investigated methods better than does simulated data created with abstract "possible data characteristics," which might reveal differences in clustering method performance. For example, we found that the three clustering algorithms identified dissimilar partitions of the empiric data for several levels of K.

The performance of the clustering algorithms on the erythemato-squamous empiric data also illustrate one of the potential drawbacks of using benchmark empiric data with "known" classes. If the classes are based on expert opinion, as they were in this case, there is no guarantee that the classes represent a "true" structure in the data, or the "best" classification of the data. For example, if the experts had labeled the psoriasis patients as two psoriasis subtypes and grouped pityriasis rosea and pityriasis rubra pilaris together, K-modes would have performed better (as measured by ARI) than K-means at identifying the disease classes. In our simulated

data based closely on possible subtype structures in the empiric data, changing the "true disease class" labels without changing the data or the partitions drastically increased the perceived performance of K-modes. It is ironic, and perhaps troublesome, that classifications based on expert opinion of benchmark data are often blindly accepted as the gold standard or "true" structure, when the purpose of a study is to evaluate cluster analysis techniques that we hope will perform better than expert opinion!

Choice of the best partitions from a set of partitions generated for different K or by different algorithms is an important aspect of using K-centroids methods (Everitt, 1979; Steinley, 2006a). This topic is beyond the scope of the present investigation. However, we point out that the issue of how to choose the best partition is unresolved in the cluster analysis literature.

This study was limited to K-centroids variations and did not consider other cluster analysis or class discovery methods that could be applied to class discovery for presence/absence data, such as latent class analysis. This choice was made based on the desire to focus on what we felt were interesting extensions of the K-means algorithm. Future studies could explore data characteristics that may have contributed to the varying performances of the clustering algorithms on the benchmark data set.

In conclusion, we found that K-means, standard K-modes, and K-modes with Jaccard distance identified different partitions of an empiric benchmark data set with presence/absence disease attributes, suggesting that application of more than one of these clustering methods to empiric data may be useful in identifying partitions or clusters with different characteristics. Our results did not provide convincing evidence to support the idea that the Jaccard distance is generally preferable to the simple matching distance for asymmetric presence/absence data. We illustrated that in the presence of sub-classes in a "true" class, alternate labeling of the sub-classes (as one class vs. as sub-classes) can drastically affect perception of the performance of clustering methods.

3.5 Figures and tables.

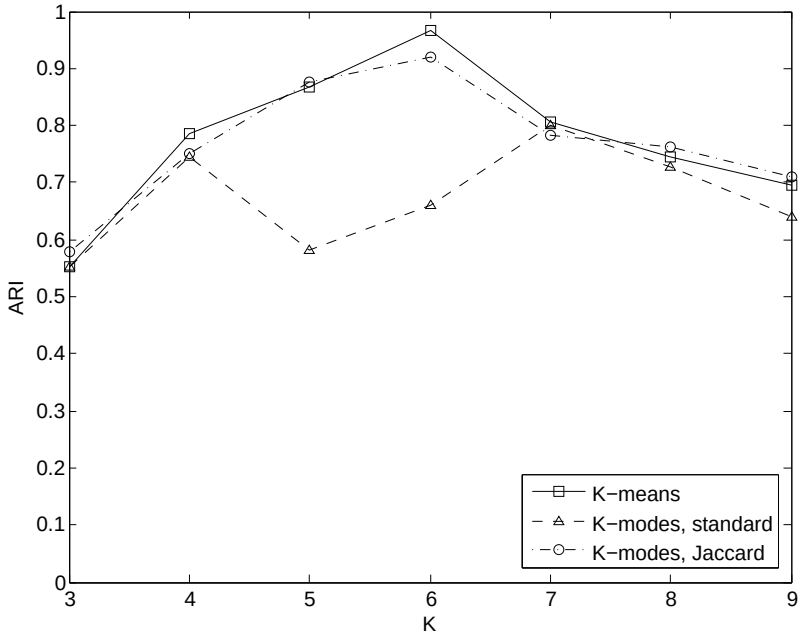

Figure 3.1: ARI by K and clustering algorithm for partitions of the empiric data set (Study 1).

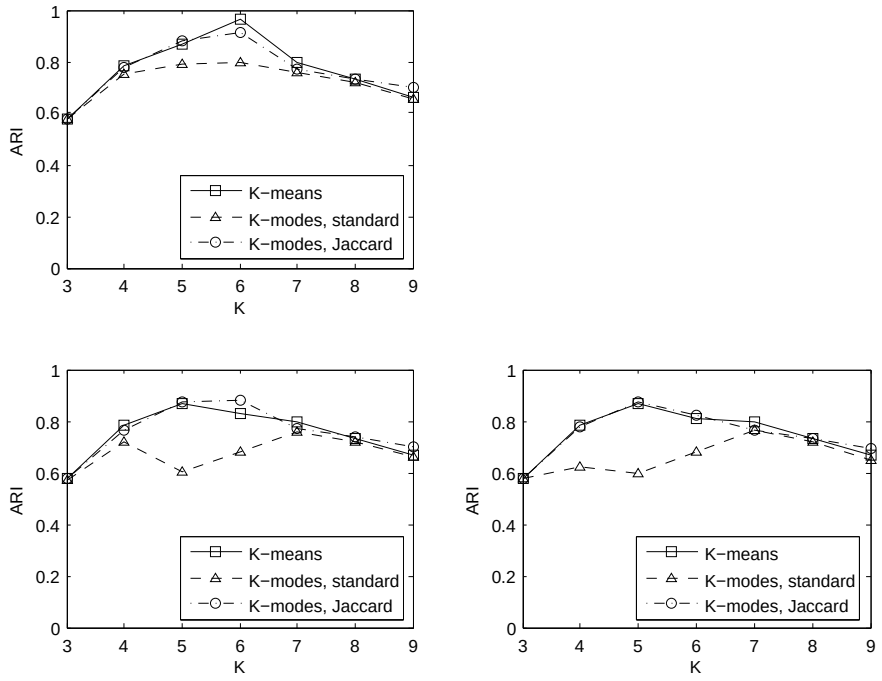

Figure 3.2: ARI by K and clustering algorithm for Study 2 simulations from Scenarios A (upper left), B (lower left), and C (lower right).

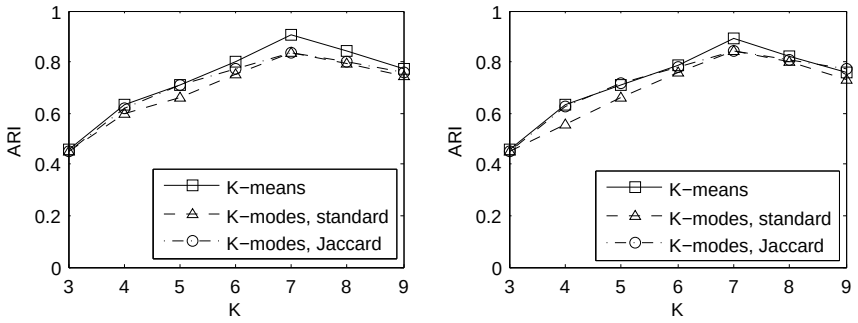

Figure 3.3: ARI by K and clustering algorithm for Study 2 simulations from Scenarios B (left) and C (right), labeling Disease 1 subtypes as distinct classes.

Table 3.1: Attributes in erythemato-squamous benchmark data set.

Clinical attributes		Histologic attributes	
1	erythema	12	melanin incontinence
2	scaling	13	eosinophils in the infiltrate
3	definite borders	14	PNL infiltrate
4	itching	15	fibrosis of the papillary dermis
5	Koebner phenomenon	16	exocytosis
6	polygonal papules	17	acanthosis
7	follicular papules	18	hyperkeratosis
8	oral mucosal involvement	19	parakeratosis
9	knee and elbow involvement	20	clubbing of the rete ridges
10	scalp involvement	21	elongation of the rete ridges
11	family history	22	thinning of the suprapapillary epidermis
		23	spongiform pustule
		24	Munro microabscess
		25	focal hypergranulosis
		26	disappearance of the granular layer
		27	vacuolisation and damage of basal layer
		28	spongiosis
		29	saw-tooth appearnace of retes
		30	follicular horn plug
		31	perifollicular parakeratosis
		32	inflammatory mononuclear infiltrate
		33	band-like infiltrate

Table 3.2: Cross tabulation of true disease classes with partitions for $K = 5$ and $K = 6$.

		Partitions with $K = 5$					Partitions with $K = 6$					
		K-means partitions										
Disease	n	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5	C_6
1	112	112	0	0	0	0	112	0	0	0	0	0
2	61	0	59	0	1	1	0	59	0	1	1	0
3	72	0	0	72	0	0	0	0	72	0	0	0
4	49	0	49	0	0	0	0	5	0	44	0	0
5	52	0	0	0	52	0	0	0	0	0	52	0
6	20	0	0	0	0	20	0	0	0	0	0	20
Total	366	112	108	72	53	21	112	64	72	45	53	20
		Standard K-modes partitions										
Disease	n	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5	C_6
1	112	60	52	0	0	0	60	52	0	0	0	0
2	61	0	0	60	0	1	0	0	60	0	1	0
3	72	0	0	0	72	0	0	0	0	72	0	0
4	49	0	0	49	0	0	0	0	49	0	0	0
5	52	0	0	7	0	45	0	0	7	0	45	0
6	20	0	1	19	0	0	0	1	0	0	0	19
Total	366	60	53	135	72	46	60	53	116	72	46	19
		Jaccard K-modes partitions										
Disease	n	C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5	C_6
1	112	108	3	0	0	1	108	3	0	0	1	0
2	61	0	59	0	2	0	0	57	0	2	2	0
3	72	0	0	72	0	0	0	0	72	0	0	0
4	49	0	5	0	44	0	0	5	0	44	0	0
5	52	0	0	0	0	52	0	0	0	0	52	0
6	20	1	7	0	7	5	1	0	0	0	0	19
Total	366	109	74	72	53	58	109	65	72	46	55	19

Table 3.3: Selected attributes and conditional probability of being present in clusters C_1, C_2 from K-modes $K = 7$ partition; these clusters mostly comprise Disease 1.

Attribute	C_1	C_2	$C_{1\&2}$
itching	0.32	0.74	0.51
Koebner phenomenon	0.12	0.79	0.43
hyperkeratosis	0.30	0.79	0.53
disappearance of granular layer	0.32	0.74	0.51
vacuolization and damage of basal layer	0.12	0.79	0.43
n	60	53	113

Table 3.4: Mean adjusted Rand index (ARI) by data set design factor levels, for best ARI from all partitions generated by each algorithm (for $K = 2, 3, \ldots 9$).

		mean ARI_{best}	
Factor level	No. of data sets	K-means	standard K-modes
All	864	0.742	0.757
No. of patients			
200	432	0.742	0.755
1000	432	0.743	0.759
Distribution of patients			
$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$	432	0.749	0.757
$\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$	432	0.736	0.756
"high" probability (H)			
0.9	288	0.878	0.889
0.8	288	0.752	0.764
0.7	288	0.597	0.617
Intra-subtype dependence			
$\rho = 0$	288	0.856	0.863
$\rho = 0.4$	288	0.720	0.741
$\rho = 0.8$	288	0.650	0.666
Subtype structure			
No overlap	288	0.782	0.793
1 attribute overlaps	288	0.753	0.769
2 attributes overlap	288	0.692	0.708

Table 3.5: Linear regression of adjusted Rand index (ARI) on data design factors.

Factor level	Response: ARI_{best} $\hat{\beta}$	p-value
K-centroids algorithm		
K-means	(ref)	
K-modes, standard	$^{-}0.007$	0.014
No. of patients		
200	(ref)	
1000	0.001	0.752
Distribution of patients		
$\frac{1}{3},\frac{1}{3},\frac{1}{3}$	(ref)	
$\frac{1}{4},\frac{1}{4},\frac{1}{2}$	$^{-}0.013$	0.001
"high" probability (H)		
0.9	(ref)	
0.8	$^{-}0.125$	< 0.001
0.7	$^{-}0.281$	< 0.001
Intra-subtype dependence (ρ)		
0.0	(ref)	
0.4	$^{-}0.136$	< 0.001
0.8	$^{-}0.206$	< 0.001
Subtype structure		
No overlap	(ref)	
1 attribute overlap	$^{-}0.029$	< 0.001
2 attributes overlap	$^{-}0.090$	< 0.001
Interactions		
K-modes $\times$ ($N = 1000$)	0.003	0.151
K-modes $\times$ (distribution $= \frac{1}{4},\frac{1}{4},\frac{1}{2}$)	0.012	< 0.001
K-modes $\times$ (H $= 0.8$)	0.001	0.581
K-modes $\times$ (H $= 0.7$)	0.009	< 0.001
K-modes $\times$ ($\rho = 0.4$)	0.014	< 0.001
K-modes $\times$ ($\rho = 0.8$)	0.009	< 0.001
K-modes $\times$ (1 attribute overlaps)	0.004	0.091
K-modes $\times$ (2 attributes overlap)	0.004	0.077

Appendix

Subtype structures. We used presence/absence attributes, representing disease manifestations, to define subtype structures. Each subtype was defined as having a high probability for a specified subset of the attributes (these were the "high-probability" or "subtype-defining" attributes) and a low probability for the remaining attributes. Thus, each subtype structure consisted of three probability distributions representing three subtypes.

Structure 1 has no overlap between subtypes, meaning that there is no attribute with a high probability of being present in more than one subtype. In this structure, subtype 2 is defined as having a high probability for attributes E, F, G, and H.

	Attributes, probability of being present											
True subtype	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
Subtype 1	H	H	H	H	-	-	-	-	-	-	-	-
Subtype 2	-	-	-	-	H	H	H	H	-	-	-	-
Subtype 3	-	-	-	-	-	-	-	-	H	H	H	H

Structure 2 has one attribute (D) that overlaps (has a high probability of being present) in subtypes 1 and 2, and one attribute (G) that overlaps in subtypes 2 and 3.

	Attributes, probability of being present									
True subtype	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Subtype 1	H	H	H	H	-	-	-	-	-	-
Subtype 2	-	-	-	H	H	H	H	-	-	-
Subtype 3	-	-	-	-	-	-	H	H	H	H

Structure 3 has two attributes (C and D) that overlap in subtypes 1 and 2, and two attributes (G) that overlap in subtypes 2 and 3.

	Attributes, prob. of being present							
True subtype	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Subtype 1	H	H	H	H	-	-	-	-
Subtype 2	-	-	H	H	H	H	-	-
Subtype 3	-	-	-	-	H	H	H	H

CHAPTER 4

Paper 2: A comparison of relative validity indices for choosing the best partition in presence-absence data

Abstract

Background. Partition-based clustering algorithms such as K-means create a partition with a pre-specified number of clusters (K). For empiric data, which presumably has an unknown number of underlying subgroups, an important step of using partition-based clustering algorithms is to determine the "best K" (number of clusters).

Aims. Our aim was to compare the ability of 14 relative validation indices to choose the best partition of the data.

Methods. We applied the validation indices to K-means and K-modes partitions of an empiric benchmark data set and to simulated data representing possible disease subtype scenarios. The "best partition" of each data set by an algorithm (or "best K") was considered to be the partition with the best recovery of true disease classes, as measured by Hubert and Arabie's adjusted Rand index (ARI). The outcomes were correct choice of the best partition, and true disease class recovery (ARI) of the partition chosen by the index.

Results. Only one of the indices chose the best partitions of the empiric data from K-modes or K-means. Our results also illustrate the pitfalls of using indices based on the determinant of the within-cluster sum of squares and cross products matrix with presence/absence data.

4.1 Introduction

Cluster analysis of presence/absence (binary) data has been used in many fields for the identification of putative subtypes. One of the key decisions to be made when performing cluster analysis is choosing the "best" partition (set of mutually exclusive clusters) of the data. In this paper, we examine this choice for the situation in which a partitioning clustering algorithm is used to create partitions over a range of K, the number of clusters. Although we do not use hierarchical clustering algorithms in this paper, a similar decision about where to "cut" a hierarchical cluster tree must be made if mutually exclusive clusters are desired, and the indices explored can be applied to that situation.

In K-centroids cluster analysis, the number of clusters K is pre-specified. The measure of clustering quality that is optimized by K-centroids (the sum of distances from objects to their cluster centroids) cannot be directly compared between partitions with different K. This is because as the number of clusters in the partition increases, the value of the criterion decreases or stays the same (i.e. has a monotonic property, Späth, 1985). Therefore, a commonly used strategy is to perform the clustering algorithm for a range of K, typically from $K = 2$ to double the number of what is "expected," then to use some index of partition validity to choose the "best" partition.

In this paper, we consider the "best partition" to be the partition that best agrees with the underlying subtype structure of the data, even if K of this partition is different from the "true" number of classes. (Measurement of "cluster recovery," or "agreement," will be discussed in a later section.) This parallels the situation of class discovery in real data, when the underlying structure of the data is unknown.

In the exploration of real data, indices such as those evaluated in this study would not be the only criteria used to guide choice of partitions for further exploration. Ultimately, it is hoped that putative subtypes identified by cluster analysis represent different etiologies of the disease, which would have different responses to treatment and different outcomes. Indices based solely on the clusters and the data used to create clusters (e.g. disease manifestations) are meant to identify the partition that is best according to the criteria used to create clusters or some other measure of quality of clustering. However, the "best" partition by these criteria might not be the partition that best captures important differences in etiology or response to treatment.

The objective of this study was to evaluate relative validity indices for choosing the best partition of the data. K-means and K-modes were used to create partitions in simulated data with known subtype structures. Indices that performed well in the simulated data (i.e. chose partitions that agreed with the underlying subtype structure) were applied to real data from patients with dermatologic diseases that have overlapping manifestations.

4.2 Methods

Each clustering algorithm was performed on each simulated data set for $K = 2, 3, \ldots, K_{\max}$. Next, the relative validity indices were calculated for each partition. After determining the "usage" of each index that most often made the correct choice, we assessed the influence of data set design factors on index performance.

Indices that performed well on the simulated data were applied to an an empiric data set from patients with overlapping erythematosquamous diseases.

4.2.1 Indices considered

We evaluated the following relative validity indices (criteria) that can be used to choose the best partition from among a set of partitions generated by the K-centroids algorithm.

Notation, terminology. We use $\mathscr{U}_K$ to denote a partition of the data with K clusters, and I_K to denote an index measure for partition $\mathscr{U}_K$. *SSW* is the within-cluster sum of squared error, which is equivalent to the trace of $\mathbf{W}$, and *SSB* is the sum of between-cluster error (see Section 3). For the present discussion, it is assumed that we have one partition of the data for each K under consideration.

1. SSW.

   $I_K = SSW$ (Equation (2.2)), or trace$(\mathbf{W})$.

   This is the optimization criterion for the K-means clustering algorithm (MacQueen, 1967). Edwards and Cavallis.Ll (1965) suggested using the minimum *SSW* as an index of the best level of a dendrogram created by divisive hierarchical clustering using the *SSW*, but did not provide an example of its use for this purpose. Friedman and Rubin (1967) explored using the maximum decrease in *SSW* as an indicator of the number of clusters.

2. Ball and Hall (Hill, 1980).

   (a) Milligan and Cooper (1985) described the Ball and Hall index as "the average distance of the items to their respective cluster centroids."

   (b) $I_K = \frac{SSW}{K}$

   Dimitriadou *et al.* (2002) presented this as the Ball and Hall index, and used the maximum second difference of the index. We used this definition of the Ball and Hall index.

3. Xu (1997).

   $I_K = D\log\left(\sqrt{\frac{SSW}{DN^2}}\right) + \log(K)$, where D is the number of attributes (Xu, 1997).

Xu (1997) designed the index to control for the number of clusters, so that the minimum value indicates the best value for K. Dimitriadou *et al.* (2002) used the maximum second difference of the index.

4. SAD.

   $I_K = SAD$, the sum of absolute differences.

   This is the optimization criterion for the K-modes clustering algorithm.

5. Ball and Hall (SAD).

   $I_K = \frac{SAD}{K}$

6. Xu (SAD).

   $I_K = D\log\left(\sqrt{\frac{SAD}{DN^2}}\right) + \log(K)$.

7. Calinski-Harabasz (Calinski, 1968).

   $I_K = \frac{SSB/(K-1)}{SSW/(N-K)}$.

8. Hartigan (1975).

   $I_K = log(\frac{SSB}{SSW})$.

9. Ratkowsky-Lance (Ratkowsky and Lance, 1978).

   (a) mean$(\sqrt{\frac{SSB_m}{SST_m}})/\sqrt{K}$, where SSB_m is the SSB for the m-th attribute, SST_m is the SST for the m-th attribute (Ratkowsky and Lance, 1978; see Hill, 1980).

   Ratkowsky and Lance intended the maximum value of the index to indicate the best number of clusters.

   (b) mean$(\sqrt{\frac{SSB_m}{SST_m}})$, where SSB_m is the SSB for the m-th attribute, SST_m is the SST for the m-th attribute (Ratkowsky and Lance, 1978; see Hill, 1980).

   This is the form of the index presented by Dimitriadou *et al.* (2002).

10. Davies-Bouldin (Davies and Bouldin, 1979).

    $I_K = \frac{1}{K}\sum_{i=1}^{K} R_i$, where SSW_K is the measure of dispersion and the squared Euclidean distance is the cluster separation measure.

    The Davies-Bouldin index measures the mean "similarity" between each cluster and its "most similar" cluster in a partition (Davies and Bouldin, 1979). The similarity, R_{ij}, between clusters C_i and C_j is a ratio of the sum of the dispersion (compactness) of each cluster to the separation of the clusters:

$$R_{ij} = \frac{S_i + S_j}{M_{ij}},$$

where S_i is the dispersion of cluster C_i, S_j is the dispersion of cluster C_j, and M_{ij} is the distance between the "characteristic vectors" of the clusters.

The form of Davies-Bouldin index used by Milligan and Cooper (1985) and Dimitriadou *et al.* (2002) used *SSW* as the cluster dispersion measure and squared Euclidean distance of cluster means as the cluster separation measure,

$$R_{ij} = \frac{SSW_i + SSW_j}{d_2(\bar{\mathbf{x}}^{(\mathbf{i})}, \bar{\mathbf{x}}^{(\mathbf{j})})}.$$

Davies and Bouldin designed the index so that the minimum value indicates the best number of clusters.

11. Davies-Bouldin (SAD).

    $I_K = \frac{1}{K}\sum_{i=1}^{K} R_i$, where SAD_K (the sum of the absolute difference from each object to the cluster mean, which is equivalent to the simple matching distance for binary data) is the measure of dispersion and the simple matching distance is the cluster separation measure.

12. Friedman and Rubin (1967).

    $\frac{|\mathbf{T}|}{|\mathbf{W}|}$.

    This measure was proposed by Friedman and Rubin as an optimization criterion for clustering, not as an index for choosing the number of clusters (see next criterion); however, it has been presented as an index for number of clusters.

13. Scott-Symons (Scott and Symons, 1971).

    $I_K = N \log \frac{|\mathbf{T}|}{|\mathbf{W}|}$.

    Friedman and Rubin proposed using the maximum increase in natural logarithm of $\frac{|\mathbf{T}|}{|\mathbf{W}|}$ as an indicator of the number of clusters (Friedman and Rubin, 1967), and $N \log \frac{|\mathbf{T}|}{|\mathbf{W}|}$ was proposed by Scott and Symons. We evaluate the latter index, noting that the indices are equivalent in practice because N is constant.

14. Marriott (1971).

    $I_K = K^2|\mathbf{W}|$.

    Marriott proposed using the minimum value of this index to indicate the number of clusters.

4.2.2 Factorial simulations data

Data were simulated with a completely crossed factorial design. The data set design factors were number of subtypes, subtype structure (structures varied in cluster overlap), relative distribution of individuals across subtypes, probability of being present for "high-probability" attributes, and conditional dependence of "high probability" attributes.

A "subtype" was designed as a sub-group of patients with a high probability of a subset of the possible disease manifestations. A "data structure" consisted of the probability distributions for a set of subtypes.

1. *Number of subtypes.* Data were created with three and six subtypes.

2. *Data structures* We used three data structures with differing amounts of overlap between subtypes in the high-probability attributes. Subtypes in structure type 1 did not share any high-probability attributes; subtypes in structure type 2 shared at most one high-probability attribute with another subtype; subtypes in structure type 3 shared at most two high-probability attributes with another subtype. The three- and six-subtype versions of these structures are depicted in the Appendix.

3. *Distribution of individuals across subtypes.* Two levels of distribution across subtypes were considered. The first level was even distribution of patients across subtypes, or the same proportion in each subtype. In the second level, one subtype was larger than the others, with the remaining individuals spread evenly over the remaining subtypes. For three-cluster partitions, the proportions used were (50%-25%-25%); for six-cluster partitions, the proportions used were ().

4. *High probability and intra-subtype dependence.* The three levels of probability for the high-probability attributes were 0.9, 0.8, and 0.7. All low-probability attributes had a probability of 0.1. The three levels of intra-subtype dependence used were $\rho = 0, 0.4$, and 0.8.

5. *Number of individuals.* All artificial data sets had 1,000 individuals.

4.2.3 Empiric erythemato-squamous diseases data

We applied selected indices with high sensitivity and high specificity (using either local optima or co-occurrence of global optima) to K-means and K-modes partitions of a benchmark data set of patients with erythematosquamous diseases.

The six-cluster K-means partition had the best true disease class recovery of all K-means partitions, with an ARI of 0.966. The seven-cluster K-modes partition had the best true disease class recovery of all K-modes partitions, with an ARI of 0.799.

4.2.4 Identification of "best" partitions

Before assessing the ability of indices to choose the best partitions, we identified the "best" K-means and K-modes partition for each data set. The best partition was defined as the partition with the best recovery of true subtype structure, as measured by Hubert and Arabie's Adjusted Rand Index (ARI). This index is a measure of agreement between two partitions of data (Hubert and Arabie, 1985). The ARI range is from -1 to 1, with 1 indicating perfect agreement between two partitions and 0 indicating chance agreement. The ARI is preferred over classification-based methods because it does not require the labelling of clusters (Steinley, 2004).

4.2.5 Evaluation of index performance

Following the procedure used be earlier studies of this type, we considered the maximum and minimum of the following three statistics as possible usages of each index:

1. The index value for partition $\mathscr{U}_k$, I_k.
2. The difference to the left: $I_k - I_{k-1}$.
3. The difference to the right: $I_{k+1} - I_k$.
4. The "second difference," or the elbow sharpness: $(I_k - I_{k+1}) - (I_k - I_{k-1})$.

We calculated these statistics for each partition of the data from $K = 2$ through $K_{\max}$, then identified cases for which the minimum or maximum of each statistic occurred at K of the best partition. In addition to these statistics, we also determined if the best K occurred at a local minimum or maximum of the index.

Indices with one or more usages making a correct choice for more than 25% of partitions were selected for further evaluation. In contrast to earlier studies, we considered that there might be multiple "best usages" of an index. For example, it is possible that a local minimum of an index indicates the best partition in some cases, and the sharpest "elbow" indicates the best partition in other cases (possibly when there is no local minimum).

To characterize index performance further, we calculated the average ARI of partitions chosen by each best usage. We used logistic regression with "correct choice" as the outcome to assess the impact of data set design factors and clustering algorithm used to create the best partition on index performance.

Cluster analysis methods

K-means and K-modes were performed on the data set for $K = 2, \ldots, 9$ (for three-subtype data) or $K = 2, \ldots, 11$ (for six-subtype data), using the combined batch and individual reassignment K-means algorithm

(Späth, 1985). To reduce the chances of finding local rather than global minima, each algorithm (i.e. K-means and K-modes) was performed 10,000 times for each level of K, with the K-th partition taken to be the partition with the lowest *SSW* or *SAD*.

Computational methods

All computations were performed in MATLAB v2008a. The MATLAB statistics toolbox `kmeans.m` function was used for K-means and K-modes. Indices were calculated by MATLAB routines written by the author.

4.3 Results

4.3.1 Factorial simulations

Correct choice of highest-ARI partition.

The percent of best partitions correctly chosen by the global minimum and maximum of the index, the minimum and maximum difference to the left, and the minimum and maximum second difference are shown in Table 4.1. The maximum second difference of the Xu index and Xu_{SAD} index performed the best, choosing 62.2% and 62.9% of the best partitions respectively. The minimum second difference of the Calinski-Harabacz index performed next best overall, choosing 61.0% of the best partitions. The maximum second difference of the *SSW* and *SAD* chose 56.7% and 59.9% of best partitions, respectively. The two versions of the Davies-Bouldin index chose correctly for more than 50% of partitions. The maximum second difference of the Davies-Bouldin (SSW) chose 51.4% correctly, and the global minimum of the Davies-Bouldin (SAD) chose 56.1% correctly.

Several best index/usage pairings agreed with the recommendations of Dimitriadou et al. The maximum second difference chose the best-ARI partition most often for the SSW, Ball-Hall, and Xu indexes, and the minimum second difference chose the best-ARI partition the most often for the Calinski-Harabasz index. For the Davies-Bouldin, Scott-Symons, and Marriott indexes, the usage recommended by Dimitriadou *et al.* (2002) was not the best index/usage combination in our study, but chose correctly almost as well as the best combination. This probably reflects the relationships between the statistics in "picking up" certain shapes. For example, the maximum difference to the left and the minimum second difference tend to indicate an elbow that is convex upward. The best usages of the Hartigan, Ratkowsky-Lance, and Friedman-Rubin indices in our study did not agree with recommendations by Dimitriadou *et al.* (2002).

Logistic regression

For index/usage pairings that chose the best-ARI partition in more than 40% of cases, we evaluated the influence of clustering algorithm used to create partitions and data design factors on correct choice by the index/usage. Logistic regression models were fit using generalized estimating equations to estimate variance

to account for the repeated measures study design. Models included two-way interactions of data design factors with clustering algorithm.

Number of underlying subtypes. The Davies-Bouldin (SAD) index performed significantly better for data with six underlying subtypes. All other indices performed worse for data sets with six underlying subtypes.

Distribution across subtypes. Most indexes performed slightly better for data with one subtype larger than the others (distribution level 2). Distribution level 2 had a significant negative effect on the Davies-Bouldin (SAD) and Scott-Symons indices.

Subtype overlap. The Davies-Bouldin-SSW, Davies-Bouldin-SAD, and Hartigan indexes were the worst performers in data with overlapping subtypes. The effect of overlap level 1 (up to one high-probability variable overlapping between subtypes) on these indexes were ($\hat{\beta} = -0.1.37$), ($\hat{\beta} = -0.73$), and ($\hat{\beta} = -0.99$), all significant. The effects of overlap level 2 (up to two high-probability variables overlapping between subtypes) on these indexes were ($\hat{\beta} = -1.78$), ($\hat{\beta} = -1.90$), and ($\hat{\beta} = -1.34$), also all significant. The SSW, SAD, Calinski-Harabasz, and Scott-Symons indexes also had a negative dose-response relationship with increasing subtype overlap, with the effect not significant for overlap level 1.

High-probability variables. Decreasing levels of high probability (the probability of being present) adversely affected the performance of all indexes. The lowest level of high probability used (0.7) had a significant negative effect on the performance of all indexes except the Davies-Bouldin index, with the strongest negative effect on the Scott-Symons, Xu, and Xu (SAD) indexes.

Intra-subtype dependence. All indexes performed worse on data with intra-subtype dependence of high-probability variables.

Clustering algorithm, main effect and interactions. Contrary to what was expected, *K*-modes had a significant negative main effect on the performance on *SAD*, Xu (SAD), and Davies-Bouldin (SAD). However, there was a positive interaction between *K*-modes and some data design factors for these and other indexes. For the Davies-Bouldin and Davies-Bouldin (SAD) indexes, there was a positive interaction between *K*-modes and distribution level 2 (one large subtype). Decreasing levels of high probability had less of a negative effect on index performance on partitions created by *K*-modes for the SAD, Xu, Xu (SAD), Davies-Bouldin, and Calinski-Harabasz indexes. The negative effect of subtype overlap on performance of the SSW, Davies-Bouldin, Hartigan, and Scott-Symons indexes were lessened for partitions created by *K*-modes. *K*-modes had a positive interaction with intra-subtype dependence, lessening the negative effect, for the Xu (SAD) and Davies-Bouldin (SAD) indexes, and a negative interaction with intra-subtype dependence

for the Scott-Symons index.

4.3.2 Increasing index sensitivity

We next considered expansion of the "best usage" statistic of a cluster to other usages that performed well. Table 4.4 shows the percent of data sets for which the highest-ARI partition was chosen by any statistic of an index that chose correctly more than 20% of the time. In several cases, inclusion of other usages increased the probability of identifying the highest-ARI partition.

We chose the Davies-Bouldin, Davies-Bouldin (SAD), Xu, and Xu (SAD) indices for further evaluation. These indices had several usages that performed well overall. Combining the best usages improved the sensitivity of the index/usage combinations, or the probability of one of the best usages of the index to choose the best partition. Either the minimum difference to the left or the maximum second difference (or both) of the Xu and Xu (SAD) indices chose the best partition in 70.5% and 69.8% of cases. For the Davies-Bouldin and Davies-Bouldin (SAD) indices, 79.7% and 78.3% of the best partitions were chosen by either the maximum second difference, the the maximum difference to the right, or the global minimum.

4.3.3 Increasing index specificity

There is an important drawback of evaluating index performance by determining the percent of best partitions identified by index/statistic combinations. Because there is always a global minimum or maximum of these statistics, index/statistic combinations have no specificity.

We explored two types of specificity for selected index/global statistic optima pairings. First, we evaluated the ability of a local shape occurring at a statistic's global optimum to identify cases in which the global optimum correctly chose the best partition. Second, we evaluated the ability of the joint occurrence of two statistics' global optima to identify cases in which the global optima chose correctly.

Local shapes

Figure 4.1 gives examples of local shapes that can be found in the plot of an index for a range of partitions: a local minimum, a local maximum, an upward-concave elbow with an overall decrease in index value, a downward-concave elbow with an overall decrease in index value, an upward-concave elbow with an overall increase in index value, a downward-concave elbow with an overall increase in index value.

In Figure 4.2, the contingency table on the left shows how local shapes can be used to create specific index usages. In this table, the "test" is that a certain local shape occurs at the same K as a certain statistic's global optimum of choice. The table on the left depicts the general contingency table for using local shapes to increase specificity, where the global optimum of statistic A is at K_A. The table on the right depicts use of local minima to identify cases in which global minima of the Davies-Bouldin index have chosen correctly.

Intersection of global optima

In Figure 4.3, the table on the left depicts the general contingency table for using the co-occurrence of two statistics' global optima to increase specificity, where the global optimum of statistic A is at K_A and the global optimum of statistic B is at K_B. The table on the right depicts use of joint occurrence of the maximum difference to the left and minimum different to the right to identify cases in which global minima of these statistics for the Calinski-Harabasz index have chosen correctly.

4.3.4 Empiric erythemato-squamous disease data

Figures 4.4 and 4.5 show plots of the the four selected indices against the K-means and K-modes partitions for $K = 2, \ldots 11$.

Table 4.7 shows the K of the partition chosen by the minimum and maximum of the four usage statistics. There was only one instance in which some usage of an index recommended the best partition. The Davies-Bouldin index had a global minimum index value and global maximum difference to the right at the six-cluster K-means partitions. In Study 1, co-occurrence of these two global optima had a high specificity for correct choice of partition. However, several global optima intersections that had high specificity in Study 1 did not choose the best partition of the benchmark data. The minimum difference to the left and maximum second difference of the Davies-Bouldin and Davies-Bouldin (SAD) indices occurred at the three-cluster K-means partition, which was not the best partition. Several co-occurrences of usages of these two indices that had high specificity in Study 1 were seen for the K-modes partitions, but did not choose the best partition. The Xu and Xu (SAD) indices had global minima of the difference to the left and global maxima of the second difference for the same partition from each set of partitions (K-means, K-modes). All four of these co-occurrences occurred at $K = 3$, which was not the best partition for K-means or K-modes.

$|\mathbf{W}|$-based indices

The Friedman-Rubin, Scott-Symons, and Marriott indices are based on the determinants of $\mathbf{W}$ and $\mathbf{T}$, the within-cluster and total sums of squares and cross products matrices (Friedman and Rubin, 1967; Scott and Symons, 1971; Marriott, 1971). Application of these indices to the benchmark data set revealed a problem that may occur when using $|\mathbf{W}|$-based indices on presence/absence data.

The determinant of a matrix is 0 if all elements of the i-th row and j-th column, where $i = j$, are 0. For binary data, this occurs in the sum of squares and cross products matrix (either total or within-cluster) when the probability of an attribute is equal to 0 or 1.00. For the benchmark data, $|\mathbf{W}|$ was equal to 0 for the five-cluster K-means partitions. Table 4.8 shows that the cluster mean for attribute 33 is either 0 or 1 for every cluster in the partition, causing $|\mathbf{W}|$ to be 0. As a result, the Scott-Symons and Friedman-Rubin indices were

undefined for the partition, and the Marriott index was 0.

4.4 Discussion

We found that several index/global optima pairs had reasonably high sensitivity in simulated data, with the pairings of the Davies-Bouldin and Xu indices among the best performers in terms of sensitivity. We explored creating index usages with specificity by using the co-occurrence of index/global optima pairs with local shapes, or by using the intersection of the global optima of several usage statistics for a given index. Several of these combinations for the Xu and Davies-Bouldin indices had high specificity for correct identification of the best partition.

We hypothesized that versions of *SSW*-based indices using *SAD*, the *K*-modes clustering criteria, would perform better than the *SSW*-based criteria for partitions created with *K*-modes. We did not find this to be the case in our simulations.

Increasing index specificity. Because there is always a global optima of an index or index statistic, any index/global statistic optima pairing cannot be "negative," and thus has a specificity of 0. In the context of application to real data, this means that the index usage does not have a positive predictive value. We investigated several approaches for creating an index usage with specificity, which may have the capability to create more complete characterizations of index performance.

The Davies-Bouldin index provides a good example of one of our approaches for increasing sensitivity, combining local shapes with global optima. The global minimum of the Davies-Bouldin index occurred at the best partition in only 13% of cases in our study. If the global minimum was also a local minimum, the specificity of the global minimum for occurring at the correct partition was near 100%.

Analytic approach. We evaluated the performance of relative validity indices in the same manner as previous studies, whereby the "best usage" of an index is determined by calculating the index for sets of partitions over a range of number of clusters, then identifying the characteristics of the index plot over all *K* that most often occur at the best partition. As previous studies had done, we evaluated the global maximum and minimum of four statistics: the index itself, the difference to the left, the difference from the right, and the second difference. There are several problems with this approach.

First, some of the indices evaluated include adjustments that were meant to allow direct comparison of the index across levels of *K*, so that the minimum or maximum of the index indicates the best partition. Evaluating the global optima of the second difference and other statistics seems counterintuitive if an index is meant to be used in a straightforward manner. (Unlike, for example, looking for "elbows" in a statistic such as the *SSW* that decreases monotonically.) Among the indices evaluated in this study, the minimum of the Xu index, the minimum of the Davies-Bouldin index, and the maximum of the Ratkowsky-Lance index are meant to indicate the best partition, respectively.

Another drawback of this type of study is that if several global statistic optima have similar performance, the choice of the "best" usage may somewhat arbitrary. By dropping usages of the index that might work well in some situations without further investigation, the ability of an index to choose the best partition might be underestimated.

Considerations for some indices. Some of these indices were actually proposed as alternative clustering criteria. For example, $|\mathbf{W}|$ is presented in several studies of relative validity indices as an index, with reference to Friedman and Rubin (1967), (e.g. Milligan, 1981a; Dimitriadou *et al.*, 2002). In fact, Friedman and Rubin (1967) proposed the use of $|\mathbf{W}|$ as an alternative criterion for the K-means algorithm, in place of $\text{tr}(\mathbf{W})$. It may be inappropriate to apply an index that is actually an alternative clustering criterion for the following reason. (It should be noted that Friedman and Rubin (1967) evaluated use of the K-means clustering criterion as an index for choosing the best partition created by $|\mathbf{W}|$).

Conclusion. We conclude that the indices described in this study have less than desirable accuracy. However, our findings suggest that it may be possible to create index usages with high specificity. An important implication of our findings is that it may unwise to apply these indices in a blind manner, where many cluster analyses will be performed and best partitions chosen automatically by an index. Finally, we wish to emphasize that we do not believe indices such as those evaluated in this study should be used as the only evidence for choosing a set of clusters to be considered as candidate subtypes, a choice which should ultimately consider differences in etiology and treatment response between clusters. Instead, these indices should be considered as suggestions of the partition that best met the K-means or K-modes clustering criteria.

4.5 Figures and tables.

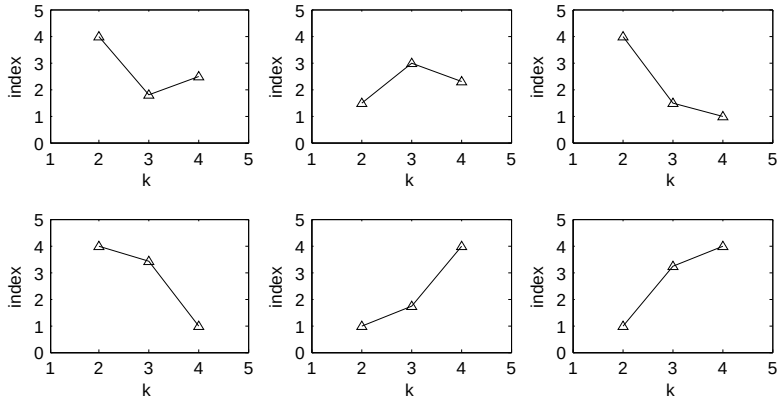

Figure 4.1: Local shapes. Upper row from left to right: local minimum, local maximum, elbow concave upward (decreasing). Bottom row from left to right: elbow concave downward (decreasing), elbow concave upward (increasing), elbow concave downward (increasing).

$K_A = K_{\text{best}}$

Local shape at K_A		*yes*	*no*	
	yes	a	b	$(a+b)$
	no	c	d	$(c+d)$
		$(a+c)$	$(b+d)$	total

$K_{\text{global min}} = K_{\text{best}}$

Local min at $K_{\text{global min}}$		*yes*	*no*	
	yes	490	449	939
	no	32	857	889
		522	1306	1728

Figure 4.2: Utilizing local shapes.

$K_A = K_B = K_{\text{best}}$

$K_A = K_B$		*yes*	*no*	
	yes	a	b	$(a+b)$
	no	c	d	$(c+d)$
		$(a+c)$	$(b+d)$	total

$K_A = K_B = K_{\text{best}}$

$K_A = K_B$		*yes*	*no*	
	yes	482	180	662
	no	0	1066	1066
		482	1246	1728

Figure 4.3: Utilizing intersections of two statistics' global optima.

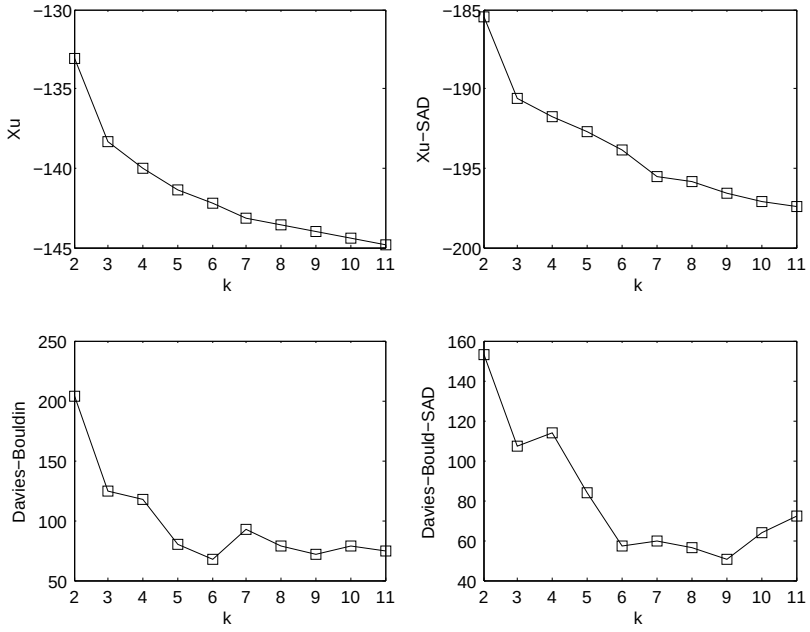

Figure 4.4: Indices for K-means partitions of benchmark data set.

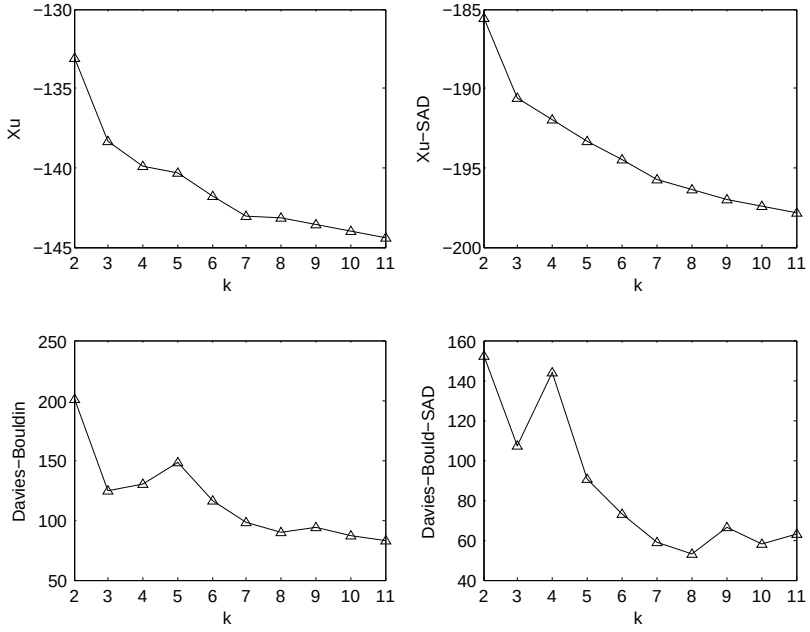

Figure 4.5: Indices for K-modes partitions of benchmark data set.

Table 4.1: Percent of best-ARI partitions chosen correctly by the minimum and maximum of each statistic, by index.

		I_k			$I_k - I_{(k-1)}$			$I_{(k+1)} - I_k$			$(I_{(k+1)} - I_k) - (I_k - I_{(k-1)})$	
Index		*min*	*max*		*min*	*max*		*min*	*max*		*min*	*max*
SSW		1.97	0.00		36.75	2.20		0.17	1.33		0.93	56.66
Ball-Hall		1.97	0.00		34.26	1.97		0.00	1.33		0.00	34.49
Xu	1.97		0.00	50.93	4.17		3.01	22.51		3.82	62.21	
SAD	1.97		0.00	40.16	2.20		1.62	1.56		2.26	59.90	
Ball-Hall (SAD)		1.97	0.00		34.26	1.97		0.00	1.33		0.00	34.43
Xu (SAD)		2.31	0.00		57.75	3.76		5.15	13.60		4.63	62.91
Calinski-Harabasz		1.50	43.87		5.44	52.03		44.91	6.02		61.00	5.15
Hartigan		0.00	1.97		2.20	35.01		1.50	0.17		47.97	1.33
Ratkowsky-Lance		2.03	0.06		34.72	2.66		0.12	2.78		2.31	36.63
Davies-Bouldin		30.21	0.00		35.19	2.37		1.33	51.16		2.72	51.39
Davies-Bould-SAD		56.13	0.58		35.59	3.36		2.89	50.00		2.95	45.43
Friedman-Rubin		0.00	1.97		28.47	3.82		2.03	1.74		11.05	3.59
Scott-Symons		0.00	1.97		1.91	37.27		2.43	7.00		44.39	8.97
Marriot		2.03	0.00		32.06	2.66		3.76	5.50		6.02	28.94

Table 4.2: Interaction models

	SSW		SAD		Xu		Xu2	
Effect	Est.		Est.		Est.		Est.	
(Intercept)	3.31,	**	2.93,	**	3.19,	**	2.99,	**
K-modes	−0.41		−0.99,	**	−0.31		−0.83,	**
$G = 6$	−1.61,	**	−0.91,	**	−0.67,	**	−0.28	
Distribution 2	0.29		0.61,	**	0.53,	*	0.68,	**
Overlap 1	−0.25		0.07		0.04		0.05	
Overlap 2	−1.67,	**	−0.75,	**	−0.44		−0.32	
$H = 0.8$	0.03		0.12		−0.76,	**	−0.24	
$H = 0.7$	−0.84,	**	−1.19,	**	−1.82,	**	−1.67,	**
$\rho = 0.4$	−1.51,	**	−1.57,	**	−2.13,	**	−1.95,	**
$\rho = 0.8$	−2.62,	**	−2.82,	**	−1.97,	**	−2.76,	**
K-modes × $G = 6$	0.29		−0.21		−0.19		−0.51,	**
K-modes × Distribution 2	0.01		0.14		0.08		0.09	
K-modes × Overlap 1	0.08		0.23		−0.21		0.06	
K-modes × Overlap 2	0.50,	*	0.18		−0.28		0.26	
K-modes × $H = 0.8$	0.28		0.73,	**	0.87,	**	0.81,	**
K-modes × $H = 0.7$	0.21		0.80,	**	0.61,	*	0.60,	**
K-modes × $\rho = 0.4$	−0.31		0.13		−0.19		0.04	
K-modes × $\rho = 0.8$	0.23		0.20		−0.00		0.56,	*

*$P < 0.01$

**$P < 0.001$

Table 4.3: Interaction models

	DB		DB2		CH		Ha		SS	
Effect	Est.		Est.		Est.		Est.		Est.	
(Intercept)	2.99,	**	2.40,	**	2.89,	**	3.18,	**	3.20,	**
K-modes	−0.58		−0.74,	*	−0.10		0.05		0.24	
$G = 6$	−2.28,	**	0.98,	**	−0.83,	**	−3.39,	**	−1.20,	**
Distribution 2	0.74,	**	−0.49,	*	0.54,	**	0.84,	**	−0.79,	**
Overlap 1	−1.37,	**	−0.73,	**	−0.17		−0.99,	**	−0.45	
Overlap 2	−1.78,	**	−1.90,	**	−1.18,	**	−1.34,	**	−1.82,	**
$H = 0.8$	−0.03		0.06		−0.33		0.00		−1.53,	**
$H = 0.7$	−0.29		−0.62,	*	−1.52,	**	−0.91,	**	−2.54,	**
$\rho = 0.4$	−0.82,	**	−1.57,	**	−1.56,	**	−1.22,	**	−1.73,	**
$\rho = 0.8$	−2.25,	**	−1.94,	**	−1.72,	**	−2.19,	**	−0.25	
K-modes × $G = 6$	−0.10		−0.11		−0.16		0.92,	**	−0.05	
K-modes × Distribution 2	0.47,	*	0.81,	**	0.07		−0.33		0.20	
K-modes × Overlap 1	0.93,	**	0.21		−0.16		0.73,	**	0.38	
K-modes × Overlap 2	0.02		0.25		0.19		0.20		1.04,	**
K-modes × $H = 0.8$	0.50,	*	−0.28		0.51,	*	−0.12		0.34	
K-modes × $H = 0.7$	0.68,	*	−0.37		0.46,	*	−0.39		0.21	
K-modes × $\rho = 0.4$	−0.29		0.00		−0.28		−0.42		−0.77,	*
K-modes × $\rho = 0.8$	−0.31		0.84,	**	−0.16		−0.00		−0.87,	**

*$P < 0.01$

**$P < 0.001$

Table 4.4: Multiple index usages.

Index	Best	ARI	>20%	# usages	ARI
SSW	56.66	0.786	62.04	2	0.769
Ball-Hall	34.49	0.787	34.49	2	0.787
Xu	62.21	0.786	74.88	3	0.768
SAD	59.90	0.782	63.77	2	0.770
Ball-Hall-SAD	34.43	0.789	34.84	2	0.785
Xu-SAD	62.91	0.784	69.79	2	0.773
Calinski-Harabasz	61.00	0.786	74.25	4	0.780
Hartigan	47.97	0.787	52.49	2	0.770
Ratkowsky-Lance	36.63	0.788	36.92	2	0.788
Davies-Bouldin	51.39	0.774	82.58	4	0.760
Davies-Bould-SAD	56.13	0.784	81.54	4	0.767
Friedman-Rubin	28.47	0.782	28.47	1	0.782
Scott-Symons	44.39	0.807	56.13	2	0.790
Marriot	32.06	0.802	34.90	2	0.793

Table 4.5: Statistic/index vs. local shapes

Index	gstat	%	locshape	a	$(a+b)$	Sn	Sp
CH	max(I_k)	43.87	locmax	43.87	66.49	100.00	59.69
CH	max(diff L)	52.03	locmax	41.38	62.91	79.53	55.13
CH	min(diff R)	44.91	locmax	34.78	51.04	77.45	70.48
CH	min(2nd diff)	61.00	locmax	46.76	72.28	76.66	34.57
CH	max(I_k)	43.87	elb-downD	0.00	0.00	NaN	NaN
CH	max(diff L)	52.03	elb-downD	9.03	18.00	17.35	81.30
CH	min(diff R)	44.91	elb-downD	10.13	22.86	22.55	76.89
CH	min(2nd diff)	61.00	elb-downD	14.12	25.75	23.15	70.18
DB	min(I_k)	30.21	locmin	28.36	48.55	93.87	71.06
DB	min(diff L)	35.19	locmin	17.30	18.23	49.18	98.57
DB	max(diff R)	51.16	locmin	48.84	78.88	95.48	38.51
DB	max(2nd diff)	51.39	locmin	34.38	41.38	66.89	85.60
DB2	min(I_k)	56.13	locmin	55.32	89.00	98.56	23.22
DB2	min(diff L)	35.59	locmin	25.17	33.04	70.73	87.78
DB2	max(diff R)	50.00	locmin	45.20	74.13	90.39	42.13
DB2	max(2nd diff)	45.43	locmin	37.85	53.88	83.31	70.63

Table 4.6: Co-occurrence of index/global optimum (statistic A) and index/global optimum (statistic B), specificity (Sp).

Index	Statistic A	$K_A = K_{\text{best}}$	Statistic B	$K_B = K_{\text{best}}$	$(a+b)$ $K_A = K_B$	$(a+c)$ $K_A = K_B = K_{\text{best}}$	Sp
Xu	min(diff L)	50.93	max(2nd diff)	62.21	61.23	42.65	67.61
Xu2	min(diff L)	57.75	max(2nd diff)	62.91	76.74	50.87	47.35
CH	max(I_k)	43.87	max(diff L)	52.03	51.10	35.88	76.26
CH	max(I_k)	43.87	min(diff R)	44.91	73.50	33.74	40.00
CH	max(I_k)	43.87	min(2nd diff)	61.00	61.98	40.51	63.91
CH	max(diff L)	52.03	min(diff R)	44.91	38.31	27.89	85.55
CH	max(diff L)	52.03	min(2nd diff)	61.00	72.63	47.11	51.75
CH	min(diff R)	44.91	min(2nd diff)	61.00	56.48	37.04	69.12
DB	min(I_k)	30.21	min(diff L)	35.19	4.80	4.75	99.94
DB	min(I_k)	30.21	max(diff R)	51.16	29.98	21.82	89.56
DB	min(I_k)	30.21	max(2nd diff)	51.39	16.15	14.47	98.04
DB	min(diff L)	35.19	max(diff R)	51.16	15.74	15.10	99.25
DB	min(diff L)	35.19	max(2nd diff)	51.39	64.70	32.70	52.45
DB	max(diff R)	51.16	max(2nd diff)	51.39	34.49	28.99	92.26
DB2	min(I_k)	56.13	min(diff L)	35.59	25.41	21.18	94.64
DB2	min(I_k)	56.13	max(diff R)	50.00	56.89	36.98	68.41
DB2	min(I_k)	56.13	max(2nd diff)	45.43	41.55	31.02	84.73
DB2	min(diff L)	35.59	max(diff R)	50.00	18.23	14.53	95.67
DB2	min(diff L)	35.59	max(2nd diff)	45.43	73.96	32.35	38.49
DB2	max(diff R)	50.00	max(2nd diff)	45.43	36.52	24.88	84.51

Table 4.7: K chosen for each set of partitions of benchmark data by the minimum and maximum of index usage statistics for each index.

K-means partition ($K_{\text{best}} = 6$)

	I_k		$I_k - I_{(k-1)}$		$I_{(k+1)} - I_k$		2nd diff	
Index	*min*	*max*	*min*	*max*	*min*	*max*	*min*	*max*
Xu	11	2	3	11	2	10	9	3
Xu2	11	2	3	11	2	10	6	3
DB	6	2	3	7	2	6	7	3
DB2	9	2	3	10	2	9	4	3

K-modes partition ($K_{\text{best}} = 7$)

	I_k		$I_k - I_{(k-1)}$		$I_{(k+1)} - I_k$		2nd diff	
Index	*min*	*max*	*min*	*max*	*min*	*max*	*min*	*max*
Xu	11	2	3	8	2	7	5	3
Xu2	11	2	3	11	2	10	6	3
DB	11	2	3	5	2	4	5	3
DB2	8	2	5	4	4	3	4	3

Table 4.8: K-means $K = 5$ partition, cluster means for selected attributes.

	Attributes				
Cluster	29	30	**31**	32	33
C_1	0.00	0.02	**0.00**	0.98	0.02
C_2	0.99	0.01	**0.00**	0.99	1.00
C_3	0.01	0.00	**0.00**	0.95	0.01
C_4	0.00	0.00	**0.00**	0.96	0.02
C_5	0.00	0.95	**1.00**	0.95	0.05

CHAPTER 5

Paper 3: Using class discovery methods to explore early lupus subtypes in the Georgia Lupus Registry

Abstract

Objective. To identify clusters of early systemic lupus erythematosus (SLE) patients with similar signs and symptoms to be considered as candidate early lupus subtypes.

Methods. The study population consisted of incident SLE patients from the Georgia Lupus Registry. K-means cluster analysis, K-modes cluster analysis, and latent class analysis (LCA) were used to create partitions, or sets of clusters of patients with similar clinical and laboratory manifestations. Internal and external validation methods were used to compare partitions.

Results. K-means, K-modes, and LCA suggested different four-cluster partitions of the patients, but there were similarities between the partitions. Each partition had a cluster characterized by proteinuria, thrombocytopenia, and hypocomplementemia that was strongly associated with death within 2 years of diagnosis and end-stage renal disease. Each of the partitions had two clusters characterized by anti-RNP and anti-DNA, one of which was also characterized by cutaneous manifestations. Each of the partitions had a non-specific cluster with no characteristic features and a mean age of diagnosis than other clusters in the partition. In regression models, the four-cluster K-means partition, the four-cluster K-modes partition, and the three-cluster LCA partition had the best fit with death and end-stage renal disease.

Conclusion. The methods and the partitions described can inform further exploration of early lupus subtypes.

5.1 Introduction

Systemic lupus erythematosus (SLE) is an autoimmune disease characterized by antibodies to components of the cell nucleus. Patients with SLE can have autoimmune-mediated activity in virtually any organ system, which may result in permanent damage such as end-stage renal disease (ESRD) or lung fibrosis. SLE is diagnosed by the recognition of a constellation of symptoms and signs, and disease activity and organ damage vary widely between patients. The etiology of SLE is poorly understood, and it has been suggested that the use of heterogeneous patients in research has hampered progress in describing the etiology of SLE. It is possible that different etiological pathways or key differences in a common etiologic pathway give rise to several syndromes with manifestations in the SLE spectrum but having greater inter-patient homogeneity than SLE patients taken as a whole. Presumably, such disease or syndrome subtypes would have different risks of clinical outcomes and different responses to treatment.

Formal class discovery methods such as cluster analysis or latent class analysis directly or indirectly identify sub-groups of patients with shared features, or subphenotypes. Consideration of such subphenotypes as possibly arising from several etiologies is analagous to the epidemiologic practice of preliminary case definition for patients with a previously undescribed disease. Typically, in this situation, a collection of features shared by the group of sick patients is used as a preliminary case definition, and then the preliminary case definition is used to investigate the etiology and outcomes of the "new" disease. The case definition is revised as information about the disease process or clinical outcomes is gained (Lasky and Stolley, 1994; Tyler and Last, 1998). Investigations of subsets has long been present in the SLE research literature, but few studies have used formal class discovery methods such as finite mixture models or cluster analysis (To and Petri, 2005; Jurencak *et al.*, 2009).

In this study, we used three formal class discovery methods to create a range of partitions (sets of mutually exclusive clusters) of early SLE patients, using clinical and laboratory disease manifestations. K-means cluster analysis, K-modes cluster analysis, and latent class analysis were used to create partitions. Internal and external validation techniques were used to suggest "candidate" partitions, or sets of clusters to explore further as prototypes of early lupus phenotypes.

5.2 Methods

5.2.1 Study population: Georgia Lupus Registry

The study population consisted of early lupus patients from the Georgia Lupus Registry. The American College of Rheumatology (ACR) Criteria for the Classification of SLE classify a patient as having SLE if 4 or more of 11 clinical and laboratory criteria are met (Hochberg, 1997). For this study, the case definition of early SLE was the presence of 3 ACR criteria and diagnosis of SLE by a rheumatologist or 4 or more ACR criteria, with first diagnosis of "possible lupus" from Janary 1, 2002 to December 31, 2004. Of the 234 patients who met the case definition, patients with race other than African-American or white were excluded (n = 6), and patients without measurements for laboratory variables to be used in class discovery analyses were excluded $(n = 77)$, leaving a study population of 151 patients. Excluded patients were similar to the study population in the distributions of age of diagnosis, race, and sex (all P-values > 0.05).

26 clinical and 11 laboratory variables were used for class discovery analysis. Demographic variables, end stage renal disease (ESRD), and death within two years of diagnosis of possible lupus were used to characterize candidate partitions. Information on demographic variables, ESRD, and death were abstracted from medical records by the Georgia Lupus Registry.

5.2.2 Analysis methods

We used two types of K-centroids cluster analysis (K-means and K-modes) and latent class analysis to create partitions, or sets of exhaustive and mutually exclusive clusters. In both of these class discovery methods, the number of clusters or latent classes is pre-specified by the researcher. The standard approach is to create partitions or models for a range of number of clusters (or latent classes), then to use validation techniques to identify the "best" partitions. Internal validation methods quantify how well a partition meets the clustering criterion (K-centroids cluster analysis) or how well a model fits the data (latent class analysis). External validation methods utilize information not used in the class discovery analysis, such as association of clusters with additional variables from the original data set, or reproducibility in a second study population. After creating partitions of the study population using K-centroids clustering and LCA, we used a combination of internal and external validation methods to suggest "best" partitions.

5.2.3 Class discovery and internal validation

K-centroids cluster analysis

K-means is perhaps the most popular ad-hoc algorithm for cluster analysis (Steinley, 2006b). The K-means algorithm places patients into a pre-specified number of clusters (K) in such a way as to minimize the sum of within-cluster squared Euclidean distances from each patient to their respective cluster mean (MacQueen,

1967). *K*-modes is a variation of *K*-means that uses the Hamming distance (simple matching distance) and cluster mode as the distance and centroid, respectively, so that the sum of the number of matches of each patient to their respective cluster mode is minimized (Huang, 1998). Relative validity indices based on adjustments of the clustering criteria have been reported to perform well in choosing the best *K*-means partition in some situations (Milligan and Cooper, 1985; Dimitriadou *et al.*, 2002) However, in our own evaluation of such indices, we found that the previously recommended indices did not reliably perform well in choosing *K*-means or *K*-modes partitions in data scenarios similar to what might be anticipated in the study population (Speckman *et al.*, in preparation). Therefore, we used a bootstrapping method inspired by the gap statistic to suggest best partitions from those created by *K*-means and *K*-modes.

***K*-centroids internal validation.** In the gap statistic of Tibshirani *et al.* (2001), application of the clustering method to numerous simulated data sets is used to create a bootstrap null distribution for each of the clustering criteria. The best *K* is chosen as the smallest *K* for which the observed value is more than one standard error less than the expected. As a modification of this approach, we compared the observed value of the criterion for the real data from each level of *K* to the null distribution for that *K*, with the 2.5th percentile used as a guideline for significance. To create a null distribution of the sum of within-cluster squared error (the *K*-means clustering criterion), we created 200 simulated data sets with the same number of patients as the study sample. Each simulated patient was "drawn" from the observed marginal probability distribution of clustering variables. For example, 71.5% of the study population had arthritis, and a simulated patient was assigned to have arthritis if a random number generated from the uniform distribution over (0, 1) was at or below 0.715. *K*-means was performed on each simulated data set using the same implementation method (number of replications, etc.) used on the real data, for each level of *K* considered. The resulting sum of within-cluster squared errors from the simulated data forms the bootstrap null distribution. The same method was used to create a bootstrap null distribution for the *K*-modes clustering criterion.

Latent class analysis

In latent class analysis, the data is modeled as being the result of a number of underlying classes, each of which has its own probability distribution for the observed variables (McCutcheon, 1987). In other words, a mixture of multivariate Bernoulli distributions is fit to the data. The expectation-maximization algorithm is used to obtain maximum-likelihood parameter estimates for the population proportion in each latent class and the multivariate Bernoulli distribution of each latent class (Dempster *et al.*, 1977). The latent class model is used to calculate posterior probabilities of membership in each class for the patients, then each observation is assigned to the class for which it has the highest posterior probability of membership to create mutually exclusive clusters.

LCA internal validation. We used two types of measures to assess model fit, the bootstrap likelihood ratio test (LRT) and information criteria (IC) measures. The standard chi-square LRT is theoretically inappropriate for use in comparing latent class models (McLachlan, 1987). In the bootstrap LRT, many data sets are generated using the parameter estimates from the latent class model with G classes, the "null" model. LCA is performed on each of these data sets for G and $(G+1)$ classes to create a null distribution for the log likelihood ratio of the $(G+1)$ model compared to the g model; then, the observed log likelihood ratio is compared to the null distribution to obtain a P-value (McLachlan, 1987).

Information criteria have the general form

$$IC = -2logL(\theta_j) + \text{penalty term(s)},$$

where $L(\theta_j)$ is the likelihood from model j (Yang and Yang, 2007). Akaike's information criterion (AIC) has the penalty term $2p$, and Schwarz's information criterion (BIC) has the penalty term $p\log(N)$, where p is the number of free parameters in the model and N is the number of observations. We used the following IC, which contain sample size adjustment terms: Draper's BIC (DBIC), Hurvich and Tsai's AIC (HT-AIC), and the BIC and CAIC with Sclove's sample size adjustment (BIC2, CAIC2). The minimum values of these criteria indicate the best model(s) (Yang and Yang, 2007).

External validation

Association of cluster membership with external variables (variables not used in the cluster analysis) was assessed using logistic regression with cluster membership as the main exposure. Cluster 1 of each partition was used as the reference group. Outcomes considered were ESRD and death within two years of date of diagnosis. An unadjusted and adjusted model were fit for each partition with each outcome. The adjusted ESRD models included race, age, and sex as covariates; the adjusted mortality models included race, age, sex, and ESRD as covariates. Logistic regression models were fit using Firth's penalized maximum likelihood estimation method to reduce bias from small sample size and allow parameter estimation in the presence of complete or quasi-separation of data (Firth, 1993). Model fit comparisons were made with the same information criteria used to compare latent class models.

5.2.4 Comparing and characterizing partitions

We evaluated agreement between partitions using a measure of partition agreement commonly used in the class discovery literature, Hubert and Arabie's adjusted Rand index (Hubert and Arabie, 1985; Steinley, 2004). Following the advice of Milligan, we did not conduct statistical tests for differences between clusters in features that were used to create clusters (Milligan, 1996). The clusters were described and compared by

the raw percentage of cluster members with each feature.

Software and statistical methods implementation

K-means and K-modes cluster analysis were performed with MATLAB v2008a using the Statistics Toolbox kmeans routine, which uses Späth's combined batch and individual reassignment algorithm (Späth, 1985). For each partition generated, we performed a large number of replications (10,000) of the algorithm with randomly selected starting seeds in order to identify global minima of the clustering criteria, as recommended by (Steinley, 2003). Validity indices for K-centroids were implemented with MATLAB routines, and 200 simulated data sets were used to make bootstrap null distributions for K-means and for K-modes. Latent class analysis and measurement of model fit were performed in MATLAB with routines written by RAS. (Several commercial software packages can perform similar analyses, for example, M-plus and LatentGold.) In order to identify global maxima of the likelihood function, 1,000 replications of the E-M algorithm were performed for each model. All other analyses were performed in SAS.

5.3 Results

5.3.1 Internal cluster validation

Figure 5.1 shows the observed clustering criteria for the K-means and K-modes partitions with $K = 2$ through 5, and the 2.5th, 50th, and 97.5th percentiles for the clustering criteria from K-means and K-modes partitions of data generated from the observed marginal probabilities, or the bootstrap null distributions of the clustering criteria. The four- and five-cluster partitions created by K-means and K-modes had observed clustering criteria less than or equal to the 2.5th percentile of the bootstrap null distribution, indicating that these partitions reflected real structure in the data and are not forced partitions of the data.

The bootstrap likelihood ratio test indicated that the latent class models for $G = 2$ through 5 successively fit the data better than the $(G-1)$ model, indicating that these partitions were not spurious (Figure 5.2). Table 5.1 shows the CAIC2, BIC2, DBIC, and AIC3 values of the latent class models for $G = 1$ through 5. The CAIC2, BIC2, and AIC3 were lowest for the four-class model, and the DBIC was lowest for the two-class model, suggesting that these two models were better fits for the data than the other models.

5.3.2 External cluster validation

Among K-means partitions, models of the four-cluster partition with ESRD and death had the best fit of the unadjusted models, the two-cluster partition had the best adjusted fit with ESRD, and the four-cluster partition had the best adjusted fit with death (Table 5.2). Among K-modes partitions, the four-cluster partition had the best fit with ESRD and death in both unadjusted and adjusted models. Among LCA partitions, the three-cluster partition had the best adjusted and unadjusted fit with ESRD, and the two-cluster partition had the

best unadjusted and adjusted fit with death. Of all ESRD models, the LCA three-cluster partition had the best fit among unadjusted models and the *K*-means two-cluster partition had the best fit among adjusted models. Of all mortality models, the LCA two-cluster partition had the best fit among both unadjusted and adjusted models.

5.3.3 Characterization of selected partitions

Based on consideration of the internal and external validation results, we selected one partition from each clustering method for further characterization. The four-cluster *K*-means partition was chosen because internal validation indicated that either the four- or five-cluster partition reflected real structure, and the four-cluster model had the best fit with external variables. The four-cluster *K*-modes partition was chosen using the same reasoning. The four-cluster LCA partition was chosen because internal validation suggested that the two-cluster and four-cluster partitions were the best models, and for the purposes of comparison with the four-cluster *K*-means and *K*-modes partitions.

All cross-tabulations of the four-cluster partitions were significantly associated by the Chi-square test; however, they had poor agreement as measured by Hubert and Arabies adjusted Rand index (ARI) (Table 5.3). The *K*-modes cluster with the greatest proportion of its members in *K*-means Cluster 1 was labelled Cluster 1, and so forth. The ARIs for the partition comparisons were 0.370 for *K*-means vs. *K*-modes, 0.390 for *K*-means vs. latent class analysis, and 0.250 for *K*-modes vs. latent class analysis, where 0 is the expected ARI for agreement due to chance, 1 is the ARI for perfect agreement, and 0.65 is often considered "good" (Steinley, 2004).

Manifestations present in greater than 80% of the clusters members were considered strongly characteristic, manifestations present in 60 to less than 80% of a cluster's members were considered "moderately characteristic," and manifestations present in a higher percentage than in other clusters of a partition were considered "relatively characteristic." Characteristic manifestations for each four-cluster partition are presented in Table 5.4.

Although the partitions do not strongly agree with each other, there are similarities across the partitions (Table 5.5). Each partition has a cluster characterized mainly by arthritis, a cluster characterized by proteinuria, a cluster characterized by anti-RNP and other laboratory features, and a cluster characterized by anti-RNP and cutaneous features. Cluster 1 of each partition has only one characteristic feature, arthritis. In these non-specific clusters, no manifestations other than arthritis are present in more than 50% of cluster members, and no manifestations are present in appreciably higher proportions than in other clusters. Cluster 2 of each partition is characterized by anti-RNP, anti-DNA, and arthritis. *K*-means Cluster 2 is also characterized by anti-Sm antibodies, and *K*-modes Cluster 2 is also characterized by lymphopenia, leukopenia, and

low complement (Table 5.4).

Cluster 3 of each partition was characterized to some degree by cutaneous manifestations, arthritis, anti-RNP, anti-DNA, lymphopenia, and low complement. In K-means Cluster 3, alopecia was present in almost all cluster members, and malar rash, discoid rash, mucosal ulcers, photosensitivity, and Raynaud's syndrome were present in a higher percentage than in the other K-means clusters. In K-modes Cluster 3, alopecia and Raynaud's syndrome were moderately characteristic, and mucosal ulcers were present in a higher proportion than in the other K-modes clusters; other cutaneous manifestations were present in very slightly higher proportions than other clusters, and thus not considered "relatively characteristic." This cluster was also strongly characterised by pleuritis, a feature that was not characteristic of any other clusters (in any partition) by our definitions. LCA Cluster 3 had a relatively higher proportion of malar rash, discoid rash, and photosensitivity than the other LCA clusters. Cluster 4 of all partitions were characterized to varying degrees by the same five features: lymphopenia, low complement, proteinuria, thrombocytopenia, and anti-DNA antibodies.

5.3.4 "Candidate" subtypes.

Of the partitions of early lupus patients described in this paper, we recommend using the patient clusters from the four-cluster K-means partition as candidate early subtypes. This choice was partially based on the fit of models of clusters with clinical outcomes; the K-means four-cluster partition had the best fit with ESRD out of the three four-cluster partitions considered. This partition was also preferred over the K-modes and LCA four-cluster partitions because it had a larger number of "highly characteristic" manifestations (Table 5.4).

5.4 Discussion

We described three partitions (sets of clusters) of early lupus patients, some of which may be considered as candidate early lupus subtypes. In particular, we felt that a set of four clusters created by K-means cluster analysis warranted description in further studies. All partitions contained versions of a proteinuria/ thrombocytopenia/ hypocomplementemia cluster, a cutaneous cluster with high anti-RNP and anti-DNA, a cluster with high anti-RNP and anti-DNA and no clinical characteristics, and a non-specific cluster.

Studies formally exploring patient subtypes typically use one class discovery method to create a set of clusters or a latent model to serve as candidate subtypes. We took a different approach in this study, using several class discovery methods in parallel. We felt this approach to be appropriate because of the exploratory analysis setting, in which the characteristics of hypothetical clusters are unknown. Different class discovery methods create partitions by optimizing different criteria or fitting different models, and thus would be expected to create different partitions of the data for a given number of clusters. In the absence of knowledge about the characteristics of subphenotypes, it may be that the use of several class discovery methods increases the possibility of finding a partition suitable to serve as prototypic subphenotypes. Interestingly, the

four-cluster partitions created by each clustering method used in our study had similar characteristic (high-probability) features, yet there was relatively poor overlap in patient membership between the clusters with similar characteristics.

A strength of our study methods is that we used a multi-pronged approach to class discovery by using several class discovery methods. A related contribution of this work is our use of K-modes cluster analysis, a method that is advocated for presence/absence data (Leisch, 2006) but we have seldom seen used for class discovery in the medical setting. An important strength of our study is the study population, which included a wider breadth of cases than previous study populations. The study population was drawn from the Georgia Lupus Registry, which is population-based and includes cases that do not meet the ACR criteria.

Data was collected from medical records, so it is possible that there was incomplete ascertainment of symptoms and signs. The size of our study population was also limited by our decision to use laboratory-based variables, which were not available for many patients meeting our case definition of early lupus. We did not use several immunologic features of interest (lupus anticoagulant and anti-cardiolipin or aPL antibodies) due to the small percentage of patients for whom these tests were performed. Because the study population was taken from a population-based registry, information of the type that might be gathered in a cohort study was not available (for example, possible etiologic risk factors such as a family history of lupus.)

Several caveats must be made about the interpretation of our results. First, by describing candidate partitions with four clusters, we do not suggest that four is the "true" number of early lupus subtypes. Choosing the best partition generated by K-means or K-modes is an issue that is unresolved in the class discovery literature, and there are conflicting opinions on the best measures to use in choosing latent class models. Bootstrapping methods suggested that the four- and five-cluster K-means and K-modes partitions were not artifacts of the cluster analysis, and we chose to describe the four-cluster partitions as candidate partitions due to the small size of some clusters in the five-cluster partitions. With a larger data set, partitions with a greater number of clusters could be evaluated, and it might be found that partitions with a different number of clusters are suggested.

Conceptually, early lupus subtypes would have different etiologies and responses to treatment. Our choice of candidate partitions was made largely with internal validation measures, which only use information that was used for the cluster analysis. Ultimately, the meaningfulness or utility of proposed early lupus subtypes will depend on the characterization of differences in etiology, response to treatment, and outcomes in lupus patients from putative subtypes.

A logical next step is extension to other study populations. Partitions of interest could be created using methods similar to what was used in the partition-generating portion of this study. Then, strength of association of a partition's clusters with health outcomes such as damage accrual could be evaluated using regression

models with clusters as the main exposure, using measures of model fit could be used to compare the overall association of candidate partitions with outcomes. This two-step approach would allow similar evaluation of partitions generated by latent class analysis and ad-hoc clustering algorithms. Alternatively, extended forms of latent class could be used to jointly model latent classes with covariates and outcomes.

In conclusion, we described three partitions (sets of clusters) of early lupus patients that can be considered as possible prototypes of early lupus subtypes, and suggested one set of patient clusters for further exploration. Further study of these patient clusters, or candidate early lupus subtypes, may be of benefit for elucidation of SLE etiology and prediction of clinical outcomes.

5.5 Figures and tables.

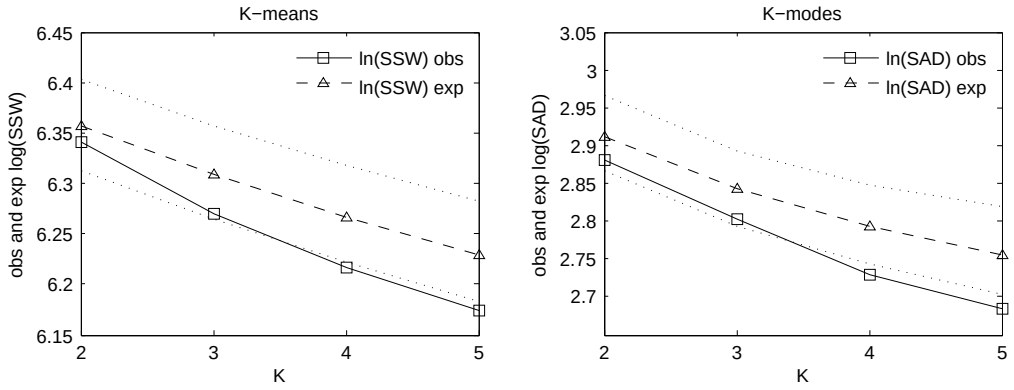

Figure 5.1: Clustering criteria for K-means (left plot) and K-modes (right plot) partitions compared to bootstrap null distributions. Solid line with squares depicts the observed criterion values, the dashed line with triangles depict the medians of the bootstrap null distributions, and the dotted lines depict the 2.5th and 97.5th %iles of the null distributions.

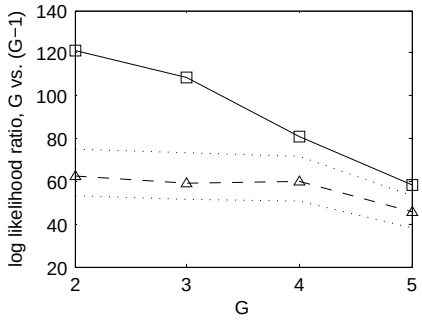

Figure 5.2: Bootstrap log likelihood ratio for latent class partitions, for G vs. $(G-1)$. Solid line with squares depicts the observed criterion values, the dashed line with triangles depict the medians of the bootstrap null distributions, and the dotted lines depict the 2.5th and 97.5th %iles of the null distributions.

Table 5.1: Model fit indices for latent class models with G classes. Minimum values are in boldface.

Latent class model, G classes	BIC2	CAIC2	DBIC	AIC3
$G = 1$	3972.69	3978.08	4021.79	4015.15
$G = 2$	3921.70	3932.63	**4021.23**	**4007.77**
$G = 3$	3883.58	3900.05	4033.54	4013.26
$G = 4$	**3873.41**	**3895.41**	4073.79	4046.70
$G = 5$	3885.27	3912.80	4136.08	4102.17

Table 5.2: Fit (BIC2) of logistic regression models with partitions. Adjusted model covariates: race, age at diagnosis, sex, and ESRD (for death as outcome).

	ESRD		Death	
	unadjusted	*adjusted*	*unadjusted*	*adjusted*
K-means partitions				
$K = 2$	53.89	**45.84***†	68.76	62.57
$K = 3$	53.61	48.68	69.65	63.65
$K = 4$	**52.86**	48.53	**66.97**	**61.59**
$K = 5$	57.70	52.18	67.94	62.66
K-modes partitions				
$K = 2$	60.22	54.14	68.67	62.32
$K = 3$	62.08	55.26	69.60	63.37
$K = 4$	**55.63**	**49.80**	**65.44**	**60.23**
$K = 5$	55.82	51.01	67.52	61.24
LCA partitions				
$K = 2$	55.11	50.01	**65.38***	**59.63***
$K = 3$	**50.73***	**46.75**	68.76	63.31
$K = 4$	57.17	51.40	67.09	61.63
$K = 5$	59.40	54.15	71.57	66.09

*Minimum value in column

†Boldface indicates minimum value in column within a clustering method.

Table 5.3: Cross-tabulations of four-cluster partitions. ARI is adjusted Rand index, P is from Chi-square test.

$ARI = 0.370$, $P < 0.001$

K-means clusters	K-modes C_1	K-modes C_2	K-modes C_3	K-modes C_4	
C_1	36	1	0	0	37
C_2	16	23	8	0	47
C_3	1	13	18	0	32
C_4	0	7	4	24	35
Total	53	44	30	24	151

$ARI = 0.390$, $P < 0.001$

K-means clusters	LCA C_1	LCA C_2	LCA C_3	LCA C_4	
C_1	24	12	1	0	37
C_2	0	35	10	2	47
C_3	1	4	23	4	32
C_4	1	7	0	27	35
Total	26	58	34	33	151

$ARI = 0.250$, $P < 0.001$

K-modes clusters	LCA C_1	LCA C_2	LCA C_3	LCA C_4	
C_1	24	25	3	1	53
C_2	0	25	14	5	44
C_3	1	7	17	5	30
C_4	1	1	0	22	24
Total	26	58	34	33	151

Table 5.4: Characteristic manifestations of four-cluster partitions.

	n	Strongly characteristic (present in $> 80\%$)	%	Moderately characteristic (present in $60 - 80\%$)	%	Relatively characteristic	%
***K*-means clusters**							
Cluster 1	37	Arthritis	84%				
Cluster 2	47	anti-RNP	96%	arthritis	77%		
		anti-DNA	70				
		anti-Sm	64				
Cluster 3	32	arthritis	84	anti-DNA	72%	malar rash	50%
		alopecia	97	anti-RNP	72%	discoid rash	4%
		lymphopenia	82			mucosal ulcers	47%
		low complement	84			photosensitivity	34%
						Raynaud's	50%
Cluster 4	35	lymphopenia	97	proteinuria	69%	seizures	20%
				thrombocytopenia	60		
				low complement	74		
				anti-DNA	74		
***K*-modes clusters**							
Cluster 1	53	Arthritis	84%				
Cluster 2	44	anti-RNP	96%	arthritis	77%		
				leukopenia	68		
				anti-DNA	80		
				anti-RNP	80		
				low complement	66		
Cluster 3	30	arthritis	83	alopecia	67%	mucosal ulcers	57%
		pleuritis	87	Raynaud's	73%		
		anti-DNA	83	lymphopenia	70		
				anti-Sm	77		
				anti-RNP	77		
				low complement	63		
Cluster 4	24	lymphopenia	96	proteinuria	75%	seizures	21%
		low complement	88	thrombocytopenia	71		
				low complement	74		
				anti-DNA	71		
LCA clusters							
Cluster 1	26			arthritis	77%		
Cluster 2	58			arthritis	79		
				anti-DNA	71		
				anti-RNP	67		
Cluster 3	34	arthritis	82%	alopecia	65%	malar rash	47%
				Raynaud's	62%	discoid rash	44%
				lymphopenia	74%	photosensitivity	53%
				anti-DNA	71		
				anti-Sm	62		
				anti-RNP	76		
				low complement	74		
Cluster 4	33	lymphopenia	97%	proteinuria	67	seizures	33
Cluster 4	33	low complement	91%	thrombocytopenia	64		
Cluster 4	33			anti-DNA	73		

Table 5.5: Common characteristics of four-cluster partitions.

	Cluster 1: No clear characteristics	**Cluster 2:** MCTD-like	**Cluster 3:** MCTD-like	**Cluster 4:**
Clinical characteristics	arthritis	arthritis alopecia Raynaud's phenomenon malar rash discoid rash	arthritis	
Autoantibodies		anti-RNP anti-DNA (anti-Sm)	anti-RNP anti-DNA (anti-Sm)	anti-DNA
Laboratory characteristics				proteinuria thrombocytopenia low complement

CHAPTER 6

Paper 4: Exploring early lupus subtypes in the Georgia Lupus Registry

Abstract

Objective. To describe clusters of early systemic lupus erythematosus (SLE) patients with similar signs and symptoms than can be considered as candidate early lupus subphenotypes.

Methods. The study population was incident patients from the Georgia Lupus Registry. K-means cluster analysis was used to create mutually exclusive clusters of patients with similar clinical and laboratory manifestations. Logistic regression was used to describe associations of cluster membership with death and ESRD while controlling for patient characteristics. Point-based classification scores were created to characterize "ease of classification."

Results. We described four clusters: a cluster of patients characterized by proteinuria, thrombocytopenia, and hypocomplementemia; a cluster characterized by mucocutenous features, lymphopenia, low complement, anti-RNP, and anti-DNA; a cluster characterized by anti-RNP and anti-DNA with no highly prevalent/shared clinical characteristics; and a cluster with no characteristic features (i.e., no features shared by most cluster members). The cluster characterized by proteinuria, thrombocytopenia, and hypocomplementemia and was strongly associated with death within 2 years of diagnosis and ESRD. Classification scores based on five features characteristically present or absent in each cluster demonstrated good discriminatory ability, with area under the ROC curves of 0.94 and higher.

Conclusion. The clusters of patients described suggest that there may be subphenotypes of early lupus.

6.1 Introduction

There is a wide variety in clinical and laboratory characteristics among patients diagnosed with SLE, and its etiology is poorly understood. Researchers in many fields of medicine have used or advocated the use of subset exploration to elucidate etiology in diseases that, like SLE, have wide inter-patient variety in disease manifestations. Edworthy and colleagues, among others, advocated the exploration of lupus subgroups as a route to identify etiological factors (Edworthy *et al.*, 1988).

There is a rich history of exploration of lupus subsets. In many studies, groups of patients delineated by clinical, immunologic, or demographic variables have been suggested as possible lupus subsets (Calvo-Alen *et al.*, 1995; Font *et al.*, 2004; Lee *et al.*, 2008). There are also examples of groups of patients with a collection of features being thought of as subtypes. For example, several recent reviews support the conceptualization of subacute cutaneous lupus erythematosus (SCLE) as a subset or subphenotype of LE characterized by widespread photosensitive, nonscarring skin lesions (Sontheimer, 2005; Werth, 2005).

Most investigations of possible lupus subsets have been based on expert clinical experience or statistical methods not considered to be formal class discovery methods, such as partition-based cluster analysis or latent class analysis. Several recent studies have used cluster analysis to explore subtypes of SLE by identifying subsets of patients with similar autoantibody profiles (To and Petri, 2005; Jurencak *et al.*, 2009). To our knowledge, no formal class discovery studies have examined adult SLE patients early in the disease course and using variables other than autoantibodies. In a previous paper, we described the implementation of several formal class discovery methods that could be used to create clusters of homogeneous patients, which in turn could be considered as models for early lupus subphenotypes. Using internal and external validation techniques, we found that a four-cluster partition created by K-means cluster analysis was a good candidate partition.

In this paper, we explain the method used to create this set of clusters, explore in more detail the relationships of cluster membership with patient characteristics and adverse events, explore the clinical implications of the clusters, and describe a series of point-based classification scores for cluster membership. These clusters of early lupus patients can serve as prototypes for early lupus phenotypes in future investigations.

6.2 Methods

6.2.1 Study population and variables

The study population was drawn from the Georgia Lupus Registry, a population-based registry designed to ascertain the prevalence in 2002 and incidence in 2002-2004 of SLE in Atlanta, Georgia. Details of the Georgia Lupus Registry are described elsewhere. Briefly, possible cases were ascertained by screening

medical records for ICD-9 diagnostic codes for SLE, discoid lupus erythematosus, antiphospholipid antibody syndrome, mixed connective tissue disease, and undifferentiated connective tissue disease. Demographic, clinical, and laboratory data were abstracted for patients with a physician-made diagnosis of "possible lupus" or "definite lupus."

The American College of Rheumatology (ACR) Criteria for the Classification of SLE classify a patient as having SLE if 4 or more of 11 criteria are met (Hochberg, 1997). For this study, the case definition of "early SLE" was the presence of 3 ACR criteria and diagnosis of SLE by a rheumatologist or 4 or more ACR criteria, with first diagnosis of possible or definite lupus from Janary 1, 2002 to December 31, 2004. Of the 234 patients who met the case definition, patients with race other than African-American or white were excluded ($n = 6$), and patients without measurements for laboratory variables to be used in class discovery analyses were excluded ($n = 77$), leaving a study population of 151 patients. Excluded patients were similar to the study population in the distributions of age of diagnosis, race, and sex (all P-values > 0.05). We used the following clinical and laboratory variables for cluster anlaysis: arthritis, myositis, malar rash, discoid rash, photosensitvity, mucosal ulcers, subacute cutaneous lupus erythematosus, panniculitis, alopecia, cutaneous vasculitis, Raynauds phenomenon, pleuritis, pericarditis, peritonitis, interstitial lung disease, pneumonitis, seizures, psychosis, mononeuritis multiplex, transverse myelitis, aseptic meningitis, chorea, peripheral or cranial neuropathy, leukopenia, lymphopenia, thrombocytopenia, proteinuria, ANA, anti-DNA, anti-Sm, anti-RNP, hemolytic anemia, antiphospholipid syndrome, and low complement. (Manifestations present in less than 5% of the study population and were not used for cluster analysis.)

We compared association of clusters with demographic variables, ESRD, and death within two years of diagnosis of lupus were used to characterize candidate partitions. Data on demographic variables, ESRD, and death were collected from medical records by the Georgia Lupus Registry.

In another paper, we described the application of several types of cluster analysis methods (K-means, K-modes, and latent class analysis) to create a variety of partitions (sets of clusters) of patients, and the use of cluster validation techniques to choose a "best" partition from each clustering method for further description (Speckman *et al.*, in preparation). Here, we describe the four-cluster partition created by K-means. While we describe this set of clusters as a candidate partition, we note that partitions not described in this paper might be of use in further explorations of lupus subtypes.

6.2.2 K-means cluster analysis

In K-means cluster analysis, K clusters of patients are created to maximize similarity of patients to their respective cluster means. The K-means algorithm minimizes the sum of within-cluster squared error, or the squared Euclidean distances from patients in a cluster to the cluster mean. A cluster's mean is a vector of

the proportion of patients in the cluster with each variable. For example, is 80% of the patients in a cluster have arthritis and 75% have anti-RNP antibodies, the cluster mean is $(0.80, 0.75)$ for those two variables. The squared Euclidean distance from a patient to a cluster mean is the sum over all variables of the squared difference between a variables cluster mean and the patients value for the variable. For example, a patient with arthritis but without anti-RNP antibodies would have the values of $(1,0)$ for these variables. The squared Euclidean distance from this patient to the cluster mean would be $((0.85-1)^2+(0.75-0)^2)$.

6.2.3 Characterizing clusters

Following the advice of Milligan (1996), we did not conduct statistical tests for differences between clusters in features that were used to create clusters. Instead, the clusters were described and compared by the raw percentage of cluster members with each feature. Crude association of cluster membership with external variables (variables not used in the cluster analysis) was described using Chi-square or Fishers exact tests of association and odds ratios for categorical variables (race, sex), and ANOVA for continuous variables (age at diagnosis). Association of cluster membership with ESRD, death within two years of diagnosis, and a combined outcome of ESRD or death within two years was assessed using logistic regression models.

6.2.4 Classification scores

We explored the ability of subsets of clinical and laboratory features to correctly predict cluster membership. If a small subset of features can be used to discriminate cluster membership, it could be considered to be indicative of the ease by which physicians might classify patients into the subgroups. Discriminatory ability was measured using the estimated area under the receiver operating characteristic (ROC) curve for logistic regression models in the following fashion. First, a model with the desired variable or set of predictor variables variables is fit, with the cluster as outcome. Then, for each member/non-member pair of patients, the predicted probabilities of membership in the cluster are calculated. If the predicted probability of membership in the cluster is higher for the patient that is actually in the cluster (the cluster member), the comparison function score is 1; if the predicted probability of membership in the cluster is higher for the patient that is not actually in the cluster, the comparison function score is 0; if the predicted probabilities are equal, the comparison function score is 0.5. The estimated area under the ROC curve (AUC_{ROC}) is the total of the comparison function scores for all member/non-members pairs over the total number of pairs.

We created classification rules for cluster membership using the features with the best discriminatory ability. For each cluster, we identified the five features with the highest AUC_{ROC}. Then, we evaluated point-based classification rules based on the 3 best, 4 best, or 5 best features. Each rule was created as follows: if a feature had a positive association with the cluster (i.e. was present in higher percentage in that cluster than in patients not in the cluster), a patient received a point for having the feature; if a feature had a negative

association with the cluster, the patient received a point for not having the feature. AUC_{ROC} for the score was evaluated by using the score as the single predictor in a logistic regression model.

6.3 Results

117 (77.5%) patients in the study population were African-American, and 23 (15.2%) were male. The mean age at diagnosis was 39.4 years (SE 1.34). 23 patients had three ACR criteria and a final diagnosis of SLE, and 128 patients had four or more ACR criteria. 110 patients had a final diagnosis of SLE, 27 SLE and another connective tissue disease (including 12 MCTD), 1 drug-induced SLE, 2 discoid lupus without systemic features, 3 chronic or subacute cutaneous non-discoid lupus, 6 other specified CTD (5 MCTD, 1 RA), 2 unspecified diffuse CTD. 70.9% of the study population had one or more mucocutaneous feature, and 51.7% had one or more mucocutaneous ACR criteria (malar rash, discoid rash, photosensitivity, or mucosal ulcers). 71.5 had arthritis, 39.1% had pleuritis, and 17.2% had one or more neuropsychiatric features. 59.6% had high titers of anti-DNA antibodies, 49.7% anti-RNP, and 35.8% anti-Sm.

Table 6.1 shows the characteristic features of the clusters. Cluster 1 had no characteristic features. Cluster 3 had an MCTD-like phenotype, with or relatively high probabilities of alopecia, arthritis, Raynaud's phenomenon, and anti-RNP antibodies. Cluster 2 had similar autoantibody profile as that of Cluster 3, with a high frequency of anti-RNP and medium-high probabilities of anti-DNA and anti-Sm antibodies. Cluster 4 had a high frequency of lymphopenia and medium-high frequencies of proteinuria, thrombocytopenia, low complement, and anti-DNA.

Cluster 1 had the lowest frequency of African-American members, 22 of the 37 (59.5%) (Table 6.2). Clusters 2 and 3 had the highest frequencies of African-American members (85.1 and 84.4%), both with significantly higher odds than Cluster 1 of being African-American. Cluster 2 had the highest frequency of male patients, ten of 47 members (21.3%). Cluster 1 had the highest mean age of diagnosis (47.0 years, SE 2.9), Clusters 2 and 3 had the lowest mean ages of diagnosis (35.7 years, SE 2.3 and 35.3 years, SE 2.9 respectively), and Cluster 4 had an intermediate age of diagnosis (40.3 years, SE 2.9).

Table 6.3 shows the frequency of ESRD and death occurring within 2 years of diagnosis, and "ESRD or death" combined as one variable. Both ESRD and death were significantly associated with cluster for all partitions, and the frequencies of ESRD and death were highest in Cluster 4 of each partition. Although there are too few adverse events in the data to draw further conclusions, it is interesting that adverse outcomes are distributed differently in the clusters with seemingly similar characteristics.

6.3.1 Point-based classification rules

Point-based classification rules using five features had high discriminatory ability for the clusters (Table 6.4). There are two types of features for the cluster classification scores: "positive" features, for which a point

is gained if the patient has the feature, and negative features, for which a point is gained if a patient does not have the feature. The area under the ROC curve was 0.96 for the Cluster 1 score, 0.94 for the Cluster 2 score, 0.96 for the Cluster 3 score, and 0.96 for the Cluster 4 score. The point-based Cluster 1 classification score has all negative features (anti-RNP, anti-DNA, anti-Sm, low complement, and leukopenia). The Cluster 2 classification score has two positive features (anti-RNP, anti-Sm), and three negative features (alopecia, lymphopenia, and mucosal ulcers). The Cluster 3 score has all positive features (alopecia, low complement, discoid rash, malar rash, anti-RNP). The Cluster 4 score has two positive features (proteinuria and thrombocytopenia) and three negative features (anti-RNP, anti-Sm, and arthritis). The sensitivity and specificity for different point cutoffs are shown for each classification score. For example, using three points as a cutoff for the Cluster 3 score, the sensitivity is 0.91 and the specificity is 0.94, with a positive likelihood ratio of 15.41. A patient with three or more of the five features would be classified as being in Cluster 3.

6.4 Discussion

We described clusters of early lupus patients, some of which can be considered as candidate early lupus subphenotypes. One cluster had an MCTD-like phenotype, with high or relatively high probabilities of alopecia, arthritis, Raynaud's phenomenon, and anti-RNP antibodies. Another cluster had a high probability of proteinuria, thrombocytopenia, and low complement. There were significant differences between clusters in ethnicity, sex, and age at diagnosis. The cluster characterized by proteinuria, thrombocytopenia, and hypocomplementemia was more likely to have ESRD or die within two years of diagnosis than patients in other clusters. Point-based classification scores for cluster membership based on five features had good discriminatory ability.

The prevalences of clinical and laboratory features in our study population were generally similar to reported prevalences at clinical presentation for patients in the Hopkins Lupus Cohort from 1960 to 1992, and in the Euro-lupus cohort (Cervera *et al.*, 1993; Petri *et al.*, 1993). An exception is malar rash, which was present in 21% of our study population and roughly 40 to 50% in these reports.

Average age of diagnosis in the study population was older than that reported in some studies; for example, the mean age of diagnosis in the EuroLupus cohort was 31 years old (Cervera *et al.*, 1993) One possible explanation is that the study population, which is taken from a population-based registry, probably captures certain types of patients that would not be included in a study population drawn from a (some) tertiary center(s). It may be that SLE patients who are less likely to be referred to an academic center on average have a later or more insidious onset of disease. Or, the study may capture patients that are more likely to have a delayed diagnosis of SLE, such as uninsured patients or patients with nonspecific symptoms. The median age of diagnosis for patient with managed care, HMO, or PPO coverage was 42.0, compared to a median age

of 33.2 for uninsured patients, suggesting that the latter is more likely to play a role in this study population. Other studies have reported subgroups of patients with differences in age of onset. Calvo-Alen *et al.* (1995) reported an average age of clinical diagnosis of roughly 35 for patients meeting the ACR criteria and roughly 45 for patients diagnosed with SLE but not meeting the ACR criteria. In the LUMINA study, age of onset, rate of accrual of ACR criteria, and time from onset to diagnosis differed between ethnic groups. Hispanic patients from Puerto Rico had an average age of onset of roughly 35 years compared to 31 years in Hispanic patients from Texas (Vila *et al.*, 2004). Caucasian patients had a later age of onset (roughly 41 years) than Hispanic and African-American patients (Alarcon *et al.*, 1999).

To and Petri (2005) used K-means cluster analysis to identify three autoantibody clusters in the Hopkins Lupus Cohort. They described a cluster with a high percentage of members with anti-Sm and anti-RNP autoantibodies relative to the other clusters, a cluster with a high or relatively high percentags of anti-dsDNA, anti-Ro, and anti-La autoantibodies, and a cluster with high percentages of anti-dsDNA and anticardiolipin autoantibodies and relatively high lupus anticoagulant. The pattern of characteristic features of Cluster 3 from each of the three candidate partitions of our study population was similar to that of the high anti-RNP/anti-Sm cluster described by To and Petri, which had a higher prevalence of mucocutaneous features (malar rash, discoid lupus, photosensitivity rash, oral ulcer, Raynauds phenomenon) and a lower prevalence of proteinuria, nephrotic syndrome, leukopenia, lymphopenia, and thrombocytopenia than other clusters. However, although high compared to other clusters from the same population, the prevalences of anti-Sm and anti-RNP in this cluster were much lower than in the anti-Sm/anti-RNP/mucocutaneous clusters in our candidate partitions. More than one-third of the study population used by To and Petri had longer than 10 years of follow-up after diagnosis with SLE, and more than two-thirds had four or more years of follow-up.

Jurencak *et al.* (2009) described three antibody-based clusters of pediatric-onset SLE patients; a cluster with a high proportion of dsDNA antibodies, a cluster with high proportions of anti-dsDNA, antichromatin, antiribosamal P, anti-U1RNP, anti-Sm, anti-Ro, and anti-La antibodies, and a cluster with a high proportion of anti-dsDNA, anti-RNP, and anti-Sm antibodies. The cluster with low percentages of all autoantibodies save anti-DNA was more likely to be Caucasion than other clusters. Similarly, each of our candidate partitions contained a cluster with a higher proportion of Caucasians than other cluster and relatively low prevalences (less than 30%) of anti-DNA, anti-Sm, and anti-RNP. Several candidate clusters were characterized by anti-RNP and Raynauds phenomenon, consistent with with reported associations of anti-RNP and Raynauds (Hoffman *et al.*, 2004).

We used both clinical and laboratory variables when performing cluster analysis, a different approach from previous class discovery efforts in SLE, and one that capitalizes on potential subtype information contained by clinical features. This may partially explain why there were greater differences in the presence of

some clinical features between clusters in our study than has been reported in cluster analyses using only autoantibodies.

Data was collected from medical records, so it is possible that there was incomplete ascertainment of symptoms and signs. The size of our study population was also limited by our decision to use laboratory-based variables, which were not available for some patients meeting our case definition of early lupus. We did not use several immunologic features of interest (lupus anticoagulant and anti-cardiolipin or aPL antibodies) due to the small fraction of patients for whom these tests were performed. Because the study population was taken from a population-based registry, information of the type that might be gathered in a cohort study was not available (for example, possible risk factors such as family history of lupus.)

Several caveats must be made about the interpretation of our results. First, by describing a candidate partition with four clusters, we do not suggest that four is the true number of early lupus subtypes. Choosing the best partition generated by K-means or K-modes is an issue that is unresolved in the class discovery literature, and there are conflicting opinions on the best measures to use in choosing latent class models. Bootstrapping methods suggested that the four- and five-cluster K-means and K-modes partitions were not artifacts of the cluster analysis, and we chose to describe the four-cluster partitions as candidate partitions due to the small size of some clusters in the five-cluster partitions. With a larger data set, partitions with a greater number of clusters could be evaluated, and it might be found that partitions with a higher/different number of clusters are suggested.

A logical next step is extension of this exploration to other study populations. Partitions of interest could be created using methods similar to what was used in the partition-generating portion of this study. Then, strength of association of a partitions clusters with health outcomes such as damage accrual could be evaluated using regression models with clusters as the main exposure, using measures of model fit could be used to compare the overall association of candidate partitions with outcomes. This two-step approach would allow similar comparison/evaluation of partitions generated by latent class analysis and ad-hoc clustering algorithms. Alternatively, extended forms of latent class analysis could be used to jointly model latent classes with covariates and outcomes.

6.5 Tables

Table 6.1: Characteristic manifestations of clusters.

	n	Strongly characteristic (present in $> 80\%$)	%	Moderately characteristic (present in $60 - 80\%$)	%	Relatively characteristic	%
Cluster 1	37	Arthritis	84%				
Cluster 2	47	anti-RNP	96%	arthritis	77%		
		anti-DNA	70				
		anti-Sm	64				
Cluster 3	32	arthritis	84%	anti-DNA	72%	malar rash	50%
		alopecia	97	anti-RNP	72	discoid rash	4
		lymphopenia	82	(anti-Sm)	(56)	mucosal ulcers	47
		low complement	84			photosensitivity	34
						Raynaud's	50
Cluster 4	35	lymphopenia	97%	proteinuria	69%	seizures	20%
				thrombocytopenia	60		
				low complement	74		
				anti-DNA	74		

Table 6.2: Association of clusters with patient characteristics. *P*-values from Chi-square test (race, Fisher's exact test (sex), ANOVA (age at diagnosis).

	n	African-American n (%)	OR (95% CI)	Male n (%)	OR (95% CI)	Age at diagnosis mean (SE)
Cluster 1	37	22 (59.5)	(Ref)	3 (8.1)	(Ref)	47.0 (2.9)
Cluster 2	47	40 (85.1)	3.90 (1.38, 11.00)	10 (21.3)	3.06 (0.78, 12.07)	35.7 (2.3)
Cluster 3	32	27 (84.4)	3.68 (1.16, 11.72)	5 (15.6)	2.10 (0.46, 9.58)	35.3 (2.9)
Cluster 4	35	28 (80.0)	2.73 (0.95, 7.85)	5 (14.3)	1.89 (0.42, 8.58)	40.3 (2.9)
P-value		0.024		0.003		0.009

Table 6.3: Association of clusters with clinical outcomes. Adjusted model covariates: race, age at diagnosis, sex, and ESRD for death as outcome. *P*-value from Fisher's exact test.

	ESRD *n* (%)	crude OR (95% CI)	adjusted OR (95% CI)	death *n* (%)	crude OR (95% CI)	adjusted OR (95% CI)
Cluster 1	0 (0)	0.28 (0.01, 7.45)	0.23 (0.00, 4.75)	2 (5.4)	(Ref)	(Ref)
Cluster 2	0 (0)	0.22 (0.01, 5.82)	0.23 (0.00, 4.46)	2 (4.3)	0.78 (0.12, 5.28)	0.83 (0.11, 6.43)
Cluster 3	1 (12.5)	(Ref)	(Ref)	0 (0.0)	0.22 (0.00, 2.82)	0.23 (0.00, 3.23)
Cluster 4	7 (20.0)	5.53 (0.87, 35.24)	4.71 (0.92, 47.24)	5 (14.3)	2.56 (0.57, 15.09)	2.64 (0.49, 17.35)
P-value	< 0.001			0.003		

	ESRD or death *n* (%)	crude OR (95% CI)	adjusted OR (95% CI)
Cluster 1	2 (5.4)	1.48 (0.19, 16.80)	1.29 (0.15, 15.37)
Cluster 2	2 (4.3)	1.15 (0.15, 13.05)	1.19 (0.15, 13.55)
Cluster 3	1 (3.1)	(Ref)	(Ref)
Cluster 4	11 (31.4)	9.86 (2.12, 95.33)	9.15 (1.92, 89.20)
P-value	< 0.001		

Table 6.4: Point-based classification scores for cluster membership.

	add point if feature present	add point if feature absent	Area under ROC curve	Points cutoff for membership	Sn	Sp	LR
Cluster 1		anti-RNP	0.96 (0.94, 0.98)	$= 5$	0.54	0.99	61.62
		anti-DNA		≤ 4	0.92	0.92	11.64
		anti-Sm		≤ 3	1.00	0.55	2.24
		low complement		≤ 2	1.00	0.29	1.41
		leukopenia		≤ 1	1.00	0.11	1.12
Cluster 2	anti-RNP	alopecia	0.94 (0.90, 0.98)	5	0.21	1.00	–
	anti-Sm	lymphopenia		4	0.68	0.97	23.60
		mucosal ulcers		3	0.98	074	3.77
				2	1.00	0.29	1.41
				1	1.00	0.02	1.02
Cluster 3	alopecia		0.96 (0.84, 1.00)	5	0.21	1.00	–
	low complement			4	0.47	1.00	–
	discoid rash			3	0.91	0.94	15.41
	malar rash			2	1.00	0.66	2.98
	anti-RNP			1	1.00	0.21	1.27
Cluster 4	proteinuria	anti-RNP	0.96 (0.94, 1.00)	5	0.14	1.00	–
	thrombocytopenia	anti-Sm		4	0.57	1.00	–
		arthritis		3	0.94	0.91	10.94
				2	1.00	0.55	2.23
				1	1.00	0.18	1.22

CHAPTER 7

Elaboration of results

7.1 Study 3

7.1.1 Relative validity index plots.

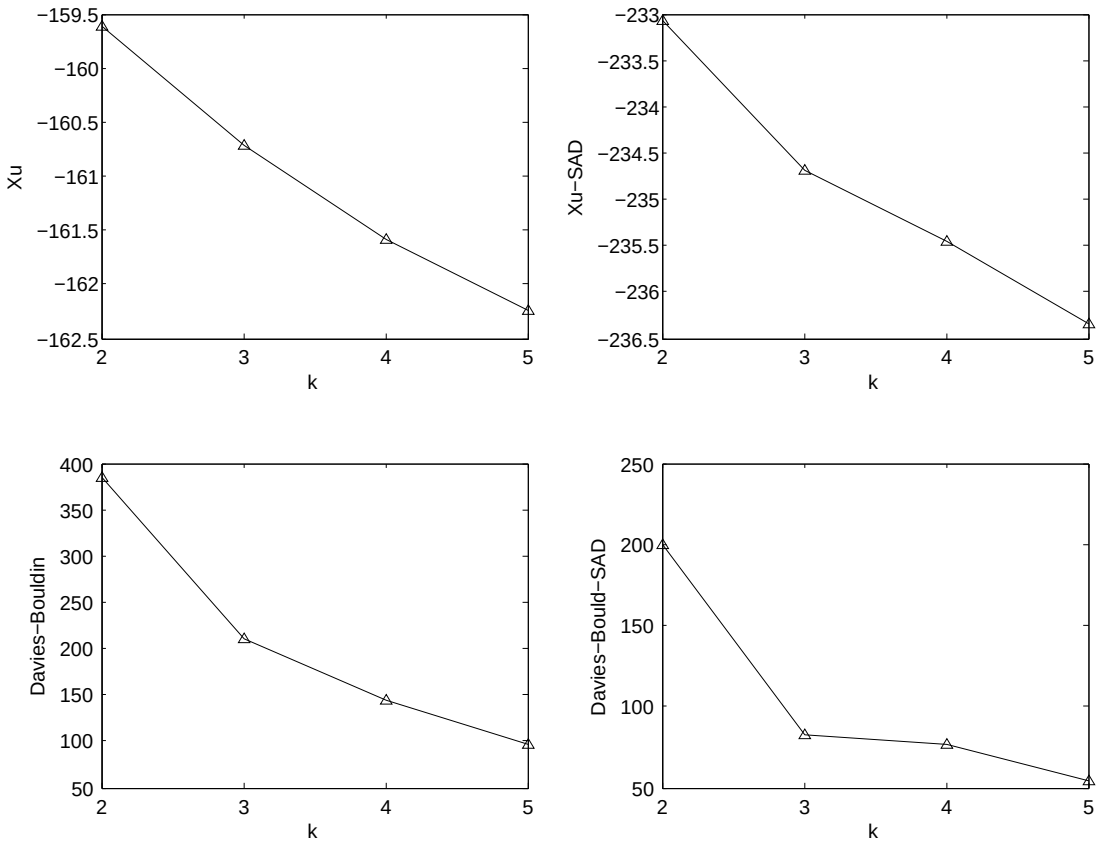

Figure 7.1: *K*-means partitions

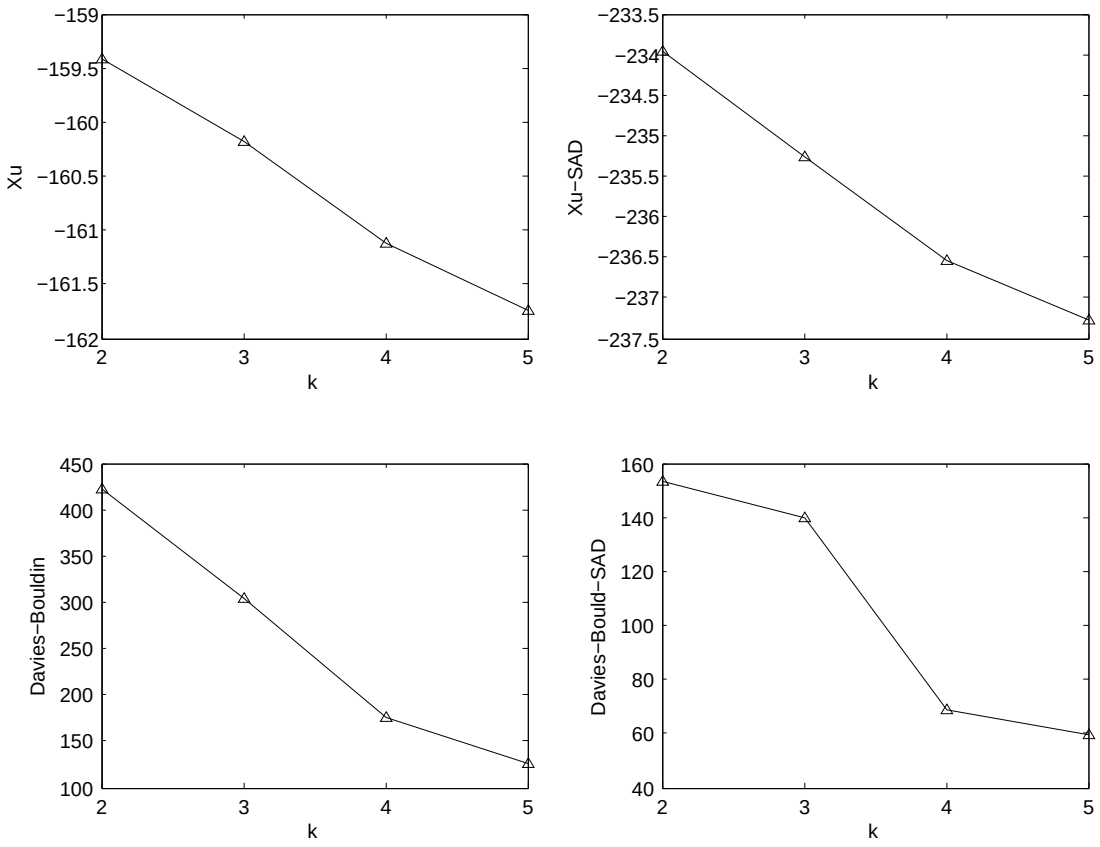

Figure 7.2: *K*-modes partitions

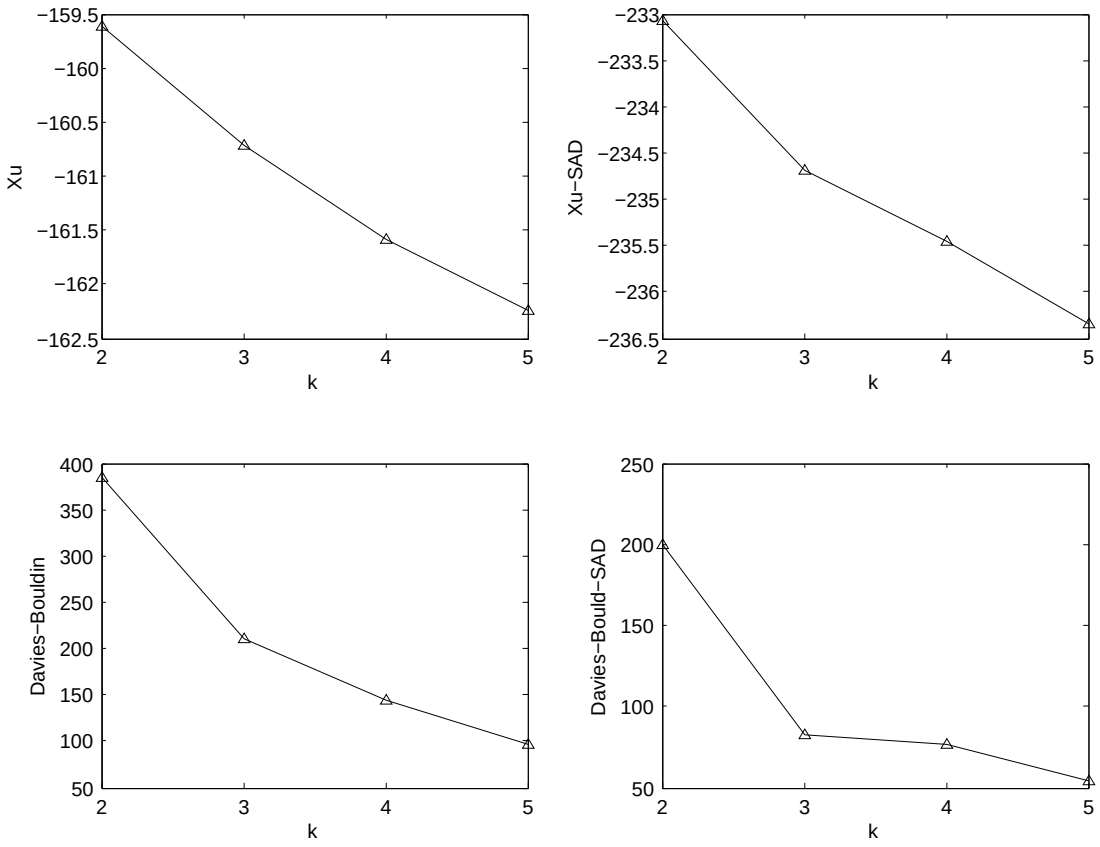

Figure 7.3: *K*-means partitions

7.1.2 Posterior probabilities of membership in latent classes.

The latent class model consists of the probability of manifestations in each cluster, and the probability of being in each cluster. From this, the probability of a patient being in each class (given their observed manifestations) is calculated: this is the posterior probability of membership in a class. To create mutually exclusive clusters, patients are assigned to the class for which they have the highest posterior probability of membership.

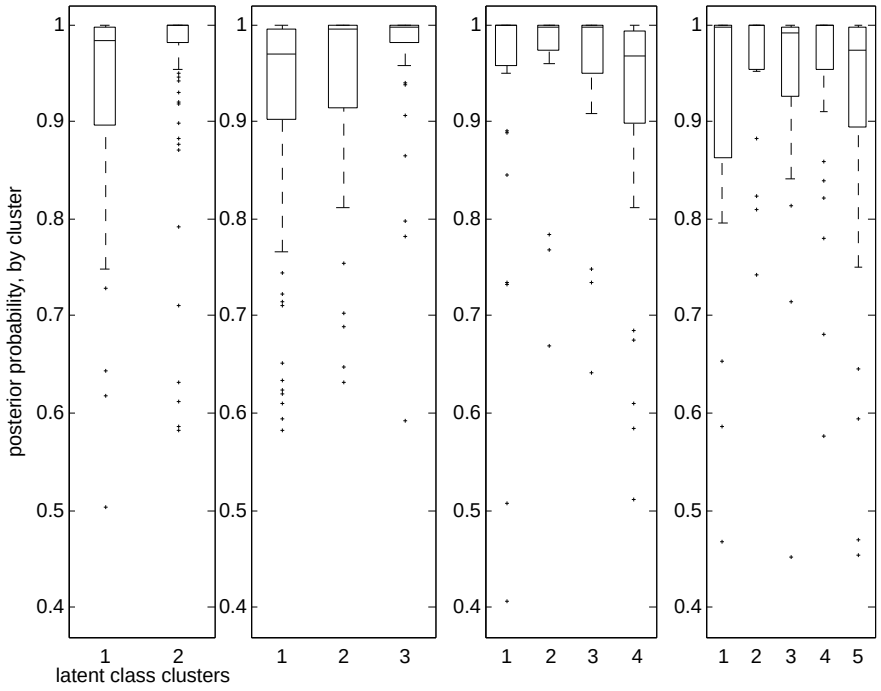

Figure 7.4: Latent class models: posterior probability of membership.

7.1.3 Frequency of features by cluster for four-cluster partitions

Table 7.1: Abbreviated feature names in tables

Abbrev-iation	Feature
Mal	malar rash
Dis	discoid rash
Pan	panniculitis
Alo	alopecia
CtV	cutaneous vasculitis
Ray	Raynaud's phenomenon
Liv	livedo reticularis
Urt	urticaria
Ph	photosensitivity
Muc	mucosal ulcers
Ser	serositis
Pl	pleuritis
Pc	pericarditis
Sz	seizures
Psy	psychosis
Mn	mononeuritis multiplex
TM	transverse myelitis
AsM	aseptic meningitis
Cho	chorea
NP	peripheral neuropathy
NC	cranial neuropathy
NN	neuropathy not specified
Art	arthritis
Myo	myositis
Leu	leukopenia
Lym	lymphopenia
Thr	thrombocytopenia
UPr	proteinuria
ANA	ANA
DNA	anti-DNA
Sm	anti-Sm
HA	hemolytic anemia
APS	antiphospholipid syndrome
Cmp	low complement
RNP	anti-RNP

K-means, four-cluster partition (selected features)

Table 7.2: Attributes, conditional probability of being present

C_k	n_k	Mal	Dis	Alo	Ray	Ph	Muc	Pl	Art	Leu	Lym	Thr	UPr	DNA	Sm
C_1	37	0.08	0.14	0.22	0.24	0.24	0.35	0.24	0.84	0.08	0.46	0.05	0.03	0.22	0.11
C_2	47	0.21	0.02	0.06	0.28	0.15	0.13	0.34	0.77	0.34	0.53	0.11	0.13	0.70	0.64
C_3	32	0.50	0.44	0.97	0.50	0.34	0.47	0.56	0.84	0.53	0.84	0.09	0.22	0.72	0.56
C_4	35	0.09	0.03	0.23	0.31	0.03	0.20	0.46	0.40	0.46	0.97	0.60	0.69	0.74	0.06

C_k	n_k	Cmp	RNP
C_1	37	0.14	0.00
C_2	47	0.36	0.96
C_3	32	0.84	0.75
C_4	35	0.74	0.17

Table 7.3: Bubble representation of conditional probability of attributes. Diameter of bubble is proportional to probability.

C_k	Mal	Dis	Alo	Ray	Ph	Muc	Pl	Art	Leu	Lym	Thr	UPr	DNA	Sm	Cmp	RNP
C_1	∘	∘	∘	∘	∘	○	∘	◯	∘	○	∘	-	∘	∘	∘	-
C_2	∘	-	∘	○	∘	∘	○	◯	○	○	∘	∘	◯	◯	○	◯
C_3	○	○	◯	○	○	○	◯	◯	○	◯	∘	○	◯	○	◯	◯
C_4	∘	-	○	○	-	○	◯	○	○	◯	◯	◯	◯	∘	◯	○

Table 7.4: Characteristic features. (Attributes in categories of frequency).

C_k	n_k	$p \geq 0.8$	$0.6 \leq p < 0.8$	$0.4 \leq p < 0.6$	$0.2 \leq p < 0.4$
C_1	37	Arth,		Lymph,	Alopec, Rayn, Photo, MucoUlc, Pleur, aDNA,
C_2	47	RNP,	Arth, aDNA, aSm,	Lymph,	Malar, Rayn, Pleur, Leuko, LowComp,
C_3	32	Alopec, Arth, Lymph, LowComp,	aDNA, RNP,	Malar, Disc, Rayn, MucoUlc, Pleur, Leuko, aSm,	Photo, UProtein,
C_4	35	Lymph,	Thrombo, UProtein, aDNA, LowComp,	Pleur, Arth, Leuko,	Alopec, Rayn, MucoUlc,

K-modes, four-cluster partition (selected features)

Table 7.5: Attributes, conditional probability of being present

		Conditional probabilities													
C_k	n_k	Mal	Dis	Alo	Ray	Ph	Muc	Pl	Art	Leu	Lym	Thr	UPr	DNA	Sm
C_1	53	0.13	0.11	0.21	0.25	0.19	0.25	0.28	0.83	0.09	0.38	0.08	0.04	0.25	0.26
C_2	44	0.27	0.16	0.30	0.23	0.18	0.16	0.18	0.73	0.68	0.89	0.14	0.20	0.80	0.36
C_3	30	0.33	0.23	0.67	0.73	0.30	0.57	0.87	0.83	0.30	0.70	0.13	0.30	0.83	0.77
C_4	24	0.12	0.04	0.25	0.17	0.04	0.17	0.42	0.29	0.33	0.96	0.71	0.75	0.71	0.04
C_k	n_k	Cmp	RNP												
C_1	53	0.11	0.28												
C_2	44	0.66	0.80												
C_3	30	0.63	0.77												
C_4	24	0.88	0.08												

Table 7.6: Bubble representation of conditional probability of attributes. Diameter of bubble is proportional to probability.

C_k	Mal	Dis	Alo	Ray	Ph	Muc	Pl	Art	Leu	Lym	Thr	UPr	DNA	Sm	Cmp	RNP
C_1	∘	∘	∘	∘	∘	∘	∘	◯	∘	∘	∘	-	∘	∘	∘	∘
C_2	∘	∘	∘	∘	∘	∘	∘	◯	◯	◯	∘	∘	◯	∘	◯	◯
C_3	∘	∘	◯	◯	∘	◯	◯	◯	∘	◯	∘	∘	◯	◯	◯	◯
C_4	∘	-	∘	∘	-	∘	∘	∘	∘	◯	◯	◯	◯	-	◯	∘

Table 7.7: Characteristic features. (Attributes in categories of frequency).

C_k	n_k	$p \geq 0.8$	$0.6 \leq p < 0.8$	$0.4 \leq p < 0.6$	$0.2 \leq p < 0.4$
C_1	53	Arth,			Alopec, Rayn, MucoUlc, Pleur, Lymph, aDNA, aSm, RNP,
C_2	44	Lymph,	Arth, Leuko, aDNA, LowComp, RNP,		Malar, Alopec, Rayn, UProtein, aSm,
C_3	30	Pleur, Arth, aDNA,	Alopec, Rayn, Lymph, aSm, LowComp, RNP,	MucoUlc,	Malar, Disc, Photo, Leuko, UProtein,
C_4	24	Lymph, LowComp,	Thrombo, UProtein, aDNA,	Pleur,	Alopec, Arth, Leuko,

LCA Partition four-cluster partition (selected features)

Table 7.8: Attributes, conditional probability of being present

		Conditional probabilities													
C_k	n_k	Mal	Dis	Alo	Ray	Ph	Muc	Pl	Art	Leu	Lym	Thr	UPr	DNA	Sm
C_1	26	0.15	0.23	0.27	0.31	0.31	0.50	0.23	0.77	0.04	0.50	0.08	0.08	0.04	0.04
C_2	58	0.12	0.00	0.17	0.17	0.02	0.05	0.33	0.79	0.33	0.57	0.10	0.17	0.71	0.45
C_3	34	0.47	0.44	0.65	0.62	0.53	0.47	0.50	0.82	0.53	0.74	0.06	0.12	0.71	0.62
C_4	33	0.15	0.00	0.33	0.30	0.03	0.27	0.52	0.42	0.42	0.97	0.64	0.67	0.73	0.18
C_k	n_k	Cmp	RNP												
C_1	26	0.12	0.00												
C_2	58	0.29	0.67												
C_3	34	0.74	0.76												
C_4	33	0.91	0.30												

Table 7.9: Bubble representation of conditional probability of attributes. Diameter of bubble is proportional to probability.

C_k	Mal	Dis	Alo	Ray	Ph	Muc	Pl	Art	Leu	Lym	Thr	UPr	DNA	Sm	Cmp	RNP
C_1	∘	∘	∘	∘	∘	○	∘	◯	-	○	∘	∘	-	-	∘	-
C_2	∘	-	∘	∘	-	∘	∘	◯	∘	○	∘	∘	◯	∘	∘	◯
C_3	○	○	◯	◯	○	○	○	◯	○	◯	∘	∘	◯	◯	◯	◯
C_4	∘	-	∘	∘	-	∘	○	○	○	◯	○	◯	◯	∘	◯	∘

Table 7.10: Characteristic features. (Attributes in categories of frequency).

C_k	n_k	$p \geq 0.8$	$0.6 \leq p < 0.8$	$0.4 \leq p < 0.6$	$0.2 \leq p < 0.4$
C_1	26		Arth,	MucoUlc, Lymph,	Disc, Alopec, Rayn, Photo, Pleur,
C_2	58		Arth, aDNA, RNP,	Lymph, aSm,	Pleur, Leuko, LowComp,
C_3	34	Arth,	Alopec, Rayn, Lymph, aDNA, aSm, LowComp, RNP,	Malar, Disc, Photo, MucoUlc, Pleur, Leuko,	
C_4	33	Lymph, LowComp,	Thrombo, UProtein, aDNA,	Pleur, Arth, Leuko,	Alopec, Rayn, MucoUlc, RNP,

K-means four-cluster partition, all features

Table 7.11: Conditional probabilities.

C_k	n_k	Mal	Dis	SC	Pan	Alo	CtV	Ray	Liv	Urt	Ph	Muc	Ser	Pl	Pc
C_1	37	0.08	0.14	0.03	0.03	0.22	0.03	0.24	0.03	0.11	0.24	0.35	0.00	0.24	0.14
C_2	47	0.21	0.02	0.02	0.00	0.06	0.09	0.28	0.00	0.02	0.15	0.13	0.02	0.34	0.19
C_3	32	0.50	0.44	0.00	0.03	0.97	0.16	0.50	0.03	0.06	0.34	0.47	0.00	0.56	0.16
C_4	35	0.09	0.03	0.03	0.00	0.23	0.17	0.31	0.06	0.09	0.03	0.20	0.00	0.46	0.14
C_k	n_k	Pn	Sz	Psy	Mn	TM	AsM	Cho	NP	NC	NN	Art	Myo	Leu	Lym
C_1	37	0.05	0.05	0.00	0.03	0.00	0.00	0.00	0.05	0.00	0.03	0.84	0.08	0.08	0.46
C_2	47	0.11	0.04	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.77	0.04	0.34	0.53
C_3	32	0.03	0.09	0.03	0.00	0.00	0.03	0.00	0.03	0.00	0.00	0.84	0.06	0.53	0.84
C_4	35	0.09	0.20	0.03	0.00	0.06	0.00	0.03	0.09	0.03	0.00	0.40	0.03	0.46	0.97
C_k	n_k	Thr	UPr	ANA	DNA	Sm	HA	APS	Cmp	RNP					
C_1	37	0.05	0.03	0.95	0.22	0.11	0.03	0.03	0.14	0.00					
C_2	47	0.11	0.13	1.00	0.70	0.64	0.11	0.06	0.36	0.96					
C_3	32	0.09	0.22	0.97	0.72	0.56	0.03	0.06	0.84	0.75					
C_4	35	0.60	0.69	0.97	0.74	0.06	0.17	0.14	0.74	0.17					

K-modes four-cluster partition, all features

Table 7.12: Conditional probabilities.

		Conditional probabilities													
C_k	n_k	Mal	Dis	SC	Pan	Alo	CtV	Ray	Liv	Urt	Ph	Muc	Ser	Pl	Pc
C_1	53	0.13	0.11	0.02	0.02	0.21	0.06	0.25	0.02	0.08	0.19	0.25	0.02	0.28	0.15
C_2	30	0.33	0.23	0.03	0.00	0.67	0.20	0.73	0.03	0.07	0.30	0.57	0.00	0.87	0.20
C_3	44	0.27	0.16	0.02	0.02	0.30	0.07	0.23	0.02	0.02	0.18	0.16	0.00	0.18	0.18
C_4	24	0.12	0.04	0.00	0.00	0.25	0.17	0.17	0.04	0.12	0.04	0.17	0.00	0.42	0.08
C_k	n_k	Pn	Sz	Psy	Mn	TM	AsM	Cho	NP	NC	NN	Art	Myo	Leu	Lym
C_1	53	0.08	0.06	0.00	0.02	0.00	0.00	0.00	0.06	0.00	0.02	0.83	0.06	0.09	0.38
C_2	30	0.10	0.13	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.83	0.00	0.30	0.70
C_3	44	0.05	0.05	0.05	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.73	0.09	0.68	0.89
C_4	24	0.08	0.21	0.04	0.00	0.08	0.00	0.04	0.12	0.04	0.00	0.29	0.04	0.33	0.96
C_k	n_k	Thr	UPr	ANA	DNA	Sm	HA	APS	Cmp	RNP					
C_1	53	0.08	0.04	0.96	0.25	0.26	0.06	0.06	0.11	0.28					
C_2	30	0.13	0.30	0.97	0.83	0.77	0.07	0.07	0.63	0.77					
C_3	44	0.14	0.20	1.00	0.80	0.36	0.07	0.07	0.66	0.80					
C_4	24	0.71	0.75	0.96	0.71	0.04	0.21	0.12	0.88	0.08					

LCA four-cluster partition, all features

Table 7.13: Conditional probabilities.

C_k	n_k	Mal	Dis	SC	Pan	Alo	CtV	Ray	Liv	Urt	Ph	Muc	Ser	Pl	Pc
C_1	33	0.15	0.00	0.03	0.00	0.33	0.21	0.30	0.06	0.12	0.03	0.27	0.00	0.52	0.18
C_2	26	0.15	0.23	0.04	0.00	0.27	0.04	0.31	0.08	0.15	0.31	0.50	0.00	0.23	0.00
C_3	58	0.12	0.00	0.02	0.03	0.17	0.02	0.17	0.00	0.00	0.02	0.05	0.02	0.33	0.19
C_4	34	0.47	0.44	0.00	0.00	0.65	0.21	0.62	0.00	0.06	0.53	0.47	0.00	0.50	0.21
C_k	n_k	Pn	Sz	Psy	Mn	TM	AsM	Cho	NP	NC	NN	Art	Myo	Leu	Lym
C_1	33	0.12	0.33	0.03	0.00	0.06	0.03	0.03	0.15	0.03	0.00	0.42	0.03	0.42	0.97
C_2	26	0.04	0.12	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.04	0.77	0.12	0.04	0.50
C_3	58	0.09	0.00	0.03	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.79	0.00	0.33	0.57
C_4	34	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.12	0.53	0.74
C_k	n_k	Thr	UPr	ANA	DNA	Sm	HA	APS	Cmp	RNP					
C_1	33	0.64	0.67	1.00	0.73	0.18	0.21	0.12	0.91	0.30					
C_2	26	0.08	0.08	0.85	0.04	0.04	0.04	0.08	0.12	0.00					
C_3	58	0.10	0.17	1.00	0.71	0.45	0.02	0.09	0.29	0.67					
C_4	34	0.06	0.12	1.00	0.71	0.62	0.12	0.00	0.74	0.76					

CHAPTER 8

Discussion

8.1 Summary and conclusions

The ultimate goal of my dissertation was to identify putative early lupus subtypes using patients from the Georgia Lupus Registry. I planned to use a type of class discovery method, K-means cluster analysis, to create homogeneous clusters of patients with similar signs and symptoms. In preparation for analysis of real patients, I explored two aspects of K-means cluster analysis relevant to the real data setting.

In my first study, I compared the performance of two K-means variations that had been recommended for presence/absence data (e.g. the presence or absence of a symptom), which is the type of real data that I would be using from the registry. I found that the two K-means variations recommended for presence/absence data did not outperform the standard K-means method. Using a benchmark data set of real data with known underlying disease subtypes and simulated data, I identified data set characteristics that resulted in worse performance by the cluster analysis methods.

In my second study, I compared the ability of a set of validity indices to choose the best partition from a group of partitions created by K-means or K-modes.

In my third study, I used K-means and K-modes cluster analysis to create clusters of early lupus patients from the Georgia Lupus Registry. Because I found in my second study that the recommended indices for choosing the best K-means or K-modes partition did not perform well, I also used latent class modeling to create partitions of the early lupus patients. (There are well-accepted methods for choosing the best latent

class model.) I described a set of four clusters created by K-means: one cluster had a phenotype similar to mixed connective tissue disease (MCTD), one cluster had laboratory features similar to MCTD but no characteristic clinical features, one cluster contained mostly patients with proteinuria, thrombocytopenia, and low complement, and one cluster did not have characteristic features. There were differences between the clusters in association with age, sex, ESRD, and death.

8.2 Strengths and Limitations

8.2.1 Cluster analysis methods evaluation

Strengths and contributions

A strength of my first study is that I objectively evaluated methods that were assumed (but not proven) to work best on the type of real data that I planned to use. A strength of my second study is that I explored extensions of recommended methods. I believe that the findings of my first two studies will be useful to researchers interested in using K-means or variations of K-means to explore subtypes in presence/absence data.

Weaknesses

The simulated data that I created in order to evaluate cluster analysis methods may have in fact limited the ability to detect differences in performance. For example, K-modes and K-means performed similarly on the simulated data, but had marked differences when applied to a real data set of overlapping erythematosquamous diseases. Another weakness is that there are many cluster analysis methods, and I only evaluated variations of one cluster analysis method. Because of the limitations in choosing the best number of clusters in K-means cluster analysis, it may be that model-based clustering methods such as latent class analysis are a better tool for subtype exploration.

8.2.2 SLE subtype exploration

Strengths and contributions

The clusters of early lupus patients that I identified and described in Study 3 may be good prototypes for early lupus subtypes. An important strength resulting from use of the Georgia Lupus Registry data is that the study population probably included milder cases of SLE and patients that would meet the ACR classification criteria for SLE but would not be diagnosed with SLE by an experienced rheumatologist, and thus included a wider breadth of cases than previous studies. Therefore, the study population met the recommendations of Nived and Sturfelt (2005), Ward (2005), and Symmons *et al.* (2006), who suggested that for studies meant to clarify classification and diagnostic criteria, the study population should include a broad range of unselected

patients.

Although subset exploration has long been part of SLE research, my study is one of the first to employ formal class discovery techniques. I believe that as a result of the knowledge gained from and by performing my first two studies, I was able to use cluster analysis in a more methodologically sound manner than previous cluster analysis studies of SLE and other rheumatologic diseases.

Weaknesses

A weakness of my third study is the limited amount of comparisons done between clusters for clinical outcomes. Another weakness is that the clusters with no characteristic clinical features may in fact be artifacts of the data source – it may be that limited medical records were available for these patients. Also, I eliminated patients with missing laboratory variables from the study population.

8.3 Future steps

8.3.1 Cluster analysis methods

I currently have no plans for future exploration of cluster analysis methods.

8.3.2 SLE early subtype exploration

Investigation of outcomes among the Georgia Lupus Registry-based early lupus clusters may give additional insight into the clinical utility of the putative subtypes. It would be interesting to explore the putative subtypes in a second population.

8.3.3 Future plans

After completing my clinical training, I will either pursue a clinical residency or move straight into academic research. Currently, I am interested in doing a physical medicine and rehabilitation residency. This field includes some disorders similar to SLE in that they are clinical syndromes encompassing patients with heterogenous symptoms. It is therefore possible that I may use the methods learned in my dissertation in future research, although I probably will pursue research based on clinical interest rather than a type of analytical method.

8.3.4 Personal gains from dissertation

Although I am not certain that will use the specific techniques from my dissertation in future work, the process of learning these techniques has been beneficial to me for several reasons. In broad terms, I increased my ability to understand statistics literature and implement new methods. I also learned several new skills that translate to other analytical methods. First, I learned how to generate artificial data. This skill will be useful if I do methods evaluation in the future, but is also applicable in other ways (such as Monte Carlo sensitivity

analyses of epidemiologic study results). Second, I did much of my work in the matrix-based MATLAB environment. I wrote routines from the ground up for a number of statistical or numerical methods, many of which are not available in commercial software. I may use this skill in the future to use newer statistical methods that I find to be appropriate, or I may simply be happy that I understand more of what SAS is doing behind the scenes, so to speak. Altogether, I feel more confident as an epidemiologist and user of statistical methods as a result of my dissertation work.

APPENDIX A

Additional methods background

A.1 Monte Carlo studies

Monte Carlo simulation is an important tool in the evaluation of class discovery procedures (Milligan, 1981b). By using simulated data, we can specify the true cluster structure underlying the data, and ascertain how a class discovery procedure recovers the true cluster structure. To do this, Hubert and Arabie's adjusted Rand index (described in a later section) is often used as the measure of agreement between the true clusters (true subtypes) and the clusters generated by the class discovery procedure.

A caveat of using simulation studies is the questionable generalizability of the results (Milligan, 1981b). Results from a simulation should not be assumed to extend to data structures that were not included in the simulation. On the other hand, the performance of a class discovery procedure in multiple simulation studies can provide insight into the robustness of the procedure under different data structures.

Monte Carlo simulation is the generation of a random sample from a "pseudo-population," in other words the creation of an artificial random sample. Two important uses of Monte Carlo simulations are (a) using simulated data to make estimates of mathematically intractable statistical quantities (b) using simulated data to learn about the performance of analytical methods (Mooney, 1997).

Estimation of statistical quantities. Monte Carlo simulations can be used to learn about the behavior of a test statistic in a random sample. The characteristics of some frequently used test statistics (e.g. sample mean) can be proved mathematically, but this is not the case for all test statistics. Manipulating characteristics

of the random sample, or the theoretical population from which it is drawn, can also help us understand the behavior of a test statistic.

Monte Carlo simulation can also be used to make estimates of other quantities that are impossible to solve for mathematically. For example, mathematically calculating the probability of having a random observation in a certain range of values, or the expected value and variance of a distribution, require integration of the density function. In our IQ example, the probability of an observation having an IQ less than 90 can be found by integrating the density function from $-\infty$ to 90. Such integration is intractable (meaning it cannot be done) for many probability distributions.

Because of the nature of statistical computing software (and the underlying computer programming), it is possible to generate random numbers from complex probability distributions (models) that are impossible to integrate or solve using statistical theory. Artificial data generated from a probability distribution can be used to make estimates about qualities of the probability distribution.

Evaluating analytical methods. Monte Carlo simulations are often used when it is desirable to have complete control over characteristics of data (data design factors), in order to study how data design factors influence the performance of an analytical method. Simulations provide a controlled data environment that is akin to an experiment rather than to observational studies. Importantly, creating simulated data allows us to characteristics of the underlying population or the data that are "unknowable" for real data.

A.1.1 Monte Carlo simulation procedure

Say that we are interested in describing the behavior of an estimator, $\hat{\theta}$, where $\hat{\theta}$ is some quantity that is calculated from observations. In inferential statistics terminology, we would like to know the *sampling distribution* of this estimator. The basic Monte Carlo procedure for constructing the sampling distribution of $\hat{\theta}$ is as follows (Mooney, 1997).

1. Specify a model for variables of interest in a pseudo-population. It is essential that this model can be computer-generated using some combination of random number generators, mathematical transformations, and fixed values.

2. Draw a sample (pseudo-sample) from the pseudo-population. The sampling method can parallel a real-life sampling scenario of interest, for example a certain sample size.

3. Calculate $\hat{\theta}$ for the sample.

4. Repeat Steps 2 and 3 t times, where t is the number of trials. (This number should be very large, on the order of thousands or millions.)

5. The relative frequency distribution of the t values for $\hat{\theta}$ is the Monte Carlo estimate of the sampling distribution of $\hat{\theta}$. This estimate of the sampling distribution is conditional on the model chosen for the pseudo-population and the sampling method.

This is best illustrated with a simple example. Let's say that we are interested in studying the behavior of a statistic $\hat{\theta}$ that is the sample median of the variable of interest (Y) divided by 2, under the conditions of a small sample. The Monte Carlo simulation procedure for this problem is:

1. Specify a model for variables of interest in a pseudo-population: In this example, we are interested in one variable and we will specify that the variable, Y, follows a normal distribution with $\mu = 30$ and $\sigma = 7$.

2. Draw a sample (pseudo-sample) from the pseudo-population: Because we are interested in the behavior of the estimator $\hat{\theta}$ with small sample sizes, we'll use a sample size of 25.

3. Calculate $\hat{\theta}$ for the sample.

   The following MATLAB program performs steps 1 through 3: (In the MATLAB programming language, the specification of the model/pseudo-population is integrated with drawing the pseudo-sample.)

```
mu = 30; sigma = 7;      %set parameters for Y
z = randn(25,1);         %draw a sample of 25 from N(0,1)
Y = (z .* sigma) + mu;   %transform to N(30,7)
Ymed = median(Y);        %this line and the next create theta hat
Ymed2 = Ymed ./ 2;
```

4. Repeat Steps 2 and 3 10,000 times. (10,000 is t, the number of trials.)

   The following MATLAB program performs this step:

```
for m = 1:10000;              %perform 10,000 times
{The code for Steps 1 to 3.}
thetahats(:,m) = Ymed2;       %save realizations in a 1 x 10000 array
end;
```

5. The histogram in Figure A.1 shows the relative frequency distribution of $\hat{\theta}$.

Random number generators. Most statistical software packages have the ability to generate random observations from the standard normal distribution $\mathcal{N}(0,1)$ and uniform distribution $\mathcal{U}(0,1)$. Observations from many other univariate distributions can be generated by transforming observations from one of these distributions.

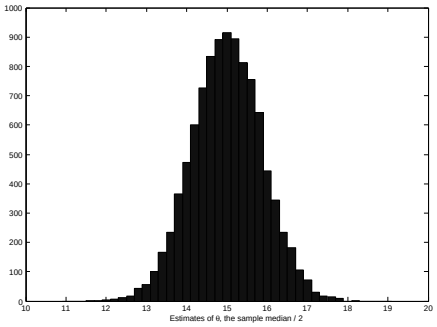

Figure A.1: Frequency distribution of $\hat{\theta}$, the sample median divided by 2.

Inversion method. When a $\mathscr{U}(0,1)$ random variable is transformed by the inverse cumulative distribution function (CDF) of a probability distribution F, the transformed random variable has the probability distribution F. (We will see in a following section that for observations drawn from a probability distribution F, the cumulative distribution function transforms the observations so that they follow a $\mathscr{U}(0,1)$ distribution.)

Figure A.2 shows a histogram from 10,000 observations generated from the $\mathscr{U}(0,1)$ distribution, then a histogram of those observations transformed into a $\mathscr{G}(3,1)$ variable. The data shown in the histograms was generated with the following MATLAB code:

```
U = rand(10000,1)      %generates 10,000 realizations from U(0,1)
X = gammainv(U,3,1);   %transform U to Gamma(3,1)
```

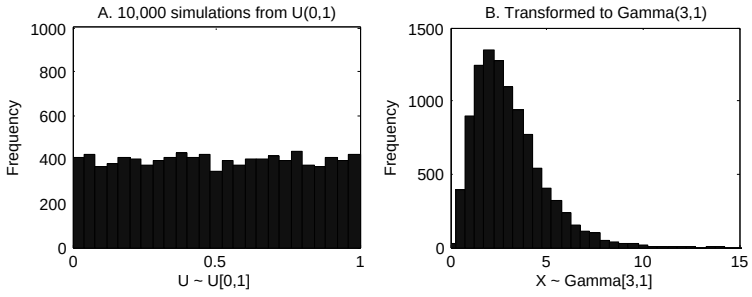

Figure A.2: $\mathscr{U}(0,1)$ random variable transformed into Gamma(3,1) random variable.

A.2 Data generation

Multivariate normal. It is relatively easy to generate data from a multivariate normal distribution, using programming tools that perform matrix operations rapidly.

To generate a multivariate normal distribution of p variables with dependence, we first generate p independent random variables from the standard normal distribution. Then, we transform the variables so that each has the desired mean and variance and so that any two variables have the desired linear correlation (ρ, rho). This transformation takes advantage of the following matrix algebra theory.

Recall from the section on multivariate normal distributions that the dependence between the variables can be described by two matrices, the covariance matrix (**cov**) and the correlation matrix (**cor**).

$$\mathbf{cov} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

$$\mathbf{cor} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$$

Because of the relationship between covariance and linear correlation (below), either of these matrices completely describes the dependence between the variables, and the covariance matrix also describes the variance of each variable.

$$\rho_{12} = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

The covariance matrix is actually a function of the correlation matrix and a vector of variances.

Cholesky decomposition. Covariance and correlation matrices have certain characteristics: they are all symmetric about the main diagonal and either positive definite or positive semi-definite. (The latter two characteristics are explained in Appendix A, but understanding of them is not necessary for this section.) By matrix theory, any symmetric positive definite matrix $\mathbf{B}$ can be "decomposed" into two triangular matrices $\mathbf{L}$ and $\mathbf{L^T}$, such that the product of $\mathbf{L}$ and $\mathbf{L^T}$ is $\mathbf{B}$. (The matrix $\mathbf{L^T}$ is the transpose of $\mathbf{L}$.)

$$\mathbf{B} = \mathbf{L}\mathbf{L^T}$$

Assume that the square matrix $\mathbf{B}$ is symmetric and either positive definite or positive semi-definite. $\mathbf{B}$

is symmetric, so $b_{21} = b_{12}$, $b_{n1} = b_{1n}$, etc. $\mathbf{L}$ and $\mathbf{L^T}$ are lower triangular and upper triangular matrices, respectively. (A lower triangular matrix is square with each element "above" the diagonal being 0, and an upper triangular matrix is square with each element "below" the diagonal being 0). The matrices of the expression $\mathbf{B} = \mathbf{LL^T}$ can be written as follows.

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}$$

$$\mathbf{LL^T} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{n1} \\ 0 & l_{22} & \cdots & l_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{nn} \end{bmatrix}$$

The matrices $\mathbf{L}$ and $\mathbf{L^T}$ of a matrix $\mathbf{B}$ are akin to the square root of a real number. The matrix $\mathbf{L}$ is often referred to as the "Cholesky decomposition" of $\mathbf{B}$, or sometimes the "Cholesky matrix" or "root matrix." *Cholesky decomposition* is actually the prevailing numerical algorithm used to decompose a positive definite (or semi-definite) matrix $\mathbf{B}$ into $\mathbf{L}$ and $\mathbf{L^T}$. In other words, the Cholesky numerical algorithm finds the matrices $\mathbf{L}$ and $\mathbf{L^T}$, such that $\mathbf{B} = \mathbf{LL^T}$.

Role of Cholesky matrices in generating MVN variables. Say that we would like to generate observations from a multivariate normal distribution of p variables, such that the vector **mu** is the means of each variables, the vector **var** is the variances of each variable, and the dependence between the variables is described by the correlation matrix **cor** and/or the covariance matrix **cov**.

Generating from standard MVN with desired correlation matrix. For the simplest scenario, say that we want to generate two variables, W^* and Z^*, from a standard MVN distribution with correlation ρ. In other words, $W^* \sim \mathcal{N}(0,1)$, $Z^* \sim \mathcal{N}(0,1)$, and W^* and Z^* are dependent with linear correlation ρ. The matrix $\mathbf{A}$ is the Cholesky decomposition of **cor**, such that $\mathbf{cor} = \mathbf{AA^T}$. The **cor** and $\mathbf{A}$ matrices are:

$$\mathbf{cor} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}$$

Here is the multiplication of $\mathbf{A}$ by its transpose, $\mathbf{A^T}$, resulting in the **cor** matrix.

$$\begin{aligned}\mathbf{AA^T} &= \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{bmatrix} \\ &= \begin{bmatrix} (1*1)+(0*0) & (1*\rho)+(0*\sqrt{1-\rho^2}) \\ (\rho*1)+(0*\sqrt{1-\rho^2}) & (\rho*\rho)+(1-\rho^2) \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\end{aligned}$$

Say that each column of the two-column $(n \times 2)$ matrix $[W\ Z]$ is n observations generated from the (univariate) standard normal distribution. In other words, $W \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{N}(0,1)$, and W and Z are independent.

$$\begin{bmatrix} W & Z \end{bmatrix} = \begin{bmatrix} w_1 & z_1 \\ w_2 & z_2 \\ \vdots & \vdots \\ w_n & z_n \end{bmatrix}$$

Multiplying the matrix $[W\ Z]$ by $\mathbf{A}$, the Cholesky decomposition of the correlation matrix that we wish to describe the variables W^* and Z^*, gives us a two-column matrix $[W^*\ Z^*]$ such that each column is still n observations from a standard normal distribution, but the columns are now observations from variables with the correlation ρ.

$$\begin{bmatrix} W^* & Z^* \end{bmatrix} = \begin{bmatrix} W & Z \end{bmatrix} \mathbf{A} = \begin{bmatrix} w_1 & z_1 \\ w_2 & z_2 \\ \vdots & \vdots \\ w_n & z_n \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix}$$

$$\begin{bmatrix} W^* & Z^* \end{bmatrix} = \begin{bmatrix} (w_1+z_1\rho) & z_1\sqrt{1-\rho^2} \\ (w_2+z_2\rho) & z_2\sqrt{1-\rho^2} \\ \vdots & \vdots \\ (w_n+z_n\rho) & z_n\sqrt{1-\rho^2} \end{bmatrix}$$

We have now generated data from a standard bivariate normal distribution with correlation ρ. The key matrix operation, the multiplication of a matrix of independent standard normal observations by the Cholesky decomposition of the desired correlation matrix, gives the same result for more than two variables.

Orientation of matrices. We have shown the matrices of n observations from a MVN distribution with p variables (or from p independent univariate normal distributions) as having dimensions $n \times p$, in other words with n rows and p columns. Frequently, this matrix operation is shown using $p \times n$ matrices for observations.

$$\begin{bmatrix} W \\ Z \end{bmatrix} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix}$$

If the matrices of observations are expressed in this manner, then the order of the matrices in multiplication must be reversed.

$$\begin{bmatrix} W^* \\ Z^* \end{bmatrix} = \mathbf{A} \begin{bmatrix} W \\ Z \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} w_1 & w_2 & \cdots & w_n \\ z_1 & z_2 & \cdots & z_n \end{bmatrix}$$

Generating from MVN with desired mean and covariance matrices. To transform observations from a standard normal variable Z into observations from a normally distributed variable X with mean μ and standard deviation σ, we use the following formula:

$$x_i = (z_i + \mu_X)\sigma_X$$

Now, we would like to generate two variables X and Y from a multivariate normal distribution. The matrix of desired means, **mu** is $[\mu_X \ \mu_Y]$ and the matrix of desired standard deviations is $[\sigma_X \ \sigma_Y]$, and the desired correlation is ρ.

We can obtain observations from X and Y by transforming the observations from random variables W^* and Z^*, which are dependent with linear correlation ρ that we desire to be shared between X and Y, in the following manner. First, recall that we used the Cholesky decomposition of the correlation matrix to obtain:

$$\begin{bmatrix} W^* & Z^* \end{bmatrix} = \begin{bmatrix} w_1^* & z_1^* \\ w_2^* & z_2^* \\ \vdots & \vdots \\ w_n^* & z_n^* \end{bmatrix} = \begin{bmatrix} (w_1 + z_1\rho) & z_1\sqrt{1-\rho^2} \\ (w_2 + z_2\rho) & z_2\sqrt{1-\rho^2} \\ \vdots & \vdots \\ (w_n + z_n\rho) & z_n\sqrt{1-\rho^2} \end{bmatrix}$$

Then, we perform the operation $x_i = (w_i^* + \mu_X)\sigma_X$ for each observation from W^*, and $y_i = (z_i^* + \mu_Y)\sigma_Y$ for each observations from Z^*. The following matrix operation performs this task:

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} = \begin{bmatrix} w_1^* & z_1^* \\ w_2^* & z_2^* \\ \vdots & \vdots \\ w_n^* & z_n^* \end{bmatrix} \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix} + \begin{bmatrix} \mu_X & \mu_Y \\ \mu_X & \mu_Y \\ \vdots & \vdots \\ \mu_X & \mu_Y \end{bmatrix}_{n\times 2}$$

This could also be written as

$$\begin{bmatrix} X & Y \end{bmatrix} = \begin{bmatrix} W^* & Z^* \end{bmatrix} \begin{bmatrix} \sigma_X & 0 \\ 0 & \sigma_Y \end{bmatrix} + \begin{bmatrix} \mu_X & \mu_Y \end{bmatrix}.$$

This gives us a matrix for X and Y with the following elements.

$$\begin{bmatrix} X & Y \end{bmatrix} = \begin{bmatrix} \sigma_X(w_1 + z_1\rho) + \mu_X & \sigma_Y z_1\sqrt{1-\rho^2} + \mu_Y \\ \sigma_X(w_2 + z_2\rho) + \mu_X & \sigma_Y z_2\sqrt{1-\rho^2} + \mu_Y \\ \vdots & \vdots \\ \sigma_X(w_n + z_n\rho) + \mu_X & \sigma_Y z_n\sqrt{1-\rho^2} + \mu_Y \end{bmatrix}_{n\times 2}$$

This method utilized the Cholesky decomposition of the correlation matrix, then transformation using the desired means and variances. Alternatively (and equivalently), X and Y can be found using the Cholesky decomposition of the covariance matrix.

The covariance matrix for X and Y is:

$$\mathbf{cov} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

$\mathbf{G}$ is the Cholesky decomposition of the covariance matrix, such that $\mathbf{cor} = \mathbf{G}\mathbf{G}^{\mathbf{T}}$. $\mathbf{G}$ for X and Y is:

$$\mathbf{G} = \begin{bmatrix} \sigma_X & 0 \\ \rho\sigma_Y & \sigma_Y\sqrt{1-\rho^2} \end{bmatrix}$$

As we did earlier, let us suppose that we have an $n \times 2$ matrix $[W\ Z]$ with each column drawn from the standard normal distribution. $W \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{N}(0,1)$, and W and Z are independent. Then, we can directly obtain X and Y using the following formula:

$$\begin{bmatrix} X & Y \end{bmatrix} = \begin{bmatrix} W & Z \end{bmatrix}\mathbf{G} + \begin{bmatrix} \mu_X & \mu_Y \end{bmatrix}$$

Showing the elements of the matrices, this is:

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} = \begin{bmatrix} w_1 & z_1 \\ w_2 & z_2 \\ \vdots & \vdots \\ w_n & z_n \end{bmatrix} \begin{bmatrix} \sigma_X & 0 \\ \rho\sigma_Y & \sigma_Y\sqrt{1-\rho^2} \end{bmatrix} + \begin{bmatrix} \mu_X & \mu_Y \\ \mu_X & \mu_Y \\ \vdots & \vdots \\ \mu_X & \mu_Y \end{bmatrix}_{n\times 2}$$

By carrying out these operations, we again arrive at:

$$\begin{bmatrix} X & Y \end{bmatrix} = \begin{bmatrix} \sigma_X(w_1+z_1\rho)+\mu_X & \sigma_Y z_1\sqrt{1-\rho^2}+\mu_Y \\ \sigma_X(w_2+z_2\rho)+\mu_X & \sigma_Y z_2\sqrt{1-\rho^2}+\mu_Y \\ \vdots & \vdots \\ \sigma_X(w_n+z_n\rho)+\mu_X & \sigma_Y z_n\sqrt{1-\rho^2}+\mu_Y \end{bmatrix}_{n\times 2}$$

If our preference is to express generated observations in $p \times n$ matrices, we could write the matrix operation as:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \mathbf{G}\begin{bmatrix} W \\ Z \end{bmatrix} + \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$

Summary. We have shown two equivalent routes to generating observations from a multivariate normal distribution with a known dependence structure.

A.2.1 Copulas

Although it is somewhat easy to generate observations from the multivariate normal distribution with dependence between the variables, the generation of simulated observations from dependent random variables is generally a difficult task. Copulas are a popular approach to this challenge.

Copulas are a class of multivariate distributions whose marginal distributions are uniform distributions on the interval $[0,1]$ (Genest and MacKay). A copula is more than simply a multivariate uniform distribution, however— a copula is a multivariate uniform distribution that has dependency, and that originates in a certain fashion. As we saw before, many distributions (for example, the Bernoulli distribution) can be generated from the uniform distribution. The utility of copulas is that we can obtain dependent variables from a variety of distributions by transforming the dependent uniform variables of a copula.

Generating copulas. Recall that for a continuous variable X_1, its cumulative density function (sometimes represented as Φ, or $F(X_1)$) calculates the probability of obtaining an observation less than a given possible value of X_1. In our example of the random normal variable IQ, the cumulative density function is:

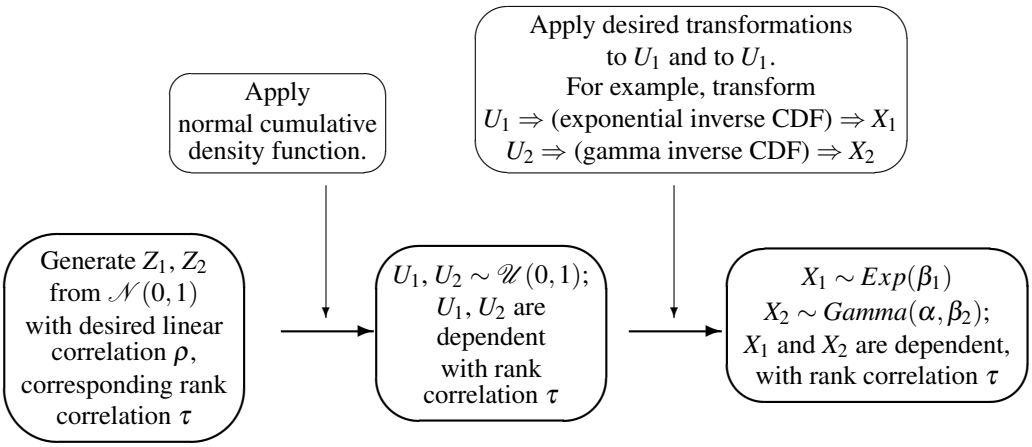

Figure A.3: Overview: generation of a Gaussian copula with two dimensions.

$$\Pr(X_1 < x_1) = \int_{-\infty}^{x_1} \frac{1}{\sqrt{20}\sqrt{2\pi}} e^{-\left(\frac{1}{2\times 20}\right)(x-99)^2}\, dx.$$

For data generated from any probability distribution, applying that distribution's cumulative density function to the data transforms the data in such a way as to make it uniform over the interval $[0,1]$. For example, Figure A.4a is a histogram of 10,000 observations generated from the standard normal distribution, $N(0,1)$. Figure A.4b is a histogram of the same data transformed by the normal CDF.

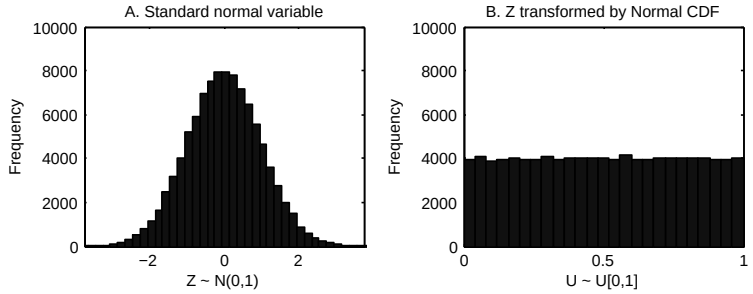

Figure A.4: Data generated from $N(0,1)$ transformed by Normal CDF.

Copula Example 1. We wish to generate observations from two distributions that are dependent: a gamma distribution with parameters $(\alpha = 3, \beta = 1)$ and an exponential distribution with parameter $(\beta = 2)$. We want the dependence between the two variables to be Kendall's τ of 0.7.

First, we find the linear correlation corresponding to a Kendall's τ of 0.7, using the formula shown previously. This linear correlation ρ is 0.89. Then, we generate two standard normal variables Z_1 and Z_2

with $\rho = 0.89$. In Figure A.5, the scatterplot between Z_1 and Z_2 illustrates their dependence. The histograms below and to the left the scatterplot show the respective marginal distributions for Z_1 and Z_2, which are both normally distributed.

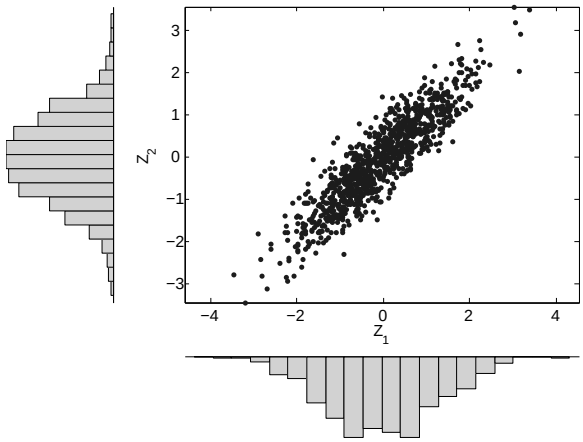

Figure A.5: Copula Example 1: U_1 and U_2; $\tau = 0.7$.

The next step is to use the normal CDF to transform Z_1 and Z_2 into variables that are uniform over the interval $[0,1]$. After this transformation, the dependency between the two sets of observations is preserved, and Kendall's *tau* remains equal to 0.7. Figure A.6 shows the scatterplot for the $\mathscr{U}(0,1)$ variables U_1 and U_2 with histograms of the marginal distributions.

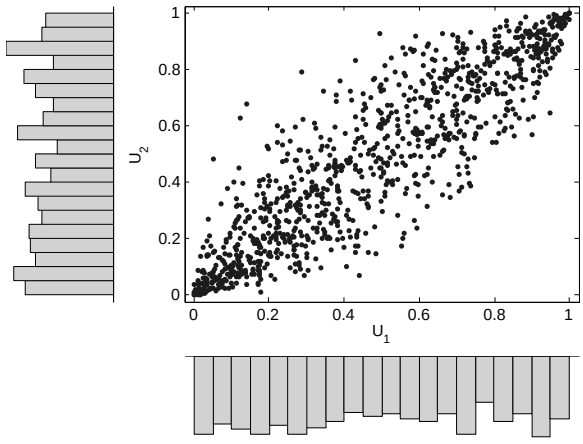

Figure A.6: Copula Example 1: U_1 and U_2; $\tau = 0.7$.

Finally, we create observations from the two desired distributions, exponential and gamma. To do this,

we apply the exponential inverse CDF to U_1 to obtain X_1, which follows an exponential distribution with a parameter of 2. We apply the gamma inverse CDF to U_2 to obtain X_2, which follows a gamma distribution with parameters $\alpha = 3$ and $\beta = 1$. Figure A.7 shows the scatterplot for X_1 and X_2 with histograms of the marginal distributions.

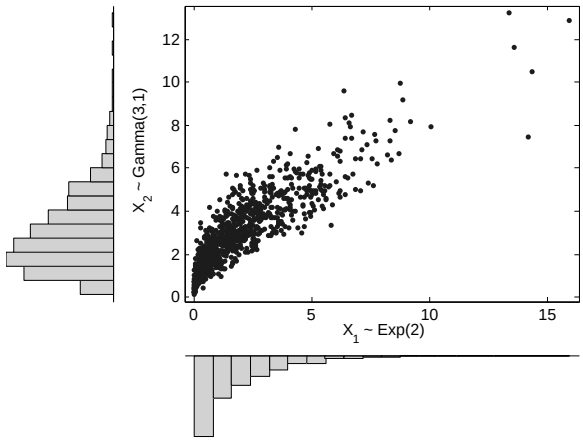

Figure A.7: Copula Example 1: U_1 and U_2; $\tau = 0.7$.

A.2.2 Simulation of dependent Bernoulli variables.

Dependent Bernoulli variables can be generated using a copula, or from similar methods. First, we will discuss generating observations from univariate (independent) Bernoulli variables.

Generating data from independent Bernoulli variables. The inversion method is a commonly used method to generate values from the Bernoulli distribution, a discrete distribution. Recall that a Bernoulli random variable can take on either the value 0 or 1, and the parameter p is the probability that an observation will have the value of 1. To transform observations from a $\mathscr{U}(0,1)$ distribution to a $\mathscr{B}ern(0,1)$ distribution, an observation is assigned the value of 1 if it is less than p and 0 otherwise. For example, if $p = 0.8$, all observations with $u_i \leq 0.8$ are assigned the value of 1, and the remaining observations are assigned the value of 0.

Similarly, observations from a binary (Bernoulli) variable A can be generated by first generating observations from a normally distributed variable (Z), then assigning the value of 1 for A if the value of Z is less than a certain cut-off point.

For example, we wish to generate observations from binary variable A, where $p_A = 0.75$. (Recall that p_A is the probability that $A = 1$). Using the inverse cumulative distribution function for the standard normal probability distribution (in MATLAB, comparable software, or an online calculator), we find that 75% of the

area under the density curve is to the left of 0.6745. Figure A.8 shows the standard normal density function with the area to the left of 0.6745 shaded.

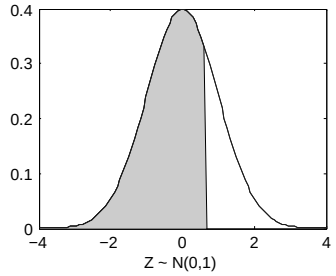

Figure A.8: Standard normal distribution: shaded area is $\Pr(Z < 0.6745) = 0.75$.

The following MATLAB code shows how we can generate observations with values of 0 and 1 drawn from a Bernoulli random variable A with $p_A = 0.75$. First, we generate 10,000 observations from the random normal distribution. If the value for Z is below 0.6745, then the observation is assigned the value of 1 for A ; otherwise, the observation is given a 0 for A.

```
%Generate 10,000 observations from random normal:
Z = normrnd(0,1,10000,1);

%Find value of Z such that Pr(Z<cutoffA) = 0.75:
cutoffA = norminv(0.75);

%Makes a template vector of 10,000 zeros:
A = zeros(10000,1);

%If the ith value of vector Z is less than cutoff, make A=1 for i=1:10000;
     if Z(i) < cutoffA
     A(i) = 1;          %Otherwise, A(i) will remain 0.
end
```

This table shows the first 7 observations generated. If the Z value generated for the i^{th} observation is below 0.6745, then that observation is assigned a 1 for the variable A. Otherwise, they are assigned a 0 for A.

generated observation (i)	Z value	A value
1	-1.3922	1
2	-0.9274	1
3	2.0773	0
4	-0.3294	1
5	0.9700	0
6	-0.0060	1
7	-2.0041	1

Figure A.9 shows a histogram of the values for Z for the observations generated by the code above.

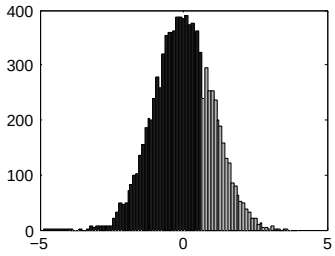

Figure A.9: 10,000 observations from standard normal distribution.

Generating data from dependent Bernoulli variables.

Numerous methods have been proposed to generate data from dependent Bernoulli variables, or dependent multivariate Bernoulli distributions (MVB) (Park *et al.*, 1996; Lunn and Davies, 1998; Kang and Jung, 2001). An important consideration when transforming data generated from continuous multivariate distributions is the ability of a method to create data with the desired correlation structure.

Emrich and Piedmonte. In Emrich and Piedmonte's method, data is first generated from a multivariate normal distribution, then transformed into binary data (Emrich and Piedmonte, 1991). For each combination of high-probability level and desired ρ of the binary variables, the necessary correlation for the multivariate normal distribution was determined using Equation (3) from Emrich and Piedmonte. (Because only one level of conditional dependence was used for each simulated dataset, this amounted to generating data from a MVN with an exchangeable correlation structure.) The data value for the m-th binary variable for the i-th individual, x_{im}, was assigned as "present" if the corresponding value generated from the multivariate normal distribution, z_{im}, was at or below a cutoff ϕ_m yielding $Pr(X < \phi_m) = 0.9$, 0.8, or 0.7, (the probability

of the attribute being present).

Bibliography

Al Attia, H. M. (2006). Borderline systemic lupus erythematosus (sle): a separate entity or a forerunner to sle? *International Journal of Dermatology* **45**, 4, 366–369.

Alarcon, G. S., Friedman, A. W., Straaton, K. V., Moulds, J. M., Lisse, J., Bastian, H. M., McGwin, G., Bartolucci, A. A., Roseman, J. M., and Reveille, J. D. (1999). Systemic lupus erythematosus in three ethnic groups: Iii a comparison of characteristics early in the natural history of the lumina cohort. *Lupus* **8**, 3, 197–209.

Alarcon, G. S., McGwin, G., Roseman, J. M., Uribe, A., Fessler, B. J., Bastian, H. M., Friedman, A. W., Baethge, B., Vila, L. M., and Reveille, J. D. (2004). Systemic lupus erythematosus in three ethnic groups. xix. natural history of the accrual of the american college of rheumatology criteria prior to the occurrence of criteria diagnosis. *Arthritis & Rheumatism-Arthritis Care & Research* **51**, 4, 609–615.

Albrecht, J., Berlin, J. A., Braverman, I. M., Callen, J. P., Costner, M. I., Dutz, J., Fivenson, D., Franks, A. G., Jorizzo, J. L., Lee, L. A., McCauliffe, D. P., Sontheimer, R. D., and Werth, V. P. (2004). Dermatology position paper on the revision of the 1982 acr criteria for systemic lupus erythematosus. *Lupus* **13**, 11, 839–849.

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press, New York,.

Arbuckle, M. R., James, J. A., Dennis, G. J., Rubertone, M. V., McClain, M. T., Kim, X. R., and Harley, J. B. (2003). Rapid clinical progression to diagnosis among african-american men with systemic lupus erythematosus. *Lupus* **12**, 2, 99–106.

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Ball, G. H. and Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science* **12**, 2, 153–&.

Baulieu, F. B. (1989). A classification of presence absence based dissimilarity coefficients. *Journal of Classification* **6**, 2, 233–246.

Bock, H.-H. (2007). A History of *k*-Means Algorithms. In *Selected Contributions in Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, 161–172. Springer Berlin Heidelberg.

Brennan, R. L. and Light, R. J. (1974). Measuring agreement when 2 observers classify people into categories. *British Journal of Mathematical & Statistical Psychology* **27**, NOV, 154–163.

Brusco, M. J. and Steinley, D. (2007). A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika* **72**, 4, 583–600.

Calinski, T. (1968). A dendrite method for cluster analysis. *Biometrics* **24**, 1, 207–&. Times Cited: 2.

Calvo-Alen, J., Bastian, H. M., Straaton, K. V., Burgard, S. L., Mikhail, I. S., and Alarcon, G. S. (1995). Identification of patient subsets among those presumptively diagnosed with, referred, and/or followed up for systemic lupus erythematosus at a large teriary care center. *Arthritis and Rheumatism* **38**, 10, 1475–1484.

CDC (1982). Update on acquired immunodeficiency syndrome (aids)–united states. *MMWR* **31**, 507–14.

CDC (1985). Current trends: revision of the case definition of acquired immunodeficiency syndrome for national reporting–united states. *MMWR* **34**, 25, 373–5.

Cervera, R., Khamashta, M. A., Font, J., Sebastiani, G. D., Gil, A., Lavilla, P., Domenech, I., Aydintug, A. O., Jedrykagoral, A., Deramon, E., Galeazzi, M., Haga, H. J., Mathieu, A., Houssiau, F., Ingelmo, M., Hughes, G. R. V., Cervera, R., Sebastiani, G. D., Font, J., Khamashta, M. A., Hughes, G. R. V., Font, J., Cervera, R., Lopezsoto, A., Vivancos, J., Ingelmo, M., Urbanomarquez, A., Khamashta, M. A., Vianna, J., Hughes, G. R. V., Gil, A., Lavilla, P., Pintado, V., Lopezdupla, M., Vazquez, J. J., Sebastiani, G. D., Deramon, E., Camps, M., Frutos, M. A., Perello, I., Santos, P. G., Abarca, M., Nebro, A. F., Domenech, I., Tokgoz, G., Aydintug, A. O., Jedrykagoral, A., Maldykowa, H., Chwalinskasadowska, H., Galeazzi, M., Haga, H. J., Mathieu, A., and Houssiau, F. (1993). Systemic lupus erythematosus: clinical and immunological patterns of disease expression in a cohort of 1,000 patients. *Medicine* **72**, 2, 113–124.

Clough, J. D., Elrazak, M., Calabrese, L. H., Valenzuela, R., Braun, W. B., and Williams, G. W. (1984). Weighted criteria for the diagnosis of systemic lupus erythematosus. *Arch Intern Med* **144**, 2, 281–5. 0003-9926 (Print) Journal Article.

Cohen, A. S., Reynolds, W. E., Franklin, E. C., Kulka, J. P., Ropes, M. W., Shulman, L. E., and Wallace, S. L. (1971). Preliminary criteria for the classification of systemic lupus erythematosus. *Bull Rheum Dis* **21**, 9, 643–648.

Costenbader, K. H., Karlson, E. W., and Mandl, L. A. (2002). Defining lupus cases for clinical studies: the boston weighted criteria for the classification of systemic lupus erythematosus. *J Rheumatol* **29**, 12, 2545–50. 0315-162X (Print) Evaluation Studies Journal Article.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **1**, 2, 224–227.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1, 1–38.

Deutsch, R., Cherner, M., and Grant, I. (2006). Significance testing of a cluster of multivariate binary variables: comparison of the tripartite t index to three common similarity measures. *Statistical Methods in Medical Research* **15**, 3, 285–299.

Dimitriadou, E., Dolnicar, S., and Weingessel, A. (2002). An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika* **67**, 1, 137–159.

Dubes, R. and Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition* **11**, 4, 235–254.

Dubois, E. L. (1966). *Lupus Erythematosus*.

Edwards, A. W. F. and Cavallis.Ll (1965). A method for cluster analysis. *Biometrics* **21**, 2, 362–&. Times Cited: 160.

Edworthy, S. M., Zatarain, E., McShane, D. J., and Bloch, D. A. (1988). Analysis of the 1982 ara lupus criteria data set by recursive partitioning methodology: new insights into the relative merit of individual criteria. *Journal of Rheumatology* **15**, 10, 1493–1498.

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *American Statistician* **45**, 4, 302–304.

Everitt, B. (1980). *Cluster analysis*. published on behalf of the Social Science Research Council by Heinemann Educational Books ; Halsted Press, London New York, 2nd edn.

Everitt, B. S. (1979). Unresolved problems in cluster analysis. *Biometrics* **35**, 1, 169–181.

Firth, D. (1993). Bias reduction of maximum-likelihood estimates. *Biometrika* **80**, 1, 27–38.

Font, J., Cervera, R., Ramos-Casals, M., Garcia-Carrasco, M., Sentis, J., Herrero, C., del Olmo, J. A., Darnell, A., and Ingelmo, M. (2004). Clusters of clinical and immunologic features in systemic lupus erythematosus: Analysis of 600 patients from a single center. *Seminars in Arthritis and Rheumatism* **33**, 4, 217–230.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing 2 hierarchical clusterings. *Journal of the American Statistical Association* **78**, 383, 553–569.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association* **62**, 320, 1159–&. Times Cited: 218.

Fries, J. F. (1987). Methodology of validation of criteria for sle. *Scand J Rheumatol Suppl* **65**, 25–30. 0301-3847 (Print) Journal Article.

Fries, J. F., Hochberg, M. C., Medsger, T. A., J., Hunder, G. G., and Bombardier, C. (1994). Criteria for rheumatic disease. different types and different functions. the american college of rheumatology diagnostic and therapeutic criteria committee. *Arthritis Rheum* **37**, 4, 454–62. 0004-3591 (Print) Journal Article Review.

Fries, J. F. and Siegel, R. C. (1973). Testing the 'preliminary criteria for classification of sle'. *Ann Rheum Dis* **32**, 2, 171–7. 0003-4967 (Print) Journal Article.

Gan, G., Ma, C., and Wu, J. (2007). *Data clustering : theory, algorithms, and applications*. SIAM American Statistical Association, Philadelphia, Pa. Alexandria, Va.

Ganczarczyk, L., Urowitz, M. B., and Gladman, D. D. (1989). "latent lupus". *J Rheumatol* **16**, 4, 475–8. 0315-162X (Print) Journal Article.

Geschwind, D. H. and Levitt, P. (2007). Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology* **17**, 1, 103–111.

Goodfellow, M. and Pirouz, T. (1982). Numerical classification of sporoactinomycetes containing meso-diaminopimelic acid in the cell-wall. *Journal of General Microbiology* **128**, MAR, 503–527.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* **49**, 268, 732–764.

Gower, J. C. and Legendre, P. (1986). Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification* **3**, 1, 5–48.

Greer, J. M. and Panush, R. S. (1989). Incomplete lupus-erythematosus. *Archives of Internal Medicine* **149**, 11, 2473–2476.

Guvenir, H. A., Demiroz, G., and Ilter, N. (1998). Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine* **13**, 3, 147–165.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. In *13th International Conference on Scientific and Statistical Database Management (SSDBM 2001)*, 107–145, Fairfax, Virginia.

Handl, J., Knowles, J., and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21**, 15, 3201–3212.

Happe, F., Ronald, A., and Plomin, R. (2006). Time to give up on a single explanation for autism. *Nature Neuroscience* **9**, 10, 1218–1220.

Hartigan, J. A. (1975). *Clustering algorithms*. Wiley, New York,.

Hill, R. S. (1980). A stopping rule for partitioning dendrograms. *Botanical Gazette* **141**, 3, 321–324.

Hochberg, M. C. (1993). The history of lupus erythematosus. *Lupus Foundation of America Newsletter* .

Hochberg, M. C. (1997). Updating the american college of rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis and Rheumatism* **40**, 9, 1725–1725.

Hoffman, I. E. A., Peene, I., Meheus, L., Huizinga, T. W. J., Cebecauer, L., Isenberg, D., De Bosschere, K., Hulstaert, F., Veys, E. M., and De Keyser, F. (2004). Specific antinuclear antibodies are associated with clinical features in systemic lupus erythematosus. *Annals of the Rheumatic Diseases* **63**, 9, 1155–1158.

Holubar, K. (1980). Terminology and iconography of lupus erythematosus. *American Journal of Dermatopathology* **2**, 239–242.

Huang, Z. X. (1998). Extensions to the *k*-Means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* **2**, 3, 283–304.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2**, 2-3, 193–218.

Hunder, G. G. (1998). The use and misuse of classification and diagnostic criteria for complex diseases. *Ann Intern Med* **129**, 5, 417–8. 0003-4819 (Print) Comment Editorial.

Iruela, M., Rubio, J., Cubero, J. I., Gil, J., and Millan, T. (2002). Phylogenetic analysis in the genus Cicer and cultivated chickpea using RAPD and ISSR markers. *Theoretical and Applied Genetics* **104**, 4, 643–651.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: A review. *Acm Computing Surveys* **31**, 3, 264–323.

Jones, E., Hodgins-Vermaas, R., McCartney, H., Everitt, B., Beech, C., Poynter, D., Palmer, I., Hyams, K., and Wessely, S. (2002). Post-combat syndromes from the boer war to the gulf war: A cluster analysis of their nature and attribution. *British Medical Journal* **324**, 7333, 321–324.

Jurencak, R., Fritzler, M., Tyrrell, P., Hiraki, L., Benseler, S., and Silverman, E. (2009). Autoantibodies in pediatric systemic lupus erythematosus: Ethnic grouping, cluster analysis, and clinical correlations. *Journal of Rheumatology* **36**, 2, 416–421.

Kang, S. H. and Jung, S. H. (2001). Generating correlated binary variables with complete specification of the joint distribution. *Biometrical Journal* **43**, 3, 263–269.

Kaufman, L. and Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge, ed., *Statistical data analysis based on the L b1 s-norm and related methods*, chap. Clustering by means of medoids, 405–416. North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co., Amsterdam ; New York New York, N.Y., U.S.A.

Klastorin, T. D. (1985). The para-median problem for cluster-analysis - a comparative test using the mixture model approach. *Management Science* **31**, 1, 84–95.

Klippel, J. H. (2001). *Primer on the rheumatic diseases*. Arthritis Foundation, Atlanta, GA, 12th edn. [edited by] John H. Klippel, MD ... [et al.]. cm.

Kogan, J. (2007). *Introduction to clustering large and high-dimensional data*. Cambridge University Press, Cambridge ; New York.

Lahita, R. G. (1987). *Systemic lupus erythematosus*. Wiley, New York. 86011007 edited by Robert G. Lahita ; with a foreword by Henry G. Kunkel. ill. ; 24 cm. A Wiley medical publication Includes bibliographies and index.

Lasky, T. and Stolley, P. D. (1994). Selection of cases and controls. *Epidemiologic Reviews* **16**, 1, 6–17.

Lee, S. S., Singh, S., Link, K., and Petri, M. (2008). High-sensitivity c-reactive protein as an associate of clinical subsets and organ damage in systemic lupus erythematosus. *Seminars in Arthritis and Rheumatism* **38**, 1, 41–54.

Leisch, F. (2006). A toolbox for *K*-centroids cluster analysis. *Computational Statistics & Data Analysis* **51**, 2, 526–544. DOI: 10.1016/j.csda.2005.10.006.

Liang, M. H., Corzillius, M., Bae, S. C., Lew, R. A., Fortin, P. R., Gordon, C., Isenberg, D., Alarcon, G. S., Straaton, K. V., Denburg, S., Esdaile, J. M., Glanz, B. I., Karlson, E. W., Khoshbin, S., Rogers, M. P., Schur, P. H., Hanly, J. G., Kozora, E., West, S., Lahita, R. G., Lockshin, M. D., McCune, J., Moore, P. M., Petri, M., Roberts, W. N., Sanchez-Guerrero, J., Veilleux, M., Brey, R., Cornblath, W. D., Filley, C. M., Fisk, J. D., Harten, P., Hay, E. M., Iverson, G., Levine, S. R., Waterhouse, E., Wallace, D. J., and Winer, J. B. (1999). The american college of rheumatology nomenclature and case definitions for neuropsychiatric lupus syndromes. *Arthritis and Rheumatism* **42**, 4, 599–608.

Lom-Orta, H., Alarcon-Segovia, D., and Diaz-Jouanen, E. (1980). Systemic lupus erythematosus. differences between patients who do, and who do not, fulfill classification criteria at the time of diagnosis. *J Rheumatol* **7**, 6, 831–7. 0315-162X (Print) Journal Article.

Lunn, A. D. and Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika* **85**, 2, 487–490.

MacQueen (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–297.

Mantel, N. (1967). Detection of disease clustering and a generalized regression approach. *Cancer Research* **27**, 2P1, 209–&.

Marriott, F. H. (1971). Practical problems in a method of cluster analysis. *Biometrics* **27**, 3, 501–&. Times Cited: 73.

McCutcheon, A. L. (1987). *Latent class analysis*. Sage University papers series. Quantitative applications in the social sciences. Sage Publications, Newbury Park.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics-Journal of the Royal Statistical Society Series C* **36**, 3, 318–324.

Milligan, G. W. (1981a). A Monte-Carlo study of 30 internal criterion measures for cluster analysis. *Psychometrika* **46**, 2, 187–199.

Milligan, G. W. (1981b). A review of Monte-Carlo tests of cluster analysis. *Multivariate Behavioral Research* **16**, 3, 379–407. Times Cited: 114.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. J. Hubert, and G. d. Soete, eds., *Clustering and classification*, 341–375. World Scientific, Singapore; River Edge, NJ.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 2, 159–179. I have hard copy.

Mooney, C. Z. (1997). *Monte Carlo simulation*. Sage university papers series. Quantitative applications in the social sciences. Sage Publications, Thousand Oaks, Calif. 96045873 Christopher Z. Mooney. ill. ; 22 cm. Includes bibliographical references (p. 99-110).

Morris, J. N. (1955). Uses of epidemiology. *British Medical Journal* **2**, AUG13, 395–401.

Morris, J. N. (1975). *Uses of epidemiology*. Churchill Livingstone : distributed in the U.S.A. by Longman, Edinburgh ; New York, 3rd edn.

Narain, S., Richards, H. B., Satoh, M., Sarmiento, M., Davidson, R., Shuster, J., Sobel, E., Hahn, P., and Reeves, W. H. (2004). Diagnostic accuracy for lupus and other systemic autoimmune diseases in the community setting. *Arch Intern Med* **164**, 22, 2435–41. 0003-9926 (Print) Journal Article.

Nived, O. and Sturfelt, G. (2005). Do we have blind spots in our diagnostic vision? *J Rheumatol* **32**, 1, 3–5. 0315-162X (Print) Comment Editorial.

Panush, R. S., Greer, J. M., and Morshedian, K. K. (1993). What is lupus - what is not lupus. *Rheumatic Disease Clinics of North America* **19**, 1, 223–234.

Park, C. G., Park, T., and Shin, D. W. (1996). A simple method for generating correlated binary variates. *American Statistician* **50**, 4, 306–310.

Petri, M., Caffentzis, E., Conroy, M., and Goldman, D. (1993). Clinical presentation of systemic lupus erythematosus (sle), 1960-92. *Arthritis and Rheumatism* **36**, 5, R22–R22. Suppl. S.

Petri, M. and Magder, L. (2004). Classification criteria for systemic lupus erythematosus: a review. *Lupus* **13**, 11, 829–37. 0961-2033 (Print) Journal Article Review.

Rand, W. M. (1971). Objective criteria for evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 336, 846–&.

Ratkowsky, D. A. and Lance, G. N. (1978). A criterion for determining the number of groups in a classification. *Australian Computer Journal* **10**, 115–117.

Sanchez, M. L., Alarcon, G. S., McGwin, G., J., Fessler, B. J., and Kimberly, R. P. (2003). Can the weighted criteria improve our ability to capture a larger number of lupus patients into observational and interventional studies? a comparison with the american college of rheumatology criteria. *Lupus* **12**, 6, 468–70. 0961-2033 (Print) Journal Article.

SAS (2009). Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

Schellenberg, G. D., Dawson, G., Sung, Y. J., Estes, A., Munson, J., Rosenthal, E., Rothstein, J., Flodman, P., Smith, M., Coon, H., Leong, L., Yu, C. E., Stodgell, C., Rodier, P. M., Spence, M. A., Minshew, N., McMahon, W. M., and Wijsman, E. M. (2006). Evidence for multiple loci from a genome scan of autism kindreds. *Molecular Psychiatry* **11**, 11, 1049–1060.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27**, 2, 387–&. Times Cited: 111.

Selim, S. Z. and Ismail, M. A. (1984). *K*-means-type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 1, 81–87.

Smith, E. L. and Shmerling, R. H. (1999). The american college of rheumatology criteria for the classification of systemic lupus erythematosus: strengths, weaknesses, and opportunities for improvement. *Lupus* **8**, 8, 586–95. 0961-2033 (Print) Journal Article.

Sontheimer, R. D. (2005). Subacute cutaneous lupus erythematosus: 25-year evolution of a prototypic subset (subphenotype) of lupus erythematosus defined by characteristic cutaneous, pathological, immunological, and genetic findings. *Autoimmunity Reviews* **4**, 5, 253–263.

Späth, H. (1980). *Cluster analysis algorithms for data reduction and classification of objects*. Computers and their applications. E. Horwood ; Halsted Press, Chichester, Eng. New York.

Späth, H. (1985). *Cluster dissection and analysis : Theory, FORTRAN programs, examples*. Horwood ; Halsted Press [distributor], Chichester New York.

Speckman, R. A., Drenkard, C., Klein, M., Bostick, R., and Lim, S. S. a. (in preparation). Using class discovery methods to explore early lupus subtypes in the georgia lupus registry. *In preparation* .

Stahl-Hallengren, C., Nived, O., and Sturfelt, G. (2004). Outcome of incomplete systemic lupus erythematosus after 10 years. *Lupus* **13**, 2, 85–8. 0961-2033 (Print) Journal Article.

Steinley, D. (2003). Local optima in *K*-means clustering: What you don't know may hurt you. *Psychological Methods* **8**, 3, 294–304.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods* **9**, 3, 386–396.

Steinley, D. (2006a). *K*-means clustering: A half-century synthesis. *British Journal of Mathematical & Statistical Psychology* **59**, 1–34.

Steinley, D. (2006b). Profiling local optima in *K*-means clustering: Developing a diagnostic techniques. *Psychological Methods* **11**, 2, 178–192.

Swaak, A. J., van de Brink, H., Smeenk, R. J., Manger, K., Kalden, J. R., Tosi, S., Marchesoni, A., Domljan, Z., Rozman, B., Logar, D., Pokorny, G., Kovacs, L., Kovacs, A., Vlachoyiannopoulos, P. G., Moutsopoulos, H. M., Chwalinska-Sadowska, H., Dratwianka, B., Kiss, E., Cikes, N., Anic, B., Schneider, M., Fischer, R., Bombardieri, S., Mosca, M., Graninger, W., and Smolen, J. S. (2001). Incomplete lupus erythematosus: results of a multicentre study under the supervision of the eular standing committee on international clinical studies including therapeutic trials (escisit). *Rheumatology (Oxford)* **40**, 1, 89–94. 1462-0324 (Print) Journal Article Multicenter Study.

Symmons, D. P., Lunt, M., Watkins, G., Helliwell, P., Jones, S., McHugh, N., and Veale, D. (2006). Developing classification criteria for peripheral joint psoriatic arthritis. step i. establishing whether the rheumatologist's opinion on the diagnosis can be used as the "gold standard". *J Rheumatol* **33**, 3, 552–7. 0315-162X (Print) Journal Article.

Tan, E. M., Cohen, A. S., Fries, J. F., Masi, A. T., McShane, D. J., Rothfield, N. F., Schaller, J. G., Talal, N., and Winchester, R. J. (1982). The 1982 revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum* **25**, 11, 1271–7.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **63**, 411–423.

To, C. H. and Petri, M. (2005). Is antibody clustering predictive of clinical subsets and damage in systemic lupus erythematosus? *Arthritis and Rheumatism* **52**, 12, 4003–4010.

Torrence, M. E. (1997). *Understanding epidemiology*, chap. Epidemiologic concepts of disease. Mosby's biomedical science series. Mosby, St. Louis. 96039467 Mary E. Torrence. ill., maps ; 24 cm. Includes bibliographical references and index.

Tyler, C. W. and Last, J. M. (1998). Epidemiology. In K. F. Maxcy, M. J. Rosenau, J. M. Last, and R. B. Wallace, eds., *Maxcy-Rosenau-Last public health & preventive medicine*. Appleton & Lange, Stamford, Conn., 14th edn.

Vila, L. M., Alarcon, G. S., McGwin, G., Friedman, A. W., Baethge, B. A., Bastian, H. M., Fessler, B. J., and Reveille, J. D. (2004). Early clinical manifestations, disease activity and damage of systemic lupus erythematosus among two distinct us hispanic subpopulations. *Rheumatology* **43**, 3, 358–363.

Vila, L. M., Mayor, A. M., Valentin, A. H., Garcia-Soberal, M., and Vila, S. (2000). Clinical outcome and predictors of disease evolution in patients with incomplete lupus erythematosus. *Lupus* **9**, 2, 110–5. 0961-2033 (Print) Journal Article.

Ward, Joe H., J. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58**, 301, 236–244.

Ward, M. M. (2005). Severity of illness in patients with systemic lupus erythematosus hospitalized at academic medical centers. *J Rheumatol* **32**, 1, 27–33. 0315-162X (Print) Journal Article.

Weiss, N. A. (2002). *Introductory statistics*. Addison-Wesley, Boston, 6th edn. 2001022689 Neil A. Weiss ; biographies by Carol A. Weiss. ill. (some col.) ; 26 cm. + 1 computer optical disc (4 3/4 in.) Includes indexes.

Werth, V. P. (2005). Clinical manifestations of cutaneous lupus erythematosus. *Autoimmunity Reviews* **4**, 5, 296–302.

Xu, L. (1997). Bayesian ying-yang machine, clustering and number of clusters. In *Pattern Recognition in Practice V Conference*, 1167–1178, Vlieland, Netherlands.

Xu, R. and Wunsch, D. C. (2009). *Clustering*. IEEE Press.

Yang, C. C. and Yang, C. C. (2007). Separating latent classes by information criteria. *Journal of Classification* **24**, 2, 183–203.