BURKE, MATTHEW JOSEPH, Ph.D. The Development of a Convergence Diagnostic for Markov Chain Monte Carlo Estimation. (2011)
Directed by Dr. Robert A. Henson. 282 pp.

The development and investigation of a convergence diagnostic for Markov Chain Monte Carlo (MCMC) posterior distributions is presented in this paper. The current method is an adaptation of an existing convergence diagnostic based on the Cumulative Sum (CUSUM, Page 1954; Yu & Mykland, 1998; Brooks, 1998c) procedure. The diagnostic under development is seen to be an improvement over the technique upon which it is based because it offers a simple way to remove one of the two major assumptions made by the previous method, namely that the shape of the distribution under consideration is symmetric. Results are mixed, but there is some evidence to indicate that the new technique is sensitive to the degree of autocorrelation present and the stability of the chains. Also, the new diagnostic behaves differently than three existing convergence diagnostics.

THE DEVELOPMENT OF A CONVERGENCE DIAGNOSTIC FOR MARKOV CHAIN MONTE CARLO ESTIMATION

by

Matthew Joseph Burke

A Dissertation Submitted to
the Faculty of the Graduate School at
the University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2011

Approved by

_____
Committee Chair

_____
Committee Co-Chair

To my mother, Anne Burke, without whose support this work would not have been possible.

APPROVAL PAGE

This dissertation has been approved by the following committee of the faculty of the Graduate School at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Co-Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Background

Bayesian approaches are commonly used in the fields of psychometrics and educational measurement (Levy and Mislevy, 2007). These approaches are based on Bayes' Theorem (Kim & Bolt, 2007) that uses probability distributions to characterize uncertainty about parameters of interest in the modeling of real world problems. Essentially, Bayesian approaches begin by stating prior beliefs (in the form of probability distributions) about characteristics of the parameters and then allow observed data to update those beliefs. The prior distributions (priors) and the observed data combine to form posterior distributions (PDs) to represent the updated information. The PDs (which are typically multivariate) are conditional probability distributions that represent the model parameters given the observed data. These PDs are then used to gain estimates of the location and dispersion of the parameters in much the same way as estimating population parameters from sample statistics (Patz & Junker, 1999a).

The Markov Chain Monte Carlo (MCMC; Patz & Junker, 1999a & 1999b) procedure is a Bayesian method of estimating model and person parameters that has been gaining popularity in psychometric modeling applications for nearly two decades (Albert, 1992). MCMC allows for the simulation of complex multivariate distributions by producing Markov chains that serve as the posterior distributions of the parameters of

interest (Chib & Greenberg, 1995). In particular, interest in applying the Metropolis-Hastings algorithm (MH; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Hastings, 1970, Green, 1995) has steadily gained momentum in recent years. This approach is extremely versatile and despite initially being confined primarily to use in the field of physics has begun to be adopted in other areas. Of particular concern in the current paper is the promise this method holds for the field of psychometrics. Patz and Junker (1999a, 1999b) showed how this method of estimating parameters can be applied in complex psychometric modeling applications. Following this work, many researchers have begun to apply MCMC sampling techniques for problems faced in testing applications (for example, Kim & Bolt, 2007; De la Torre, Stark, & Chernyshenko, 2006; Sinharay, 2004; McLeod, Lewis, and Thissen, 2003; Glas and Meijer, 2003; Fox and Glas, 2001; Beguin and Glas, 2001; is a brief list).

An MCMC technique is an alternative to the Expectation-Maximization (EM) algorithm used in marginal maximum likelihood estimation (MMLE) approaches (Bock and Aiken, 1981), for example. While MMLE is widely accepted for use in high stakes situations, it is not always possible to extend the approach to new models. An advertised benefit of applying the MCMC procedure is that it can potentially resolve the intractability of parameter estimation for complex psychometric models by way of the generally accepted maximum likelihood procedure (Patz & Junker, 1999a).

However, there is still some hesitancy among experts and practitioners concerning whether or not MCMC procedures are acceptable approaches to estimation in psychometrics. For example, in operational settings MCMC is still often seen as less

desirable than traditional estimation techniques for largely practical reasons such as estimation time and supporting research (Luecht, 2010, personal communication). A more pressing theoretical concern for the use of MCMC has to do with the quality of the estimates obtained from MCMC procedures. Regarding the quality of the estimates, probably the most difficult fact to ascertain when employing an MCMC estimation procedure is whether or not the chains have converged to stable posterior distributions. The E-M algorithm has a similar problem. It is not always easy to tell whether the solution is just a local optimum.

Convergence is directly related to the stability and trustworthiness of model parameter estimates obtained. In MCMC, ideally it would be the case that the values used for the estimation of parameters are indistinguishable from random draws from a posterior distribution that accurately characterizes the parameter being estimated given the observed responses. Thus, it seems reasonable to expect the entries in the posterior distribution to act like values sampled directly from a given distribution (i.e., these are converged by definition).

Previous authors have demonstrated the complications faced when trying to characterize convergence of MCMC samplers (see Cowles and Carlin, 1996; Sinharay, 2004 for summaries). As there is still currently no well accepted a priori method for determining how many iterations of a sampler are needed to produce converged chains, researchers must rely on ad hoc convergence diagnostics to evaluate the output of MCMC samplers. There are numerous diagnostic methods available to assess convergence of PDs. These convergence diagnostics rely on various strategies of

characterizing the output of MCMC samplers (Cowles & Carlin, 1996). The two general categories of convergence diagnostics are visual and quantitative.

Visual inspection techniques produce graphical representations of the Markov chains (or some transformation of them). These graphical representations are then 'visually inspected' to determine if there are any obvious violations of what would be expected if the chain was indeed converged. For example, if there is a continually increasing trend in the values in the Markov chain, this would be an indication that the process has not settled to a stable location. Visual inspection is a commonly used and useful technique. The primary appeal of the technique is its ease of implementation. Visual inspection of convergence is inherently subjective in that convergence is decided by plotting the values of the posterior distribution and seeing if the distribution 'looks' converged. This method is less desirable than a situation in which we have a more objective criterion to provide evidence of convergence (or lack thereof) for a Markov chain.

Quantitative indices are based upon some underlying theory or framework which describes the behavior of the chain or multiple chains produced for each parameter being estimated. The end result is a numerical value that indicates convergence or lack thereof when compared to a criterion. Quantitative techniques range from simple to complex in terms of calculation and formulation, focus on bias or variance (or both), can apply to different types of MCMC samplers, and are based on any of several different characterizations (e.g., large sample normal theory, spectral analysis, etc). In this study, the quantitative indices that will receive direct consideration are the Raftery and Lewis

(RL; 1992) diagnostic, the Geweke (G; 1992) diagnostic, and the Heidelberger and Welch (HW; 1983) diagnostic. These methods will be discussed in detail in the Literature Review. One purpose of the current research is to compare these methods to a new method under development.

In a comparison study, Cowles and Carlin (1996) described 13 different convergence diagnostics. These methods are described (e.g., visual versus quantitative indicators) and explained in enough detail so as to allow for their differences to be evident. Each of these methods for assessing convergence is then compared in different estimation settings. The authors report a common finding of how the diagnostics often disagree with one another. While the work of these authors is a thorough treatment of the diagnostic techniques, the examples to which the methods are applied are not psychometric models so the inferences to be drawn by educational researchers may be limited. Comparisons of convergence diagnostics in psychometric examples may be informative in this regard.

More recently, Sinharay (2004) summarized and reviewed five methods of assessing convergence in the context of psychometric models. Methods were chosen that were both easy to understand and implement, so as to foster a greater understanding of how to assess convergence and why convergence is of such great importance. The contribution of Sinharay's (2004) work to the field is straightforward; if MCMC is to gain even more momentum, psychometricians must make themselves aware of knowledge that allows for proper use of the MCMC technique. While this study addresses convergence in the context of psychometric models, there are some techniques

that it does not include which may be of interest to readers. Investigation of other methods than those included in this work is necessary for further implementation of MCMC methods.

When investigating convergence, a common conclusion is that different diagnostics tend to have different strengths and weaknesses (see Cowles and Carlin, 1996, and Sinharay, 2004, for examples), hence, there is a great deal of disagreement among the methods when it comes to making claims of convergence. The commonsense recommendation of these authors is to use multiple diagnostics sensitive to various violations of convergence. Multiple indices of both general types (visual and quantitative) of diagnostics should be produced to provide assurance that a chain is suitable for use in estimation. In light of these facts, it may be advantageous to explore the possibility of developing and/or refining new techniques that offer the opportunity to do the job of assessing convergence in situations that are problematic for existing methods. Additionally, methods that combine both visual and quantitative components have the potential to be particularly informative. In particular, the purpose of the current research is to investigate the potential usefulness of a new version of an existing method of characterizing the convergence of posterior distributions obtained in MCMC estimation, the cumulative sum procedure (CUSUM, Yu and Mykland, 1996). The development of the method will be described, and the usefulness will be subsequently investigated by comparing it to several other established methods.

Brief description of the new method

The method under consideration in this research is an extension of the cumulative sum (CUSUM) path plot procedure (Yu and Mykland, 1994). This technique was adopted by MCMC researchers from the field of Statistical Process Control (SPC; Page, 1954). As originally designed, the CUSUM procedure is an effective way of detecting small shifts in the mean of a distribution, and has been successfully used to monitor the output of production processes (Page, 1954) and psychometrics (person-fit; van Krimpen-Stoop and Meijer, 2000). Several investigators have had a hand in adapting this approach for use as a MCMC convergence diagnostic (Yu and Mykland, 1994, Brooks, 1998, Burke and Shu, 2010).

The cumulative sum procedure described by Page (1954) is a technique that is sensitive to changes in the mean of a distribution. Essentially, the value for each unit in a sample of production units is compared to the desired production mean (i.e., the value the unit is supposed to have as a result of the production process). When the selected sample units exhibit consecutive, same signed deviations from the production standard that exceeds a pre-specified threshold, it is an indication that the mean of the distribution is not stationary. On a surface level this looks to be an appealing method to characterize the entries of a posterior distribution in an MCMC sampling chain. However, in the case of production procedures we are in the desirable state of having a meaningful, clearly defined idea of what the mean and variance of that distribution should be. Adapting this approach to the modeling of psychometric model parameters requires some modifications. Most notably, there is no clear idea of what the value of the parameter

should be (i.e., the desired mean) under the null hypothesis, and there is no clear idea of what the variance should be so that thresholds can be established.

Yu and Mykland (1994) developed the CUSUM procedure as a visual method for assessing convergence. These authors applied the CUSUM to assessing convergence by plotting the observed values of an accumulating sum of the deviations (correcting them for the mean of the chain) for the elements in the Markov chain used to represent the PD. This plot is called the CUSUM path plot. These authors argued that the 'smoothness' (as opposed to 'hairiness') of the CUSUM plot and the distance it travels away from the mean are both indicative of the mixing rate. Mixing rate is the term used to describe how quickly the sampler is moving from its initial state to the underlying, stable distribution. These plots are compared to an 'ideal' path which is created by plotting a path based on an independent and identically distributed (i.i.d.) sequence. When the observed and 'ideal' plots behave similarly, it is seen as an indication of convergence.

Brooks (1998c) modified the CUSUM approach by adding a quantitative measure of 'hairiness' to make the method more objective. Brooks explains how each value in the observed CUSUM path plot can be transformed into a 0, 1 indicator statistic, d_i, in an attempt to capture the essence of 'hairiness'. Simply put, if a given element (S_i) in a CUSUM sequence is larger than its two immediate neighbors (S_{i-1}, S_{i+1}) it satisfies this 'hairiness' condition (because a plot of these points connected by line segments would require that the line segments have slopes with alternating signs); also, if a given element is smaller than its two immediate neighbors it satisfies this condition. If a given element is not larger or smaller than its two immediate neighbors, then it does not satisfy the

'hairiness' condition because the plot of these elements would appear smooth. Brooks (1998c) suggested that the values in a given CUSUM chain be transformed according to this rule, and then the accumulating average plotted over time, $\underline{D_t}$. This value is interpreted as the proportion of times an element in the Markov chain is on the opposite side of the mean as the previous element. If the observed value of this summary of the indicator statistic falls within prescribed thresholds, then the chain is behaving as if 'converged.' Brooks (1998c) bases his technique on the assumption that the distribution characterizing the CUSUM chain is symmetric.

Burke and Shu (2010) further modified the CUSUM approach by adapting Brooks' (1998) technique in three ways. First, the large degree of linear dependence (a result of the Markov property) in the observed chain needs to be removed by using autocorrelations to thin the chain before it is characterized by the indicator statistic. This removal of linear dependence is necessary because the thresholds used in characterizing the accumulating average of the indicator statistic are based upon an assumption that the elements are independently and identically distributed (i.i.d.). The autocorrelations (dependencies among the elements in the Markov chain) affect the value of the indicator statistic. So, the Markov chain must be thinned, by taking every n^{th} element, so that the remaining chain elements are not linearly dependent upon one another. Second, the chain of values to which the indicator statistic is applied is different than that proposed by Brooks (1998c). The indicator statistic is applied to the observed Markov chain, not the CUSUM chain. This results in a different expected value for the summary of the indicator statistic, $\underline{D_t}$. This argument is presented in the Methods section. Third, contrary to Brooks

(1998c), no assumption needs to be made about the shape of the posterior distribution characterizing the Markov chain. An argument for the lack of needing an assumption about shape is provided in the Methods section.

The current version of the CUSUM procedure is in need of thorough investigation. The strengths and weaknesses of the technique as it relates to convergence diagnosis must be revealed. Also, the similarities and differences to existing techniques must be demonstrated. For the method to gain recognition, it must be shown that the technique provides a beneficial alternative to existing techniques. In order to show how the technique compares to existing methods, it is necessary to create a situation in which the chains being diagnosed for convergence have known characteristics.

Simulating Markov Chains

To provide for controlled comparisons among the diagnostics considered in this study, a method for simulating chains with controlled amounts of autocorrelation among elements and controlled movement of the mean is needed. In this way, it is known ahead of time how the chain is behaving so that the effectiveness of the methods can be compared accurately.

First, it is necessary to simulate the chains so that they can range from completely independent draws to strongly dependent draws. This range of dependency is accomplished by controlling the degree of autocorrelation present in the simulated chains. Second, it is also desirable to simulate chains where the mean is stable and those where the mean is fluctuating. The stability, or lack thereof, is controlled by the random sampling component of the simulated chains that will allow for control of the stability of

the mean of the simulated chains. This second component of the simulated chains is referred to as the balance of the random component of the simulated chains. A thorough description of the method of simulating chains will be provided in the Methods section.

Purpose of the current research

Increased use of the MCMC procedure in the future goes hand in hand with greater understanding of the details of its implementation. The current research focuses on quality control checks for the output of MCMC samplers. The purpose of this paper is to investigate the usefulness of a new method of characterizing the convergence of posterior distributions obtained in MCMC procedures. The specific goal of the currently proposed research is to describe, modify, and subsequently investigate the capability of a procedure (relatively un-researched in regards to psychometric models) to assess the stability of posterior distributions of model parameters estimated with MCMC methods. This document continues the work of Burke and Shu's (2010) adaptation of the CUSUM method. In regards to the purpose of the current research, there remains a great deal of work to be done to compare and contrast the many methods for assessing convergence in the context of psychometric applications. Even with such work being done (Sinharay, 2004), many convergence diagnostics remain relatively un-researched, especially in regards to psychometric models. In order for MCMC methods to continue to gain acceptance, evidence must be provided that estimates obtained from these procedures are stable and sensible. In general, this research aims to add to the wealth of growing evidence that MCMC estimation offers a practical alternative to more familiar forms of estimation (i.e., E-M) when confronted with complex dimensionality by addressing one

of the biggest concerns with the approach: How confident are we in saying that the posterior chains obtained from this procedure have *converged* to a stationary distribution?

Research questions

Now that the modified CUSUM convergence diagnostic has been introduced and the goals for the current research have been provided, the specific research questions to be addressed are described.

The first research question that will be addressed is: What is the relationship of the degree of autocorrelation among chain elements, the balance of the random component in the chain simulator, and the value that the summary of the indicator statistic, $\underline{D_t}$, takes on in the case where the indicator statistic is applied to the observed Markov chain? To answer this question, the distribution of the indicator statistic, $\underline{D_t}$, for the case of the continuous uniform distribution as the random component of the chain simulator is derived.

The second research question that will be addressed is: What effect does thinning the Markov chain have on the 'diagnosis' of convergence/non-convergence for the CUSUM method and the method as directly applied to the Markov chains? Answering this question can be achieved by simulating chains with varying degrees of AC and balance and comparing the value of the summary of the indicator statistic for thinned and un-thinned chains for the two methods. This research question can also be addressed by applying the CUSUM convergence diagnostic and the current method to the thinned and un-thinned chains from real MCMC samplers with varying ratios of variances for the proposal and target distributions.

The third research question that will be addressed is: How does the CUSUM method compare to the Geweke (1992), Heidelberger and Welch (1983), and Raftery and Lewis (1992) in terms of rates of convergence/non-convergence of simulated chains? This question can be answered directly. Specifically, chains of varying AC and balance will be generated and then convergence will be diagnosed by each method. The methods will be compared in terms of their agreement. The conditions for this simulation study will be described in the methods section

CHAPTER II

LITERATURE REVIEW

Markov Chains and Monte Carlo procedures

Markov chains are random processes that produce sequences of random variables in which the elements of the chain have the Markov property (Sinharay, 2003). The Markov property implies that the value of each new element in the sequence is influenced only by the previous element. More formally, a Markov chain is a sequence of random variables, M_k, k= 1, 2,…, n, in which the value of each variable partially depends only on the previous variable (Patz & Junker, 1999a). Specifically, the conditional probability distribution for an element in the chain, $P (M_k = x|M_{k-1} = y)$, depends only on the preceding element. Markov chains can be used in conjunction with Monte Carlo experiments to produce numerical solutions to problems where analytical ones aren't possible.

Monte Carlo integration (i.e. numerical integration using random numbers) provides posterior expectations of functions of the parameters being approximated (Sinharay, 2003). The term 'Monte Carlo' implies that there is repeated random sampling used to generate the values in the chain. Monte Carlo simulations use computational algorithms to repeatedly sample from probability distributions for the purpose of providing approximate solutions in situations where closed form solutions are impossible or intractable (Geyer, 1992). To obtain estimates of model parameters, a Markov chain is

constructed from which a sample of observations can be generated in a random fashion by the repeated simulation of random numbers. Taken together, Markov chains and Monte Carlo methods provide a powerful tool for psychometricians.

Markov Chain Monte Carlo (MCMC)

MCMC sampling is used for estimation of multivariate distributions (Chib & Greenberg, 1995) and has become very popular in the field of Bayesian Analysis, especially when dealing with highly dimensional statistical models (Patz & Junker, 1999b). The multivariate distribution of interest in psychometric applications is the posterior distribution of the model parameters given the observed response data (Kim & Bolt, 2007). MCMC refers generally to a number of algorithms designed to sample from probability distributions in order to create a chain of random variables that will eventually be interpretable as random draws from a stable target distribution. The chain of values acts as a sample to provide an approximation of the distribution believed to describe the model parameters of interest. MCMC provides a way to repeatedly sample values from a convenient distribution that can eventually represent the joint posterior distribution of the unknown parameters of interest for a chosen psychometric model (Patz & Junker, 1999a). The sampled observations are then used to estimate the parameters of the model in use in much the same way that population parameters are estimated from sample statistics (Patz & Junker, 1999a).

In MCMC, each element of the Markov chain represents a unique state. When generating the next element in the chain, the current state is taken into consideration when making a decision about the transition to the next state. This decision is controlled

by the transition kernel. The transition kernel is a conditional distribution function, and describes the probability that the current state of the chain is equal to the sampled value given the value of the previous element (Chib & Greenberg, 1995). Thus, it describes the probability that the chain will move from its current state to the following step (Chib & Greenberg, 1995). For example, the transition kernel could take on the form, $t[(\Omega^0), (\Omega^1)] = P[M_{k+1} = (\Omega^1)|M_k = (\Omega^0)]$. Here, Ω^0 refers to the value observed for the parameter Ω at state 0, and Ω^1 refers to the value of the parameter at the following state, 1. For a more specific example, Patz & Junker (1999a) demonstrate the transition kernel in the context of an IRT framework.

Logic of MCMC for estimating model parameters

The logic of MCMC for estimating model parameters lies in defining the transition kernel in such a way that the underlying stationary distribution, $\pi(\boldsymbol{\Omega})$ (where $\boldsymbol{\Omega}$ is multivariate and describes all parameters of interest), of the chain is equal to the PD, $f(\boldsymbol{\Omega}|\boldsymbol{X})$, we are trying to estimate (i.e., the distribution of the model parameters given the observed data). Thus, given a sufficient number of elements in the chain have been produced, the Markov chain will act as a random sample from the posterior distribution in question because the elements in the chain should be distributed in the same fashion as the posterior we are trying to estimate. For example, the mean of the values in the chain is treated as an estimate of the parameter in question.

The general formula for MCMC estimation is:

$$f(\Omega|X) = \frac{f(X|\Omega) \cdot f(\Omega)}{\int_{\Omega} f(X|\Omega) \cdot f(\Omega) d(\Omega)} \qquad (1)$$

where $f(\Omega|X)$, represents the PD of the model parameters given the data, $f(X|\Omega)$ represents the likelihood of the data given the model parameters, $f(\Omega)$ represents the prior distribution for the model parameters, and $\int_{\Omega} f(X|\Omega) \cdot f(\Omega) d(\Omega)$ is a normalizing constant to ensure that the PD is a proper probability density function (pdf). The term *f* is used to represent general functions of the terms in parentheses, $(\cdot)$, and will be used interchangeably with the more specific term, *p*, that represents pdfs. Typically, the model parameters and observed data are represented as vectors as this can be implemented for multivariate distributions. The likelihood of the data is related to the particular model employed and observed response data, and the priors are selected by the practitioner. The only stipulation on the particular model in place is that it is identifiable. So, if an identified model is used to describe the likelihood and priors are selected, the posterior only relies on calculation of the normalizing constant, but as has been shown, this is not necessary to implement a MCMC sampling procedure (Patz & Junker, 1999a and 1999b). In this case MCMC estimation can still be implemented because the PD is proportional to the product of the likelihood of the data given the parameters and the priors, which can be written as:

$$f(\Omega|X) \propto f(x|\Omega) \cdot f(\Omega) \qquad (2)$$

where all the terms are similar to the previous equation, which is all the information necessary to evaluate the relative likelihoods of different sets of parameter values. As a result, an MCMC sampling procedure can proceed (Kim & Bolt, 2007).

Priors

MCMC is used in Bayesian frameworks, so it involves beliefs about the likely values those parameters are to take on, and these beliefs are implemented by way of prior distributions describing the model parameters. The specification of priors is commonly done in IRT applications with ML estimation (e.g. EAP and MAP in BILOG; Kim & Bolt, 2007). Priors allow us to incorporate information believed to be true about items and persons to aid in estimation of those parameters. The inclusion of priors is sometimes necessary, such as when the data are not very informative about the value of the parameters (e.g., the c parameter in the 3PL is a good example of this). In MCMC, the specification of priors is absolutely necessary (Kim & Bolt, 2007), however, they do not have to be specified in such a way as to be informative (i.e. indicate that any one value of the parameter is more likely than another). Informative priors are such that certain possible values have a greater probability of being observed (e.g., a normal prior is informative in that we expect to sample more values near the mean than near the tails). Non-informative priors are such that each and every possible value is equally likely. When non-informative priors are specified the resulting estimates are similar to those

obtained via maximum likelihood. These distributions are said to be non-informative in that the prior does not influence the value of the posterior towards any one value more than another, rather the data is providing most, if not all, of the information as the final estimate of the posterior.

There are several properties of priors that are of concern to practitioners of MCMC. Conjugacy, strength, and the number of levels at which to apply priors are three concerns worthy of describing briefly (Kim & Bolt, 2007). First, when priors have the property of conjugacy, the posterior density returned from the estimation procedure belongs to the same family of distributions as the prior. The implication is that the distributional form of the posterior has been correctly specified, and this is directly related to the computational efficiency of a sampler. When a conjugate prior is chosen, the sampler will be more efficient. Efficiency will be described in greater detail below. The possibility of incorporating conjugate priors is related to the particular model chosen and the observed data. Second, the strength of the prior is related to its specified variance. The term 'hyper-parameter' is used for the values specified for the parameters of the prior distribution. The PD is a combination of the likelihood of the data and the influence of the prior densities. As the variance of the prior shrinks, the influence of the prior usually increases, because it places a smaller range on the values expected to be observed. This reduction in variance of the prior in effect reduces the influence of the data on the *posterior* density of the parameters. However, with enough data, the influence of the prior wanes and eventually is minimized—for very large data sets. With a large variance, a wider range of values are expected with greater probability, allowing the data to be more

influential in the final posterior observed. Third, prior beliefs can be incorporated at multiple levels. For example, hyper-priors are prior distributions used to describe the possible values for the hyper-parameters in the priors. When there is less certainty about the values of hyper-parameters, hyper-priors can be used to reflect this uncertainty. The specification of hyper-priors acts to reduce the strength of the priors on the final estimate of the PD.

There are numerous variations of MCMC samplers. Two of the most commonly used MCMC samplers in psychometrics are Gibbs samplers and Metropolis-Hastings (MH) samplers. These are closely related and complementary techniques, and the Gibbs sampler has been shown to be a special case of the MH approach (Gelman, 1992).

Gibbs samplers

When an MCMC sampler is created that has the transition kernel defined by way of the complete conditional distributions, it is said to be a Gibbs sampler (Geman and Geman, 1984). The complete conditional distributions represent the probability of each model parameter given the data and all other model parameters. In practice, Gibbs samplers are commonly set up to estimate the posteriors for one parameter at a time, taking draws from univariate complete conditional distributions, $p\left(\Omega_{\mathrm{p}}\middle|\mathrm{X},\Omega_{-\mathrm{p}}\right)$. Here, Ω_p represents the particular parameter, p, being estimated, X is the data, and Ω_{-p} represents all other model parameters. Each model parameter is estimated as if the other parameter values are fixed, which is not conceptually different than the 'divide and conquer'

strategy employed in MLE approaches (Patz & Junker, 1999a).Thus, the transition kernel in a Gibbs sampler takes the form:

$$t_{Gibbs}\left[\left(\Omega_p^0\right),\left(\Omega_p^1\right)\right] = p\left(\Omega_p^1 \middle| \Omega_{-p}^0, X\right) p\left(\Omega_{-p}^1 \middle| \Omega_p^1, X\right) \qquad (3)$$

and has $\pi(\Omega) = p(\Omega|X)$ as its stationary distribution. A value for a given parameter is sampled from the complete conditional distribution for that parameter with all of the values for parameters upon which it is conditional fixed to their value from the previous step. Then, a value for another parameter is sampled treating all other parameters as fixed. Thus, each estimated model parameter is updated at each step. This process is continued until a sufficient number of iterations have occurred. The WinBUGS software package (Spiegelhalter, Thomas, Best, and Lunn, 2003) implements Gibbs samplers.

In a Gibbs sampler, it is required that the normalizing constants for each parameter can be calculated. As mentioned earlier, the normalizing constants represent the integration across the product of the complete conditionals and the prior distributions on those parameters with respect to the parameter in question. These are used to correct the complete conditional distributions in order to make them proper densities (i.e., probability distribution functions that have a total area of one). This same integration dilemma occurs with marginal maximum likelihood solutions—see, for example, Bock & Aiken (1981). It is sometimes possible to simplify the calculation of these normalizing constants. For example, Tanner and Wong (1987) provide a data augmentation approach to simplifying the calculation of the normalizing constants. However, if determining the

normalizing constants is not possible, then other MCMC techniques must be employed in which the constants are not necessary to carry out estimation. One way to avoid calculation of normalizing constants is to create a rejection sampler, which can be done within a Gibbs sampling framework (Ripley, 1987). Rejection samplers use proposal distributions, which are any convenient distribution from which to sample, to provide potential candidate members for the posterior distribution. The notation for the proposal distribution for values of a single parameter, *p*, is $q(\Omega_p^*)$. The candidate value for the parameter in question is referred to as Ω_p^*. A mechanism is put in place to accept candidate draws that exceed some minimum acceptance probability. This mechanism is a likelihood ratio where we define the acceptance probability α as,

$$\alpha = C \cdot \frac{p\left(\Omega_p^* \middle| X, \Omega_{-p}\right)}{q\left(\Omega_p^*\right)} \qquad (4)$$

In this acceptance ratio, *C* is a fixed constant which subsumes the normalizing constant necessary for a Gibbs sampler to function. *C* is chosen to be as large as possible as long as $0 \leq \alpha \leq 1$. The value $p\left(\Omega_p^* \middle| X, \Omega_{-p}\right)$ again refers to the univariate complete conditional distribution for Ω_p^*, and the proposal distribution, $q\left(\Omega_p^*\right)$, is in the denominator. When a draw is made, we calculate the probability of its acceptance and compare it to a random outcome with probability equal to α, for example. If the draw meets our criteria for acceptance (i.e., the flip is Heads), it is added to the posterior. If the draw does not meet the criteria, then it is discarded and another draw is made. This process is continued until

the desired number of elements of the posterior distribution has been achieved. The dimensionality of the model parameters and the similarity of the proposal distribution to the true posterior density affect the speed of rejection samplers. As dimensionality decreases and similarity of the proposal and posterior increase, the efficiency of the sampler increases. A particularly popular form of rejection sampling employs the Metropolis-Hastings algorithm (MH; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller, 1953; Hastings, 1970), and it has been shown that Gibbs samplers are a special case of the MH algorithm (Gelman, 1992).

Metropolis Hastings (MH) samplers

Rejection samplers can be implemented directly through the use of the MH algorithm (von Neumann, 1951; Patz & Junker, 1998a). Arguably the simplest approach to constructing a chain to estimate the posterior (Hanson and Cunningham, 1998), MH samplers only require the specification of priors, a choice of model to define the likelihood of the data given the model parameters, and the specification of a convenient proposal transition kernel. Similar to the rejection method just described, in MH to step from one state in the parameter space to the next we sample a candidate step, (Ω^*), from a convenient proposal transition kernel, $q\,[(\Omega^0),(\Omega^1)]$ and take the step, $(\Omega^k) = (\Omega^*)$ with acceptance probability:

$$\alpha[(\Omega^0, \Omega^*)] = \min\left\{\frac{\pi(\Omega^*)q[(\Omega^*, \Omega^0)]}{\pi(\Omega^0)q[(\Omega^0, \Omega^*)]}, 1\right\}, \qquad (5)$$

and takes the step $(\Omega^k) = (\Omega^{k-1})$ otherwise. Stated simply, this new candidate value is compared to the previously accepted value to determine its acceptance into the chain, if not accepted, the previous value is retained. The transition kernel for a MH sampler is:

$$t_{MH}[(\Omega^0),(\Omega^1)] = q[(\Omega^0),(\Omega^1)]\alpha[(\Omega^0),(\Omega^1)] + \delta_{\Omega^0}(\Omega^1)\left[1 - \int q[(\Omega^0),(\Omega^1)]\alpha[(\Omega^0),(\Omega^1)]\, d(\Omega^1)\right], \quad (6)$$

where δ_{Ω^0} is a point mass at (Ω^0) , and $1 - \int q[(\Omega^0),(\Omega^1)]\alpha[(\Omega^0),(\Omega^1)]\, d(\Omega^1)$ is the probability of not moving to the candidate step. This transition kernel for the MH algorithm has stationary distribution $\pi(\Omega) = p(\Omega|X)$.

Proposal distributions, also referred to as 'candidate generating densities' and represented generally by $q(x, y)$ (Chib & Greenberg, 1995), can be any proper density function (i.e. integrate to one over the range of y). When the Markov chain is at point x, the proposal distribution produces a candidate value y from $q(x, y)$.Typically, in MH applications the proposal density will not satisfy the condition of reversibility (the probability of going from state x to state y is equal to that of going from y to x), which is necessary if the chain is to converge to the invariant distribution (Chib & Greenberg, 1995). Thus, not every candidate from the proposal density can be accepted. To control the reversibility of the process, some candidates will have to be excluded from the chain (i.e. some of the moves from state x to state y will not be allowed).This is why the acceptance ratio α is used. This probability $\alpha(x, y) < 1$ is called the probability of move (Chib & Greenberg, 1995), and controls entry of candidates into the chain. Thus, the

probability of moving from state x to state y is the product of $q\,(x, y)$ and $\alpha\,(x, y)$.Thus, in MH algorithm applications, the acceptance ratio functions to ensure that a chain produced will have the necessary quality of reversibility.

The MH algorithm allows for sampling from a probability density function that is proportional to the posterior probability density function, and does not require that the normalizing constant be known. When the normalizing constant does not need to be known, then all that needs to be known is the likelihood function based on the model under consideration and the priors on the parameters of the model. The posterior density is proportional to the product of these two known quantities. In many psychometric applications of MCMC—especially multidimensional applications—it tends to be true that calculation of the normalizing constants is impossible or intractable, so the MH algorithm extends the reach of researchers interested in applying it to estimation problems. As with other rejection samplers, MH makes use of a more convenient distribution to create a proposal transition kernel to provide potential candidates for entry into the Markov chain. These candidates are then evaluated as to the likelihood of their membership as compared to the previously accepted member of the chain (this is done by way of a likelihood ratio that includes the density of the candidate in the numerator, and the density of the previous step in the denominator; if the density of the candidate is larger than that of the previous element, the candidate is always accepted, if not the candidate is accepted with probability equal to the value of the ratio). If the candidates are deemed ‘acceptable’ members then they are entered as part of the sample. If not, the previous entry is retained (i.e., entered again) and another candidate is generated. For an

efficient sampler, the acceptance rate should neither be too high or too low. Tuning of the acceptance rates is related to the dimensionality of the model being estimated as well as the appropriateness of the proposal distribution. In most psychometric applications, the dimensionality of models is high, which makes a pure MH approach challenging. In its favor, MH is a robust sampling technique. It allows for fairly general unimodal target posteriors, but the tradeoff is that it can be fairly inefficient (Hanson and Cunningham, 1998). It is possible to incorporate MH steps within a Gibbs sampler (Patz & Junker, 1999a).

MH within Gibbs

The Gibbs technique and the MH technique can be combined to work together in a sampler (Patz & Junker, 1999a) and still produce a stable underlying distribution, $\pi(\Omega)$, which is equal to the posterior distribution, $p(\Omega|X)$. As its name implies, we use a Gibbs strategy to sample from the complete conditionals where possible and use single iterations of the MH algorithm to deal with the cases where the complete conditionals are unknown. Patz and Junker (1999a) describe the implementation of a MH within Gibbs sampler. Using the proposal distribution for the parameter in question, $q_{\Omega}(\Omega^{0}, \Omega^{1})$, try to draw Ω_p^{k} from the complete conditional distribution, $p(\Omega_p|\ \Omega_{-p}^{k-1}, X)$ by drawing Ω_p* from $q_{\Omega}(\Omega_p^{k-1}, \Omega_p)$ and accepting with probability equal to:

$$\alpha\left(\Omega_p^{k-1}, \Omega_p^{*}\right) = min\left\{\frac{p\left(X|\Omega_p^{*}, \Omega_{-p}^{k-1}\right)p\left(\Omega_p^{*}, \Omega_{-p}^{k-1}\right)q_{\Omega}\left(\Omega_p^{*}, \Omega_p^{k-1}\right)}{p\left(X|\Omega_p^{k-1}, \Omega_{-p}^{k-1}\right)p\left(\Omega_p^{k-1}, \Omega_{-p}^{k-1}\right)q_{\Omega}\left(\Omega_p^{k-1}, \Omega_p^{*}\right)}, 1\right\}, \qquad (7)$$

otherwise set Ω_p^k equal to Ω_p^{k-1}. When the proposal distribution is symmetric, it cancels out of the acceptance probability, simplifying the calculation.

Considerations when constructing a sampler

When constructing a sampler, there are many things that should be taken into consideration. The selection of proposal distribution, the blocking of parameters, the acceptance rate, burn-in, mixing rate, and covariance structure of the model parameters are all important factors in the decisions to be made. Each of these will receive further treatment before proceeding with discussions of efficiency and convergence.

One of the most important choices to be made when constructing an MCMC sampler is the choice of the proposal distribution used to produce candidate elements for the Markov Chain. The characteristics of the proposal distributions that are important to consider are the shape that the distribution has, as well as the values of the parameters for the distribution. The choice of the specific form of the proposal distribution for use in an MCMC sampler has a great deal of influence on the behavior of the chains produced (Hanson and Cunningham, 1996, Chib and Greenberg, 1995). Proposal distributions have a great deal of influence on the efficiency of the sampler, the acceptance rates of candidate elements and the degree of AC present in the chain. For example, the location and scale parameters of the proposal distribution control the tuning of the sampler. Choosing the right location and choosing the variability we see in generated candidates can both influence how often candidates are accepted into the chain. Acceptance rate then has a great deal of influence on the behavior of the chain, which will be discussed below. In addition, the particular family that the proposal distribution belongs to directly

influences the behavior of the chain. For example, Chib and Greenberg (1995) describe five families of candidate generating densities. Two of these five general types of MH samplers are *random walk samplers* and *independence samplers*. A random walk MH is created by specifying that the proposal distribution is symmetric and centered at the value of the previous accepted candidate with a variance chosen to influence the acceptance probability in a desired fashion (Hanson and Cunningham, 1998). A common choice is a normal distribution with mean equal to the value of the previous state; $X \sim N(M_{k-1}, c\sigma^2)$, where $c\sigma^2$ is the variance chosen specifically to provide a desirable acceptance rate (Sinharay, 2003). Alternatively, an independence MH chooses a proposal distribution that is not necessarily symmetric (which typically means that there is not as great a degree of simplification of the acceptance probabilities) and is centered not at the previous value accepted into the chain, but rather at some estimate of the of the parameter being estimated (using the raw score to create an estimate of ability, for example; Chib and Greenberg, 1995). In this sense, the candidates generated are independent of the previous step. A very convenient common proposal density is a continuous uniform distribution centered at the current state of the chain with a finite width (restricted support, Chib and Greenberg, 1995). The fact that MH samplers are so robust makes this a typical choice which is usually successful.

Another consideration when constructing a sampler is whether or not to block parameters. Blocking entails grouping parameters that will be updated together at each step of the sampling. In the simplest case, each parameter is seen as independent from every other parameter, thus they are each updated independently via their own proposal

distribution which only reflects that parameter. Treating each parameter as independent from all other parameters simplifies the form of the acceptance probability. Blocking parameters simply means simulating multiple parameters simultaneously from a multivariate proposal distribution with the inclusion of a specified covariance structure describing the relationships among the parameters (as compared to treating each parameter in a univariate sense). The decision to block parameters is a trade-off between efficiency and accuracy. When large blocks of parameters can be incorporated via multivariate candidate distributions or complete conditional distributions it makes sampling more efficient. Small blocks of parameters mean that there are more individual sampling steps taken (which can reduce efficiency) but it allows for easier tuning of the sampler via analysis of acceptance rates (Patz & Junker, 1999a). In an IRT setting, Patz and Junker (1999a&b) describe the procedure for blocking parameters together to improve the efficiency of a sampler. For example, in IRT applications, it is convenient to block model parameters by individual persons and individual items. In the case of a 3PL, each person parameter, θ, can be treated as one dimensional and each vector of item parameters, β, can be treated as 3 dimensional. Each item and person will then have a respective sampling distribution. In the example provided, the 3PL is used for the sake of simplicity. MMLE works very well for this sort of estimation problem. When the dimensionality becomes more complex, MCMC is in an advantaged position.

Another concern when constructing a sampler is the acceptance rate for proposed candidates. The number of proposed candidates that get accepted into the chain influences the behavior of the sampler. The acceptance rate is used as an index to tune the

sampler to obtain optimal efficiency (tuning is done specifically by manipulating the location and spread of the proposal density). When the acceptance rate is too high or too low it can have a negative impact on the efficiency of the sampler to produce a chain suitable for use as an estimate of the posterior distribution. For example, the acceptance rate is commonly 'tuned' by adjusting the variance of the proposal distribution. When the variance is large, proposed candidates can vary greatly from the previous value. A large variance results in a low acceptance rate because large differences can exist between a new proposed value and the previously accepted value which influences the value of the acceptance ratio. Specifying a variance that is too large can result in 'sticky' samplers where the same value is retained on multiple successive steps and it induces a large degree of autocorrelation. When new values are accepted, it is possible that large jumps can occur, resulting in incomplete exploration of the parameter space. On the other hand, when the variance of the proposal distribution is too small, the new candidates are very similar to the previous step leading to a high acceptance rate. This type of sampler is slow to explore the parameter space because it takes small steps though the distribution. Specifying a variance that is too small also leads to a high degree of dependence among elements in the chain, which affects the mixing rate, and ultimately convergence to the stable underlying distribution. When the variance of a proposal distribution is called large or small, it is always relative to the variance of the target distribution. An ideal sampler will have a proposal distribution whose variance closely matches that of the target distribution. Acceptance rates between 25% (multivariate cases) and 50% (univariate cases) often produce efficient samplers, all other things being held constant (Patz &

Junker, 1999a, Hanson and Cunningham, 1998).

Another important consideration when constructing a sampler is determining the length of the burn in. Burn-in refers to the beginning portion of a chain that is discarded. The initial sampled values are not considered to provide good estimates of the parameters, due to strong autocorrelations. That is, the initial estimates are strongly related to earlier values in the MCMC chains, and possibly influenced by choices of starting values. Therefore, the initial n_0 burn-in values in the chain are discarded before any attempt to characterize the posterior distribution is made (Hanson and Cunningham, 1998). The post-burn-in draws are only regarded as a sample from the invariant posterior distribution after the chain has moved sufficiently far away from its arbitrary initial state (Chib & Greenberg, 1995). It is known that treating the length of the burn in as an increasing function of the first order serial correlation is a useful heuristic in many situations (Chib & Greenberg, 1995). A simple strategy is to use the AC as a guide to decide how many elements to remove from the beginning of the chain. First, determine the lag necessary so that the AC for the observed chain in question goes to zero. Second, remove at least that many elements from the beginning of the chain (Raftery and Lewis, 1992b).

Another consideration when constructing a MCMC sampler is the mixing rate of the chains. The mixing rate describes the speed with which the Markov chain is moving towards the equilibrium distribution. Fast mixing chains require shorter run lengths before achieving stability. Slow moving chains require longer run lengths before achieving stability. Mixing rate is directly related to convergence. In fact, the CUSUM

path plot convergence diagnostic incorporates the notion of mixing rates in its treatment of convergence. Mixing rate can be thought of as the number of steps necessary before the chain reaches the underlying stationary distribution and can be treated synonymously with efficiency.

Efficiency

"The statistical efficiency of an MCMC sequence is defined as the reciprocal of the ratio of the number of MCMC trials needed to achieve the same variance in an estimated quantity as are required for independent draws from the target probability distribution (Hanson and Cunningham, 1998, p. 373)." There are many things that affect the efficiency of a sampler.

Efficiency is related to the AC and can be estimated from it. Efficiency is defined in terms of the variance of an estimated quantity. When a strong degree of AC is present among a sequence of variables it reduces the apparent variability of those observations. When the AC is strong, more elements would have to be discarded to leave only independent elements. When more elements have to be discarded to leave only independent elements, it is indicative that more elements would have to be generated. Hanson and Cunningham (1998) show how the statistical efficiency of a sampler can be calculated from the AC present in a chain. Autocorrelation tells us about the degree of dependence among a string of consecutive numbers. In MCMC, the AC is a naturally occurring byproduct of the way in which the method works. Essentially, typical MCMC estimation for psychometric purposes builds AC into the resulting chain. This AC is informative about the behavior of the chain, but can also be a hindrance to estimation of

parameters and inference based upon them. AC is closely tied to mixing rate, decisions about burn in, sampler tuning, parameter estimation, and assessment of convergence. As AC increases, convergence slows (Kim & Bolt, 2007). A high degree of AC can be caused by poor parameterization and/or over parameterization. Over parameterization is capable of producing 'ridges' (i.e., local maxima) in the likelihood surface. AC adds difficulty to the estimation of variances associated with the parameter in question. A strong degree of positive linear relationship means that variance is underestimated. Corrections for this exist, however (CODA; Best, Cowles, and Vines, 1996). When the acceptance rate is too high or too low, large amounts of AC will exist in the chain (Chib & Greenberg, 1995)

Also, the relationship between the degree of correlation among parameters and whether or not that dependence is taken into account in the sampling mechanism can affect efficiency. For example, if several model parameters are highly correlated, but the sampling mechanism treats them as independent, this will result in an inefficient sampler. As stated earlier, blocking parameters can improve efficiency when the relationship among those parameters can be accurately captured in the proposal mechanism. When there are dependencies among parameters, this can be handled by proposing values for each group of associated parameters based on a proposal distribution that incorporates the covariance matrix representing the dependencies that exist. For example, Hanson and Cunningham (1998) developed a method to estimate the covariance matrix of the posterior distribution in order to aid in sampling efficiency with success. Additionally, Patz and Junker (1999a) provide an example of how re-parameterizing the model so that

the covariance matrix has zeroes on the off diagonals and is isotropic (i.e., all variables have similar variances). Thus, re-parameterization of a model in such a way that would allow construction of a sampler that specifies all parameters as independent will be more efficient than a sampler based on a model whose parameterization allows for dependence.

Also, the degree of similarity between the proposal distribution and the target posterior has a great influence on the efficiency of a sampler. For example, to develop an efficient sampler, the shape of the proposal distribution should match that of the target distribution. Correct specification can improve accuracy up to the point of a sampler producing independent draws from the target pdf (Chib & Greenberg, 1995). Misspecification can result in extreme inefficiency (e.g., specifying a normal proposal when target is exponential means you will do a poor job of estimating the tails of the target distribution).

In addition to correctly specifying the shape of the target distribution, the variability must also be correctly specified. The variability of the proposal distribution has a great deal of influence on the characteristics of a sampler (e.g., AC, acceptance rates, etc). For the utmost efficiency, the variance of a proposal distribution should be similar to that of the underlying target distribution. The influence of this similarity upon acceptance rates and AC is covered thoroughly in the Methods section. However, it is worthwhile to briefly address how the ratio of proposal distribution variance to target distribution variance can affect the behavior and appearance of a chain. Hanson and Cunningham (1998) show that when proposal variance is smaller (e.g., ¼ of the target distribution variance), the resulting chain takes on the characteristics of Brownian motion

(a purely random walk). When the variance of the proposal distribution matches that of the target distribution, the resulting chain looks more like an independent sampler. When variance of the proposal distribution is much larger than that of the target distribution, the resulting chain will have many elements that are equal to one another for successive iterations then there will be a large 'jump' in the value of the following element. This type of chain is often referred to as a 'sticky' sampler (Hanson and Cunningham, 1998).

The efficiency of a sampler is indicative of how long it might take for a sampler to converge to the stationary distribution. However, trying to assess convergence of the sampler is an entirely different matter than influencing its efficiency.

Convergence

Ideally, a chain used for estimation should be indistinguishable from a sequence of random draws from a distribution with known form. Cowles and Carlin (1996) point out that there are different connotations of convergence. In a very simple sense, once a single element is chosen from the target distribution, technically all following elements will be from the target. Thus, it could be argued that convergence occurs at a given step in the chain, and all subsequent draws are by definition converged. In a more thorough sense, convergence is taken to mean that the sampler has successfully explored the complete parameter space of the posterior distribution and has roughly revealed its shape and configuration (which is much more likely to happen with unimodal target densities). This definition is a preferable notion of convergence in that we have more information concerning estimates of the parameter of interest.

Convergence implies that the distribution is stable, which means that the distribution can be described well by its parameters. For example, a stable distribution should have location and dispersion parameters that can be adequately described by a single value each. So, if a sample is being drawn from a stable distribution it would mean that the mean and variance do not fluctuate any more than is to be expected due to sampling error.

In MCMC samplers, convergence relies upon several conditions. First, the model describing the likelihood must be an identified model. By identification we mean that there exists a unique set of parameter values relating to some set of observed data. In other words, there are not multiple sets of parameters that could describe the data equally well (in which case there would be indeterminacy). Different values of model parameters should lead to unique probability distributions for the observed variables. Sinharay (2004, and references therein) points out that many psychometric models have identifiability problems that make parameter estimation troublesome. For example, the well-known 3 parameter logistic IRT model (3PL; Hambleton and Swaminathan, 1985) is known to have identification problems due to the association between 'discrimination' and 'pseudo-guessing' parameters, and has been claimed to be slightly over-parameterized (Holland, 1990). It is not uncommon in practice to run into problems with parameter calibration for this (3PL) model.

Second, an appropriate sampling mechanism must be put in place. Two of these methods (and the combination of them) have already been discussed in some detail. There are two general criteria for creating an appropriate sampling mechanism. The first

criterion is that it must be the case that the sampler constructed is known to have a stable underlying distribution that is equal to the PD we are trying to estimate (which is done via the transition kernel). The second criterion for implementing a successful MCMC sampling mechanism is that the method must satisfy the requirements for creating an ergodic string of states. This additional criterion will be described below.

Third, an appropriate chain length must be observed. The length must be long enough to overcome the effect of some arbitrary starting value. Also, the chain must be long enough to provide a stable estimate of the parameter and its variability (i.e., enough elements must be observed to 'rough out' the parameter space). Simply put, the more elements in the chain, the better the quality of the estimate we expect to see, but at a point the estimates will not be of any greater quality by including more sample entries.

The primary challenge in assessing convergence in MCMC is that convergence is from one distribution to another distribution. Adding to the complexity of assessing convergence is that we only produce a sample on which to base our assessment (we only see a piece, or one possible realization, of the distribution). Thus, sampling error is mixed in with our estimates of the parameters. This sampling error is referred to as Monte Carlo Standard Error (MCSE; Geyer, 1992).MCSE is the error introduced due to the fact that we are sampling from a distribution. This sampling error needs to be taken into account. It is easy to deal with MCSE because running the chain to more steps always reduces the sampling error. A rule of thumb is that MCSE should be less than 5% of the standard error of estimate (i.e., the standard deviation of the observed values) (Spiegelhalter, Thomas, Best, and Lunn, 2003).

Convergence is only guaranteed if a sampler produces a chain that is ergodic. Generally speaking, ergodicity is defined to mean that a system observed over a long enough duration will produce new states that are similar to previous states. In probabilistic systems, ergodicity means that in the limit new states will be independent of the initial states. This characteristic is important because an ergodic chain has only one stationary distribution. If in application we define an ergodic chain, we have confidence that the resulting estimates produced from a chain that has run long enough will be suitable for inference. Once a chain has reached this state, any further sampling will produce a chain that is invariant from the current chain.

A Markov Chain is ergodic if each and every element is aperiodic, irreducible, and positive recurrent. First, aperiodicity means that a state in a Markov chain is reproduced at irregular intervals. That is, there will be no regularity with which a given element in the chain will be equal to a previous state (i.e., a particular state does not occur systematically, rather it occurs randomly). Second, a Markov chain is irreducible if any state in the chain can be reached from any other state. In other words, you could go from observing any one element to any one other element. With a converged chain, every element should be a plausible member of some distribution. So, a sampling mechanism drawing from a distribution could produce any element at each step regardless of the previous step. Third, a chain is said to be positive recurrent if a state in a Markov chain has a non-zero probability of occurring again in a future state. In other words, an element in a converged chain has a chance of being observed again if the sampler were continued.

Assessing convergence

Now that convergence has been defined and the conditions necessary for it to be observed have been described, it is important to discuss how convergence is assessed. There are two general categories for describing approaches to assessing convergence (Cowles & Carlin, 1996). These general categories are theoretical treatments of convergence and diagnostic approaches applied to the output of MCMC samplers.

In theoretical treatments of convergence, analysis of the transition kernel is necessary to predetermine the number of iterations necessary for a sampler to achieve convergence (within a pre-specified tolerance) to the stationary distribution. While promising, these approaches involve complicated mathematics and 'laborious' calculations which must be revised in light of each model considered. These methods also tend to produce bounds that are quite 'loose' and would require far more iterations than are typically practical (Cowles & Carlin, 1996).

Diagnostic approaches are far more common. Generally speaking, diagnostics examine the output of samplers in an attempt to determine whether or not the chains are behaving as might be expected if convergence had been achieved. No claim is made that the diagnostics clearly indicate whether or not a chain has converged, rather the diagnostics provide evidence in support of claims that the chains may or may have not converged. In the case where there is not a method to determine that the chain is indeed converged, it must be determined whether or not a chain has the qualities expected if it were indeed converged. It may not be possible to get at the truth of whether or not a chain has converged, but it should at least be addressed whether or not a chain appears to have

converged.

Cowles and Carlin (1996) point out that many researchers deem all diagnostic attempts to assess convergence as ‘fundamentally flawed’. This is because it is not possible to know what the underlying stationary distribution is, therefore, the chain is being compared to something other than the true distribution to which it is converging (if it has indeed been constructed and implemented correctly). Additionally, the diagnostics used typically analyze the output of the sampler (or compare multiple outputs of the sampler) in an attempt to determine if the sampler has moved to the true posterior density. In other words, it is not possible to compare the sample to the true posterior (because if the true posterior was known there would be no need for MCMC in the first place) so we assess convergence by looking at the product of the sampler only. Despite this criticism of being fundamentally unsound, the authors argue that a ‘weak diagnostic’ is better than no diagnostic at all. If the diagnostic can at least be used to rule out chains that may in a brief examination appear to be converged, then it can help guard against improper estimates and further inference.

There are numerous ways in which diagnostics can differ in their approach to assessing convergence. For example, Cowles and Carlin (1996) distinguish among 13 diagnostics according to seven dimensions. The dimensions used to distinguish among diagnostics are: whether the method is visual or quantitative, whether they are applied to single or multiple chains, the theoretical foundation on which the method is based, whether the diagnostics focus on univariate or multivariate distributions for the parameters, whether the diagnostics characterize convergence in terms of bias or

variance, the types of sampler to which the method can be applied, and the ease of implementation.

There are two general forms of the post hoc diagnostics. Cowles and Carlin (1996) distinguish between visual and quantitative methods for assessing convergence. These methods approach the assessment of convergence from different standpoints, and are known to perform differently from one another.

Visual diagnostics

Visual methods involve graphical representations of Markov chains or some transformation of them. Common visual methods include, but are not limited to, time series plots, running mean plots, AC plots, and CUSUM path plots.

Time series plots (Sinharay, 2003) are probably the most common way to check for convergence. These are simply plots of the value of each element in the chain (on the y-axis) and the number of the step (on the x-axis). Each point is connected by a line segment so that the 'path' the chain has traversed is evident. While not foolproof, it can indicate situations where the chain has clearly not converged (e.g., continually increasing trend, wandering up and down over different parts of the chain) or provide an estimate of the number of burn in iterations to remove. If multiple chains are run, it is an easy way to investigate if they are in agreement. Simply plot them all on one graph to inspect their similarity (if they all begin to overlap at a certain point and remain similar, it is evidence in favor of convergence). Plotting the log of the posterior density over the course of the chain can be informative as well (Sinharay, 2003). If there is an increasing trend it can be taken as evidence that the chain is moving towards the mode of the parameter space. If

there is a decreasing trend, then the sampler may have explored a part of the space with little area and is moving towards a potentially more dense area.

Mean plots, as the name implies, represent the mean of the chain at various points in the sequence. A plot of the running mean provides visual evidence about the stability of the location of a chain. At every n^{th} step of the chain, the mean is calculated and plotted. If a chain has converged, there should be little change in the means at each n^{th} step. Mean plots are very simple indicators and can easily identify cases where convergence has clearly not been achieved. However, it ignores important aspects of convergence like variance, for example.

AC plots provide indirect evidence about convergence. Inspection of the plots of ACs for each parameter's chain is informative about the behavior of the sampler. The greater the degree of AC, the longer it will take a sampler to fully explore the parameter space. Slow moving chains or multiple chains that stay in different areas of the sample space can be due to high AC or multiple modes, so it is common practice to view time series plots in light of observed AC.

CUSUM path plots were originally created to be a simple way of assessing chains diagnostically. Cumulative sum plots represent accumulating deviations from the mean. When there are a large consecutive number of same signed deviations about the mean, the resulting CUSUM plots will be smooth and will 'wander' away from the mean of the overall chain. The smoothness of the CUSUM path plots and the excursions from the mean are indicative of mixing rate, which is indirectly informative about convergence. When plotted against an 'ideal path', these plots can provide information about the

behavior of a chain over time. The current method under development is a modification of the CUSUM path plots and will be discussed in greater detail shortly.

Quantitative diagnostics

Generally speaking, quantitative diagnostics differ from visual diagnostics in that convergence is represented by way of a statistical test or confidence interval. They involve representing convergence in numerical form. A chain (or chains) is transformed into the numerical representation for the sake of comparison to a null hypothesis. The null hypothesis is meant to represent the case of convergence. Thus, these approaches attempt to treat convergence as a form of hypothesis testing. When the null hypothesis is rejected, then we favor the alternative hypothesis that the chain(s) has not converged.

Given that the purpose of the current research is to compare a modified version of the CUSUM diagnostic, the following description of several quantitative diagnostics will focus on those that have characteristics similar enough to the CUSUM so as to render them amenable to direct comparison. In particular, the Raftery and Lewis (1992) diagnostic, the Heidelberger and Welch (1983) diagnostic, and the Geweke (1992) diagnostic will be described.

The Raftery and Lewis (RL; 1992) method is intended to diagnose convergence as well as provide bounds on the variance of estimates of quantiles of functions of parameters. This approach uses as input the output of any MCMC sampler that is at least 'Nmin' iterations long (where 'Nmin' is the minimum number of iterations to achieve the desired level of accuracy of estimation if the samples were independent). After providing *q*, the quantile of interest to be estimated (perhaps .025), and *r*, the accuracy desired (say

$\pm$.005), the required probability, *s*, of obtaining the accuracy desired, and a convergence tolerance, δ, (which is usually .001) the pre-written code provides the values for: 1) 'nprec', which is the total number of iterations that should be run, 2) 'nburn', which is the number of iterations to throw away as burn in, and 3) 'k', indicating the number of intervening iterations to discard when making inference based on the chain ('k' is a thinning estimate). The largest obtained value of 'nprec' should be used for all chains.

This diagnostic is based on two-state Markov chain theory as well as sample size estimation based on binomial variance. A binary sequence is created, Z, as a 0/1 indicator equal to the length of the chain, determined by whether or not the value in the original chain is less than a particular cutoff. The approach returns an index, 'I'. If the index is greater than 5 it is an indication that there are problems with convergence. Raftery and Lewis (1992) emphasize that the strength of this method lies in being able to specify the desired accuracy of estimation at each quantile of the distribution desired. Thus, the specification of accuracy at selected quantiles of the PD allows for the estimation of the shape of the target distribution very well. Thus, it allows for good estimation of center and spread, two critical components of good estimation.

Critics have emphasized that different input chain values for the exact same parameter can result in largely variable estimates of 'nprec'. Also, RL is a univariate procedure, which may be overlooking the complexities present when trying to characterize multivariate quantities. Additionally, this technique provides an estimate, 'k', of the thinning that should be done. MacEachern and Berliner (1994) point out that any estimation procedure is degraded by throwing away iterations. This particular

criticism is not unique to the RL diagnostic, and will be revisited in the Methods section.

Geweke (1992) used methods from spectral analysis to approach the assessment of convergence for Gibbs samplers. When the purpose of the analysis is to estimate the mean of a function of the model parameters being estimated after each step of the sampler, $g(\theta^{(k)})$, the Markov chain can be treated as a time series. The method assumes that the MCMC procedure and the "….function g imply the existence of a spectral density $S_g(\omega)$ for this time series that has no discontinuities at frequency zero (Cowles & Carlin, 1996, p.886)." When the assumption holds, the expected value of $g(\theta)$ can be estimated by:

$$\bar{g}_n = \frac{\sum_{i=1}^{n} g\left(\theta^{(i)}\right)}{n} \qquad (7)$$

and the asymptotic variance is $S_g(0)/n$. The numerical standard error (NSE) is the square root of this variance, and can be interpreted as an estimate of the standard error of the mean (Cowles & Carlin, 1996).

Essentially the Geweke approach (G) tests whether or not the mean at the beginning of the chain is equal to the mean at the end of the chain. Two subsections of the chain are taken, reasonably separated by some distance to assure their independence, and transformed into a value conceptually similar to a z score. The mean at the beginning of the chain is subtracted from the mean at the end of the chain, and this difference is divided by the asymptotic standard error of the difference. The diagnostic is calculated by

taking the difference between means of the first 10% and the last 50% of the elements in the chain and dividing by the pooled estimate of dispersion. When a chain produces values between -1.96 and 1.96, the interpretation is that the means from the beginning and end of the chain are not different from one another, thus it is seen as evidence that the chain has converged (because the mean is stable).

This method assesses both bias and variance, is readily available in a free software package (CODA; Best, Cowles, and Vines, 1995), is univariate (but can be extended to a multivariate treatment with ease), and requires only a single chain for its implementation. The primary disadvantage is that the value of the statistic is sensitive to the specification of the spectral window (Cowles & Carlin, 1996).

The Heidelberger and Welch (HW; 1983) diagnostic tests whether or not the last part of a Markov chain has achieved stationarity, and it assesses whether or not a pre-specified level of accuracy has been achieved. It is based on Brownian bridge theory and spectral analysis, is a univariate approach, only requires a single chain, assesses both bias and variance, and is applicable to any type of MCMC sampler (Cowles & Carlin, 1996). It is a comprehensive procedure that combines the procedures for detecting nonstationarity presented in Schruben (1982) and Schruben, Singh, and Tierney (1983). These procedures use a spectral analysis approach to estimate the variance of the sample mean, and rely on the Cramer-von Mises statistic (von Mises, 1931) to test the null hypothesis that the chain is stationary. Essentially, a confidence interval is created that has a pre-specified half-width, ε. This diagnostic is freely available in the BOA software package in R (Smith, 2001).

The HW diagnostic applies the stationarity test of Schruben (1982) and Schruben et al (1983) in an iterative fashion. If the null hypothesis is rejected for the whole chain, the first 10% of chain elements are removed, and the test is repeated. If the null hypothesis is rejected again, another 10% of the elements from the beginning of the chain are removed and the procedure repeated. This process continues until the null hypothesis is not rejected or half of the iterations have been eliminated. If half of the elements have been eliminated, the chain will need to be run longer, and the process started again. When a portion of a chain is deemed stationary, a half-width test is performed. With the spectral density estimate of the standard error of the mean, an estimated half-width is created. If this estimate is less than ε times the sample mean of the retained portion of the chain, then the process stops. The sample mean and confidence interval are reported.

CUSUM path plots

The focus of the current research is on a modification of the CUSUM technique for assessing convergence. Therefore, the method deserves a thorough description before the modifications are discussed. The original method of using CUSUM path plots and the later addition of incorporating a quantitative component will be described here, and the new modifications to the technique currently under investigation will be provided in the Methods section.

Yu and Mykland's (1996) technique developed to assess convergence comes from a related method in the field of Statistical Process Control (SPC). SPC is often used to control production processes to achieve a desired output within some desired margin of error (e.g., to ensure that each tea bag produced by a company has the proper amount of

tea in it). The CUSUM procedure (Page, 1954) is sensitive to consecutive strings of positive or negative deviations about the center of a distribution. The CUSUM procedure is an effective method for detecting small shifts in the mean of a distribution, and has been successfully used in the field of SPC (Page, 1954) as well as psychometrics (person-fit; van Krimpen-Stoop and Meijer, 2000).

The procedure, as originally adapted as a convergence diagnostic, is potentially capable of characterizing convergence in terms of detecting trends of consecutive positive (and/or negative) deviations about the mean in a posterior chain. The underlying nature of the CUSUM procedure provides a unique alternative characterization of convergence. This technique focuses directly on the posterior chains, and characterizes each element in the chain as a deviation about the mean of those values. These deviations are then considered from a global perspective in terms of their behavior across the chain. This characterization of the elements in a production sequence provides an intuitive way to address the convergence of Markov chains. This technique was developed over a decade ago by Yu and Mykland (1998). These authors proposed a slight modification of the CUSUM as described for SPC by Page (1954) so that it becomes more appropriate in the context of evaluating convergence.

Visual inspection of the posterior chains is a common means of determining convergence (as it is informative about behavior of the chain over time). However, the traditional sequential plots provide less information about the mixing behavior of the posterior chain than might be desirable. Mixing behavior describes the shift from the chain's initial state towards the stationary distribution presumed to be the end result of

the sampling procedure. In this spirit, Yu and Mykland (1996) proposed graphing a meaningful transformation of the observed chain (e.g., accumulating deviations about the mean) over time to investigate the chain's behavior. Specifically, the values in the chain are transformed into a chain of accumulating deviations about the mean. Each element in the observed Markov chain has the mean subtracted from it, and then these deviations are summed sequentially across the chain (the first element is added to the second, the new second element is added to the third and so on). These accumulating deviations are plotted and the points are connected by line segments. The smoothness of the CUSUM plot and the distance it travels away from the mean are both indicative of the mixing rate. If the resulting plot is smooth (with many consecutive line segments having positive slopes, for example) and makes large excursions away from the mean, then the chain is mixing slowly and a large number of steps will be required before the chain reaches it stationary distribution.

Specifically, given a sequence of observations obtained from an MCMC procedure, $X_1, \ldots, X_n$, begin by discarding the first n_0 observations as burn in. After removing the burn in, the average of the observations in each chain is calculated. Then an accumulating deviation from the mean, $\hat{S}_t$, is calculated and plotted over the length of the chain to visualize the CUSUM path. A benchmark path is also plotted for comparison (this is described later). These CUSUM paths are similar to traditional sequential plots, but have the nice feature of beginning and ending at zero, and emphasize consecutive same signed deviations about the mean. Consecutive same signed deviations from the mean are an indication that the mean of the distribution is changing.

The method as originally proposed was to first calculate the mean of the posterior beyond the burn in chain, $\hat{\mu}$,:

$$\hat{\mu} = \frac{1}{n - n_0} \sum_{i=n_0+1}^{n} \theta(x_i) \qquad (8)$$

where *n* is the full length of the chain, n_0 is the portion of the chain that is discarded as burn in, and $\theta(x_i)$ is the one-dimensional summary statistic that is being monitored (in most applications, this is representing the observed value of element *i* in the chain).The value $\hat{\mu}$ is simply the mean value of the observed chain beyond the burn-in. The value $\hat{\mu}$ is then used to characterize each element in the chain as a deviation to be summed sequentially over the length of the chain to obtain the CUSUM, $\widehat{S}_t$:

$$\hat{S}_t = \sum_{i=n_0+1}^{t} [\theta(x_i - \hat{\mu})], for\ t = n_0 + 1, \dots, n \qquad (9)$$

where all notation is the same as in the previous equation. Simply put, each observation in the chain has the value of the mean subtracted from it, and these deviations about the mean are summed at each step over the length of the chain. These $\hat{S}_t$ values are then plotted for the length of the chain (excluding the burn in) and connected by line segments.

Yu and Mykland (1998) explained that smooth CUSUM plots that took large excursions away from the mean were indicative of slow mixing behavior for the chain,

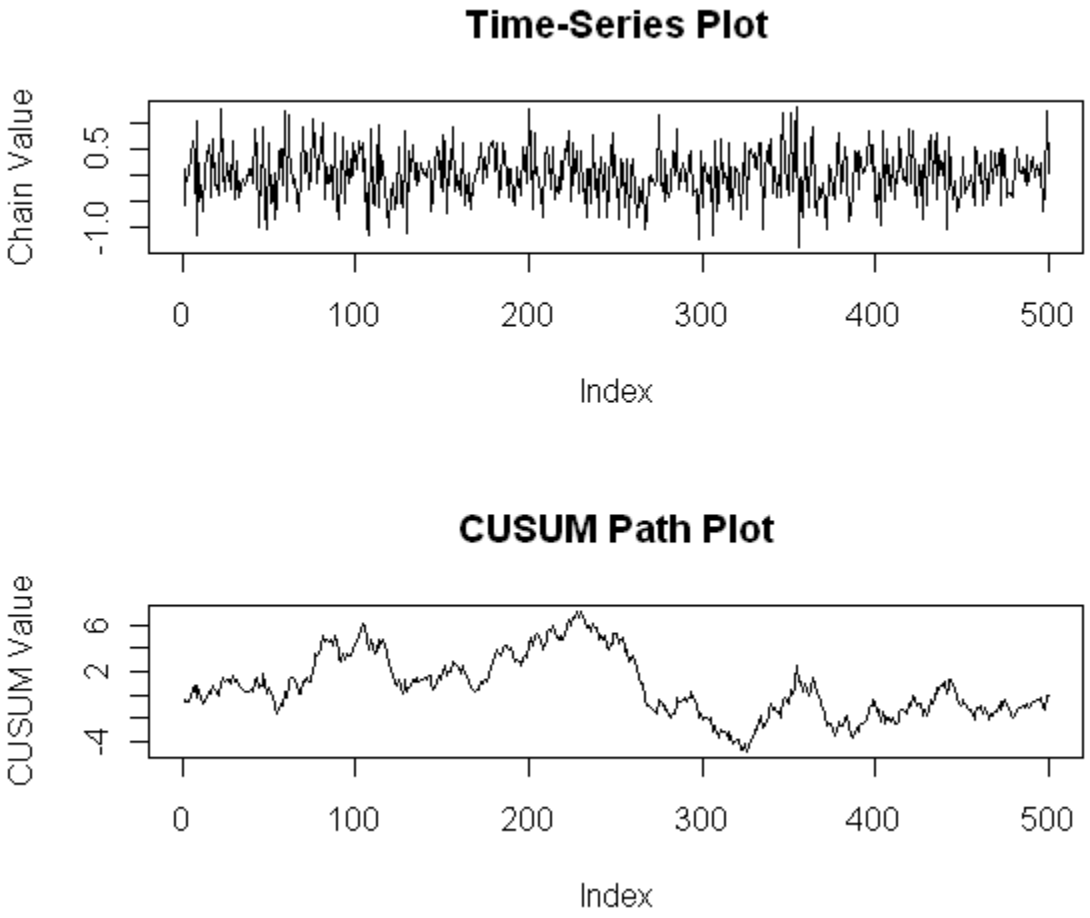
while jagged (or ‘hairy’) CUSUM plots that stayed close to the mean were indicative of fast mixing behavior. These authors also suggested that observed CUSUM plots be compared to ideal, or ’benchmark’, plots. These ideal plots are based on a simulated i.i.d. sequence from a normal distribution with mean and variance equal to that of the observed Markov chain. The i.i.d. sequence is then transformed into a CUSUM path plot. The ideal paths allow for the assessment of the behavior of the observed CUSUM paths through direct comparison in the same graphic device by acting as a rough version of a null hypothesis. These ideal paths are second order approximations to an ideal CUSUM path from the target distribution (Yu and Mykland, 1998). If the observed and ideal paths are similar, it is taken as evidence that the observed plot is behaving in a similar fashion as a converged sequence.

Of critical importance in this paper, Yu and Mykland (1998), expanding on Lin’s (1992) work with the behavior of partial sums in mixing sequences, argued that when there was rapid mixing of the distributions (i.e., when the chain is moving quickly to the stationary posterior we want to estimate) the CUSUM plot would be very ‘hairy.’ In other words, the plot would essentially be connecting points on the plot by line segments with alternating positive and negative slopes, hence the ‘hairy’ description of the resulting plot. In terms of assessing convergence, a fast mixing sequence is an indication that a shorter chain is necessary to reach a stationary distribution. Of course, this is only true when that ‘hairy’ plot stays very close to the overall mean of the chain. Taken together, a ‘hairy’ plot with a small excursion is a sign that the sampling procedure is moving quickly from its initial starting values and settling in to the presumed underlying

distribution. The CUSUM plot also provides evidence about these excursions.

For an example of what these authors mean by ‘hairy’, see the plots immediately below.

Figure 1: Example of a ‘hairy’ plot



The first of these two plots is a time-series plot for a sequence of values generated from MH within Gibbs sampler. The second plot is the CUSUM path plot for the same sequence. There are two things in the second plot that indicate fast mixing. First, the ‘hairiness’ of the plot above is evidenced by the fact that for the most part the CUSUM

observations are connected by line segments with alternating positive and negative slopes. Second, in terms of excursion, these values tend to alternate evenly on either side of the center of the observed values, and don't tend to spend extended amounts of time on any one side of the mean. Thus, when the points are plotted and connected by line segments, it results in a series of successive line segments that typically alternate in the signs of their slope while also staying close to the value of the mean. This pattern is indicative of a chain that is mixing rapidly, and this is a characteristic of chains that are moving to convergence. In this example, the chain used was deemed converged by a variety of other criteria.

Quantifying 'hairiness'

In a modification of this work, Brooks (1998c) adds a quantitative measure of 'hairiness' to this method of characterizing the posterior chain obtained from an MCMC sampler in an effort to reduce the subjectivity of Yu and Mykland's (1998) method. Brooks (1998c) explains how the posterior, when characterized by way of Yu and Mykland's (1998) CUSUM path plot, can be transformed again into an indicator statistic that tries to capture the essence of 'hairiness'. Simply put, if a given element (S_i) in a CUSUM sequence is larger than its two immediate neighbors (S_{i-1}, S_{i+1}) it satisfies this 'hairiness' condition; also, if a given element is smaller than its two immediate neighbors it satisfies this condition. If a given element is not larger or smaller than its two immediate neighbors, then it does not satisfy the 'hairiness' condition. If the current element under consideration, S_i, is larger or smaller than its two immediate neighbors, a plot of these three points joined by two line segments would require that the line

segments have differently signed slopes. More formally, the indicator statistic, d_i, can be stated as:

$$d_i = \begin{cases} 1 & S_{i-1} < S_i, and\ S_i > S_{i+1}, \quad or \\ & S_{i-1} > S_i, and\ S_i < S_{i+1} \\ 0 & else \end{cases} \qquad (10)$$

Brooks (1998c) then describes how this indicator statistic can be evaluated as an accumulating average. More specifically, the summary of the indicator statistic, $\underline{D_t}$, is:

$$D_t = \frac{1}{t - n_0 - 1} \sum_{i=n_0+1}^{t-1} d_i, \quad for\ i = n_0 + 2 \leq t \leq n \qquad (11)$$

This summary of the indicator statistic is simply the proportion of elements in the CUSUM sequence that are characterized as being either larger or smaller than their immediate neighbors (i.e., the proportion of all S_i values in a chain classified by the indicator statistic as $d_i = 1$) Brooks (1998c) notes that the sum of the d_i values can be interpreted as the number of times that $\theta(x_i)$ crosses $\hat{\mu}$. D_t is calculated at each successive step of the chain, and can be plotted to inspect its behavior over the length of a chain. For D_t to function as a formal diagnostic, its characteristics must be evaluated under the assumption that the sequence is i.i.d. and symmetric about its mean. When these conditions are met, it is argued that the expected value of this statistic is ½ (see Brooks, 1998c for a proof). So, when a chain being inspected has reached convergence, its d_i

sequence should be centered at ½. Therefore, a hypothesis test can be performed, where the null hypothesis under consideration is that the expected value of d_i is equal to ½. This can be achieved by treating D_t as a binomial outcome with a mean of ½ and a variance of $¼(t - n_0 - 1)$.

This hypothesis test can be expressed in graphical form by plotting the value of D_t across the length of the chain. By plotting the D_t statistic over time against thresholds determined by a confidence band about the expected value of the statistic under the assumption that the null hypothesis is true (i.e., for a converged, i.i.d., symmetric distribution) the proportion of elements in the chain that satisfy the conditions that Brooks (1998c) associates with 'hairiness' can be seen. When the plot of a D_t sequence from an observed posterior falls within the boundaries implied by the null hypothesis of a stationary distribution, the chain from which it is derived is indistinguishable from a converged chain.

Brooks (1998c) bases this claim on the assumption that the summary of the indicator statistic can be described by a binomial distribution. The boundaries implied by the null hypothesis of a stationary distribution can be approximated by,

$$p \pm Z_{\alpha/2}\sqrt{pq\frac{1}{(t-n_0-1)}} \qquad (12)$$

where p is equal to $P(d_i=1)$, q is equal to $1 - p$, t is the total number of elements in the chain, and n_0 is the number of elements in the burn-in. This can be interpreted as a 100(1-

$\alpha/2$) % confidence interval because D_t will be approximately normal for large values of t by the law of large numbers. While observing a plot of D_t that falls within these boundaries is not necessarily indicative of convergence, a plot that falls outside the boundaries is indicative of lack of convergence. In other words, if a CUSUM sequence is characterized by way of the indicator statistic and the plot of its accumulating average for these indicator statistics falls within the boundaries implied by the confidence interval, then we have evidence to suggest that this particular pattern of ones and zeroes is converged. This ability to indicate chains which are not behaving as if they are converged can be used to ensure against the misuse of MCMC output in inferential settings.

In the case of MCMC chains it is typically not the case that the d_i sequence is i.i.d. because it is based on a dependent sequence of values from the Markov chain. Also, it is often not the case that the distributions in question are not symmetric about the mean. So, the bounds described above are not exact but only a rough approximation. In order to remove the two assumptions upon which the method rests (and make the bounds exact rather than approximate), Brooks (1998c) proposes two modifications. First, to remove the assumption that the sequence being characterized by the method is i.i.d., the observed Markov chain can be thinned. Thinning the chain to remove the dependence among elements makes the first assumption 'approximately' true. Second, to remove the assumption of symmetry, it is argued that $P(d_i=1)$ can be modified to reflect the asymmetry of the distribution in question. The modification is done by integrating across the transition kernel. Specifically, the number of observations expected to be greater than the mean is used to weight the transition kernel from the mean to the upper limit, and the

number of observations below the mean is used to weight the transition kernel from the lower limit up to the mean.

Brooks (1998c) argues that his quantification of Yu and Mykland's (1998) CUSUM path plot is a significant improvement over the original method. The new method removes some of the subjectivity involved in the original method and offers a quantitative assessment of how close an observed sequence is to convergence. Rather than simply assessing mixing rate, the approach is now capable of providing a hypothesis test for the convergence of Markov chains. Also, the method can be extended to determining the length of the burn in as well as estimating a thinning parameter for obtaining approximately i.i.d. sequences for use in drawing inferences.

Studies comparing diagnostics

Two studies will be addressed directly. Results of Cowles and Carlin (1996) and Sinharay (2003 and 2004) will be reported. Then general findings of attempts to assess convergence will be discussed.

Cowles and Carlin (1996) used both a Gibbs sampler and a reversible jump sampler to create Markov chains, and then applied 13 different convergence diagnostics to the output of the samplers. In this study, the emphasis was not on comparing the methods directly; rather it was to describe how each sampler characterized a chain. It is evident from this study that the different diagnostics weren't always applicable to the output of the samplers, the diagnostics often disagreed with one another, and there was no clear indication of which diagnostic made the most sense to apply to a chain. Even when a diagnostic was put into use in a situation that it was designed to be sensitive to, it did

not always succeed. This finding is especially alarming in that the chains assessed in this paper were from low-dimensional, highly idealized situations. If the techniques fail in these relatively simple settings, it does not bode well for their performance in more realistic, high dimensional problems. Additionally, a number of the diagnostics are not easily implemented due to their problem specific nature. Cowles and Carlin (1995) in their summary call for more research into the theoretical and applied aspects of MCMC algorithms.

These authors claim that multiple diagnostics (both visual and quantitative) should be employed and multiple chains be run for each parameter to be estimated. This 'blanket' type approach will help prevent a researcher from 'blindly' making statements about the quality of a chain (or chains) for estimation purposes. Indeed, it is recommended that the visual and quantitative diagnostics be considered simultaneously by adding the quantitative indices to the plots of the time series, for example. In addition, it is wise to consider multiple parameters simultaneously so as to shed light on the relationships among model parameters that may be influencing the estimation procedure.

These authors also advocate making revisions to the way a sampler is created to ensure that quality estimation is performed. Reference is made to strategies for creating the Markov chain that may help avoid some of the potential pitfalls known to exist. For example, Mykland, Tierney, and Yu (1995) insert an independent MH step every so often within a very long Gibbs sampling chain. In effect, when an independent MH candidate is accepted, this is equivalent to running multiple chains and allows for the application of diagnostics requiring multiple chains, but can also be treated as one very long chain

(which is preferred over multiple shorter chains).

Similarly, the authors encourage researchers to consider multiple models, multiple types of samplers, and a good deal of 'up front' work to investigate the target distribution before applying MCMC techniques. By taking the time to investigate the data from multiple perspectives, it is possible to gain a clearer picture of likelihood or posterior surface.

Sinharay (2003 and 2004) investigated several convergence diagnostics in the context of two psychometric examples. The practical motivation for this study is that any inference to be derived on the basis of estimated parameters is only justifiable if those estimated parameters are sensible. That is to say, the method for estimating the parameters must be valid in both principle and application. Any breakdown in the estimation procedure has potentially dire consequences for inferences to be made. In the context of MCMC estimation, the technique has been shown to be sound in principle, but there are few universal guidelines for exactly how to proceed in application. Thus, the research (comparing convergence diagnostics) is justified by way of argument that convergence diagnostics are an ideal source of information when trying to determine whether or not a sampler is behaving as it should. The focus of this research is not necessarily to determine when MCMC algorithms converge; rather it attempts to demonstrate the differences that exist when applying multiple diagnostics to assessing convergence.

To aid in the understanding and accessibility of the research, Sinharay (2004) limits the diagnostics investigated to those that are conceptually easier to understand and

that are easily implemented. This research includes the RL, HW and G diagnostics described previously. One motivation for the current research is to extend the analysis of these three diagnostics and to do it along with the inclusion of the CUSUM method as well as the technique being developed in this research.

A generally agreed upon finding is that no one method works well all the time. Multiple diagnostics should be applied to any chain intended to be used as the basis for an estimate. Diagnostics address necessity, and not sufficiency of qualities for a chain to be deemed converged.

Constructing the posteriors

When the chains have been constructed, some work still remains to be done before moving forward with estimation and inference. Thinning the chain, checking model fit, and performing model comparisons are all considerations that should be undertaken (Kim & Bolt, 2007).

Thinning the chain involves removing some of the elements from the final sequence. Thinning is not to be confused with burn in. Thinning should be done after the burn in has been removed Thinning is done to deal with the AC built into the chain by the MCMC procedure. For example, taking every n^{th} element from the chain reduces the amount of AC among the remaining elements, rendering them at least somewhat linearly independent. Thinning will received additional treatment in the methods section.

Evaluating model fit is prudent before proceeding with any inference based on the MCMC estimates. A benefit of engaging in MCMC sampling is that it is possible to use posterior predictive checks. Generally speaking, it is possible to create additional

posterior distributions from different data sets than those used to produce the parameters. This allows comparison of any of a relevant set of discrepancy statistics for the original data and the new data to determine whether or not the outcomes in the observed data are likely to happen given what is observed in the replicate datasets. This is similar to a re-sampling approach.

Model comparisons are very easy to implement in an MCMC framework. Two model comparison indices are the Pseudo-Bayes Factor criterion (PBFC) and the Deviance Information Criterion (DIC). A Bayes Factor criterion forms a ratio of likelihoods involving the data conditional upon either of two potential models. The value of the ratio tells you which model fits the data better (with the customary caveat that it doesn't tell you if the model fits well, just better than the other model under consideration). The Bayes Factor criterion can be approximated by a Pseudo-Bayes factor criterion. An example of this is the Conditional Predictive Ordinate (CPO). The CPO at the level of the individual item response is given as:

$$CPO^{-1} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{p(x|\Omega_k)} \qquad (13)$$

where *K* is the total number of states in the chain, *k* represents the particular state in the chain, and $p(x|\Omega_k)$is the conditional likelihood of the data given the parameters at state *k*. To summarize across all item responses, the CPOs can be multiplied and the log then taken of the product. A model producing a higher log product is preferred over one

producing a lower log product. The DIC is similar to Akaike Information Criterion and Bayesian Information Criterion in that it is based on the -2 log likelihood and model complexity is taken into account when assessing fit. Models with more parameter should fit better, so when comparing models with different numbers of parameters the model with more parameters is penalized more.

Why use MCMC

Now that a description of MCMC has been given and some of the important considerations that need to be addressed when implementing the technique have been discussed, it is worthwhile to briefly describe why MCMC should be used. In addition to discussing its strengths and utility, some of the weaknesses of the method will also be addressed.

When an analytical solution to a function is impossible or intractable, sampling strategies like MCMC allow for numerical solutions to calculations that are otherwise unobtainable. In psychometric settings the models used are often complex and have a high degree of dimensionality. MCMC allows the user to reduce complex multidimensional problems to a sequence of lower dimensional problems (Cowles & Carlin, 1996). Thus, the traditionally accepted approach of Marginal Maximum Likelihood Estimation is stymied by these complexities and an alternative is needed. For example, the E-M algorithm can become difficult to implement with complicated models. Not only is the method itself easier to implement, MCMC extends easily to more complicated modeling situations. For example, when data is missing and augmentation is needed, this is straightforward to do in MCMC (Patz & Junker 1999b).

Also, in MCMC we are estimating distributions, not single point estimates. The fact that the entire posterior is available provides richer information about parameter estimates, although a point estimate is often the final destination for most practitioners in EM. For example, a histogram of the posterior distribution can be very informative when estimating parameters.

Finally, MCMC is relatively indifferent to the presence or absence of conjugate structure between the likelihood and the priors (Cowles & Carlin, 1996). Thus, even if the priors are mis-specified, MCMC samplers are still likely to converge to the stable, underlying distribution. So, even if we have little or no idea what kind of distribution we are trying to estimate, we can still proceed with estimation.

The MCMC method also has some weaknesses. The primary drawback of the method is that it isn't entirely clear when and if these methods reach the stationary chain that is supposed to represent the true joint posterior distribution for the parameters we wish to estimate. Convergence here is much more general than in competing estimation procedures such as MLE, where convergence is to a point. Convergence in MCMC is to a distribution (of which only a sample is ever observed). Adding to the difficulty of making statements about convergence is the reality that samples estimated in this fashion typically are linearly dependent upon one another. The result is longer estimation runs (as the procedure is limited in its efficiency in exploring the parameter space) and an unclear estimate of the variance of the chains produced (the strong linear dependence results in underestimated values of the variance). The large amount of AC (dependence within a chain) and cross correlation (dependence across chains of separate parameters) can be

caused by poorly or overly parameterized models. Re-parameterizing a model is not unique to MCMC settings, but does potentially pose additional problems before estimation can be deemed trustworthy.

MCMC is also computationally demanding. Run times for MCMC samplers can easily extend to hours and days. In situations where time is of the essence, MCMC approaches may be prohibitively time-consuming and therefore impractical. Although, as computing speed increases, the fact that MCMC is computationally demanding will become less of a concern.

MCMC is less well understood than MLE approaches. So, in cases where existing software is available using an MLE approach, it obviates the need to take an approach like this, and can save a considerable amount of time. Also, in high stake testing situations where the results of testing must be legally defensible, much work remains to be done to guarantee the legitimacy of MCMC procedures to the public at large.

Motivation for the current research

MCMC is being used as a method of estimation and as a result it is important that it is used correctly. Part of using MCMC correctly is having solid evidence to know that chains are converged and estimation can be substantiated. Many complex models exist and more are being developed for which MLE approaches will be difficult to implement. Because of the increased commonality of its use, there is an increasing need to be sure that the estimates obtained via MCMC are stable and trustworthy. The estimates must be stable and trustworthy if it is to be applied in real testing situations.

Despite convergence diagnostics having been characterized as 'fundamentally unsound', it is still necessary to investigate the characteristics of the chain associated with its suitability to act as a sample on which to base an estimate that is then to be used for inference regarding examinee responses. If convergence cannot be solved for in an a priori theoretical fashion, the next best thing is a diagnostic (or group of them) that can rule out potentially 'bad' chains. Although it is never possible to say with certainty that a finite sample from an MCMC algorithm is representative of an underlying stationary distribution, convergence diagnostics "…may offer a worthwhile check on the algorithm's progress (Cowles & Carlin, 1995, p. 903)." If researchers are to use MCMC methods and interpret the results, it is desirable that (at the very least) the output should look somewhat like what would be expected if indeed stability had been achieved and the sample obtained was indistinguishable from one obtained from a sensible distribution. Convergence diagnostics are formalized statements of 'what we would expect to see.'

Existing convergence diagnostics aren't perfect, and no one method is a panacea. All methods developed to date work well in some situations and not in others. Not all methods are easily implemented or efficient. Methods have different theoretical justifications which may or may not make sense in a given context. The relatively limited literature describing convergence diagnostics doesn't always address when certain diagnostics are more or less appropriate and/or successful. It is commonly noted that both visual and quantitative techniques should be used (Cowles & Carlin, 1996, Sinharay, 2004). This suggestion is a good motivation to develop a method that incorporates both visual and quantitative components, like the CUSUM or the technique being developed.

Visual inspection of the chains is commonly done; however, this is a time intensive prospect. A method that identifies chains which are likely suspect can greatly simplify the demand on the researcher.

Existing studies comparing and contrasting convergence diagnostics aren't exhaustive. The particular sampling methods are typically not compared directly to one another. There is a dearth of these types of studies in general, and especially for those relating specifically to psychometric applications. Thus, the particular goals of this research endeavor are to describe the development of a new convergence diagnostic (based on modifications of the CUSUM diagnostic), describe the development of a technique to generate chains without running a sampler, describe clearly the relationship between AC, balance, and the expected value of the indicator statistic under development, and to compare this method to other comparable diagnostics with simulated chains as well as with chains from real samplers.

The importance of modifying the existing CUSUM technique rests on removing the assumption of an underlying symmetric distribution. A straightforward way to modify the technique that alleviates the need for an assumption of symmetry will make the technique more widely applicable, and is discussed thoroughly in the Methods section. A technique for generating chains will also be described thoroughly in the Methods section. Generating chains allows for a convenient descriptive mechanism by which to discuss the relationship between AC, balance, and the expected value of the summary of the indicator statistic, D_t. Also, generating chains allows for a controlled simulation study to

directly compare methods of assessing convergence.

CHAPTER III

METHODS

The primary goal of this study is to modify and subsequently investigate a convergence diagnostic for posterior distributions obtained through MCMC estimation and compare it to some existing diagnostics. The new method under consideration is based on the cumulative sum (CUSUM) procedure of Yu and Mykland (1998) and the subsequent modifications made by Brooks (1998c). In particular, it will be shown that this new index is similar to the aforementioned index discussed by Brooks without requiring that the posterior distribution be symmetric.

In order to achieve the goal of this study, two things must be accomplished. First, the method under development must be thoroughly described so as to allow for a complete understanding of its implementation and implications of use. For example, the properties of the convergence diagnostic must be described for the cases of independent versus dependent sequences of elements and for the cases of stable versus unstable generating distributions for the sequences of elements. Second, a method for simulating chains with controlled amounts of autocorrelation among elements and controlled movement of the mean is needed. In this way, it is known ahead of time how the chain is behaving so that the effectiveness of the methods can be compared accurately. It is necessary to simulate the chains so that they can range from completely independent draws to completely dependent draws. This is accomplished by controlling the degree of

autocorrelation present in the simulated chains. It is also desirable to simulate chains where the mean is stable and those where the mean is fluctuating. This stability, or balance, is controlled by the random sampling component of the simulated chains that will allow for control of the stability of the mean of the simulated chains.

Modification of the CUSUM procedure

The motivation of this research is to develop a convergence diagnostic that is directly informative about the stability of the distribution producing a Markov chain. If a technique can be developed that is sensitive to a conditional distribution that is unchanging, the implication is that the next random draw in the sequence must be from the same distribution. The current diagnostic is different than the CUSUM indicator statistic developed by Brooks (1998c). The primary modification is that the quantitative indicator statistic, d_i, and its summary, D_t, should be applied to the observed Markov chain, rather than the CUSUM chain. Because the indicator statistic is no longer computed using the CUSUM chain, the expected value of the summary of the indicator statistic under the null hypothesis of an i.i.d. sequence must be derived. In addition, unlike the statistic described by Brooks (1998c), the assumption of symmetry for the distribution describing the chain under consideration of the indicator statistic is no longer necessary. Additionally, this study will also consider the effect of thinning of the chains before characterization by way of the indicator statistic.

The first modification to Brooks' (1998c) method is the particular chain to which the indicator statistic is applied. The current technique applies Brooks (1998c) d_i statistic to the observed Markov chain rather than the CUSUM chain. This affects the expected

value of the indicator statistic under the null hypothesis (i.e., an i.i.d., stationary sequence is achieved) in addition to its properties. When the indicator statistic is applied to the observed Markov chain, the expected value of d_i is equal to 2/3 rather than 1/2. Brooks (1998c) makes the argument that the sum of the indicator statistic, d_i, can be interpreted as the number of times that the CUSUM plot crosses the mean. Thus, in the case of a converged, symmetric distribution it would be expected that each new observation is equally likely to be above or below the mean. However, this interpretation is due to the fact that the indicator statistic is applied to the CUSUM chain, which is an accumulating sum of mean centered values. However, when the indicator statistic is applied to the observed chain, it is not necessary that any of the values actually cross the mean to be coded as a 1, it simply needs to be greater than or less than its two immediate neighbors in the chain. The indicator statistic is only concerned with rank ordering, and does not directly involve comparison to the mean when applied to the observed Markov chain. The rank ordering of the elements is indicative of the stability of the distribution used to generate them. Applying the indicator statistic directly to the observed Markov chain is providing information about the probability of observing particular rank orderings of the chain elements, and this is indicative of the stability of the random process generating the chain elements.

The justification for the value of 2/3 comes from an argumentative proof. When any group of three i.i.d. variables is considered in terms of their rank orderings (as is done with d_i), the middlemost element is capable of taking on the 1st ranking (largest value), the 2nd ranking (neither the largest or smallest value), or the 3rd ranking (smallest

value). When there is a converged i.i.d. sequence each rank ordering is equally possible for any of the three variables under consideration. When an observation takes on a 1st or 3rd place ranking in relation to the observations immediately before and after it, it satisfies the condition of being set equal to 1 for the indicator statistic. When an observation takes on a 2nd place ranking in relation to the observations immediately before and after it, it satisfies the condition of being set equal to 0 for the indicator statistic. Considering these facts, for any i.i.d. sequence of three, each element is equally likely to take on each of the possible rankings of 1st, 2nd, or 3rd. In other words, it is expected that two out of the three equally likely possible rankings of the sequentially middlemost value to satisfy the condition of being coded a 1, while one in three outcomes would be coded a 0. Thus, rather than centering the threshold about the value of 1/2 as previously suggested, it is argued that the expected value of the $\underline{D_t}$ statistic under the null hypothesis is 2/3. So, when we apply the indicator statistic to the observed Markov chain, we have a different expectation about what value it should take on if the process has indeed converged.

The second modification to Brooks (1998c) has to do with the thresholds about the summary of the indicator statistic. Under the null hypothesis, the standard error of the estimator of the indicator statistic, *p*, is given as, SE_p:

$$SE_p = \sqrt{pq\frac{1}{t - n_0 - 1}} \qquad (14)$$

where *t* indicates the step in the sequence, and all other terms are as described previously. The thresholds are estimated in a similar fashion as that suggested by Brooks (1998c). The only difference is the value that *p* (i.e., $P(d_i=1)$) is expected to take on under the null hypothesis. Thus the threshold is centered about a different value (2/3 rather than 1/2), and this also affects the width of the interval, as this is a binomial variable and takes on maximum variance at the value of $P(d_i=1) = .5$. This results in a slightly more conservative set of bounds than that originally proposed as $p\ (1-p)$ will decrease in size as p moves away from .5.

$$p \pm \left(z_{\alpha/2} \cdot SE_p\right) \qquad (15)$$

This equation shows how the thresholds are calculated. The threshold values are plotted on the same graph as the observed value of D_t across the chain. This allows for direct visual comparison of the observed chain to thresholds representing the null hypothesis throughout the length of a chain. Also, as further evidence that this value of 2/3 is the expected value in the case of convergence; data will be generated from several well-known distributions (both symmetric and asymmetric) and characterized by way of the indicator statistic. In all of these cases, the value of D_t goes to 2/3 as the number of observations increase.

The third modification of Brooks (1998c) is specific to the assumptions made about the shape of the posterior distribution. Brooks (1998c) showed how it was possible to remove the assumption of symmetry from his technique. However, this involves

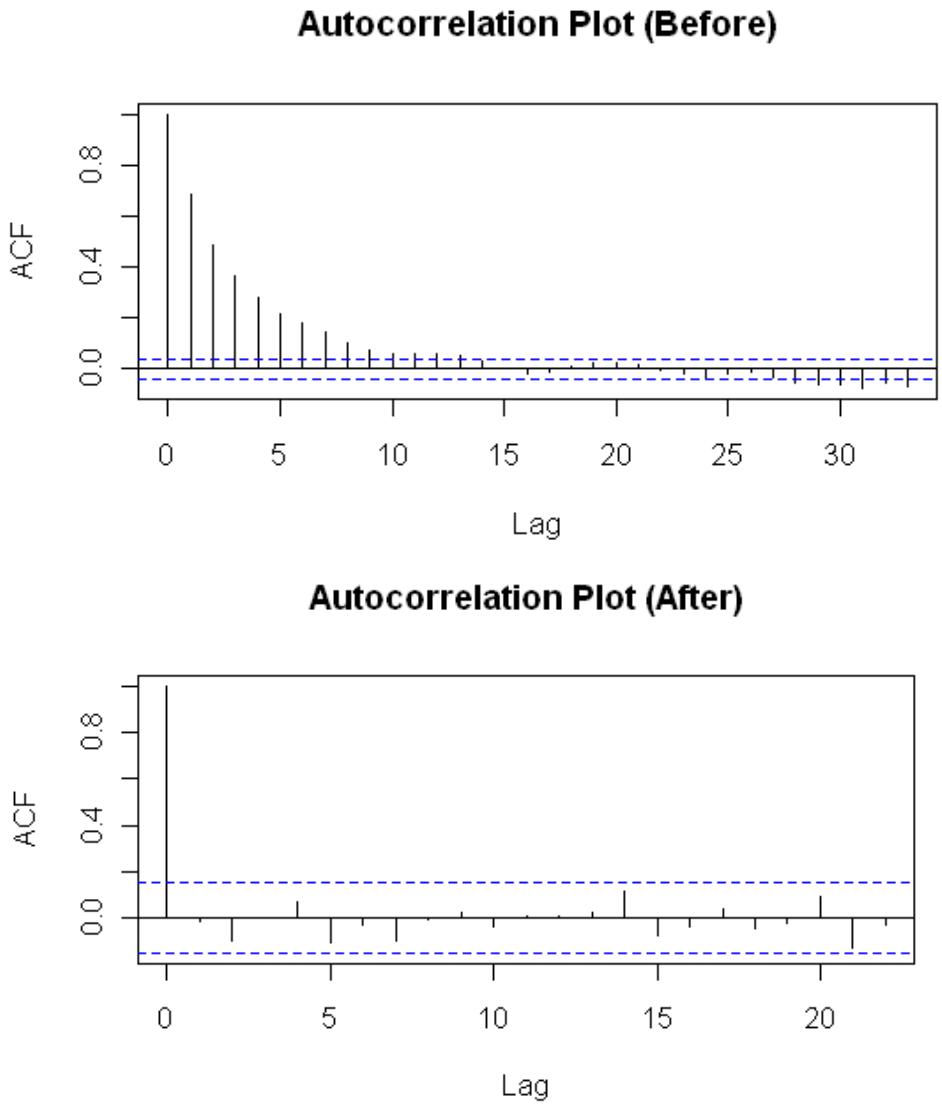
integrating across ranges of the transition kernel which may not be straightforward. The current research provides an alternative way to remove the assumption of symmetry. While it was argued that the shape of the posterior has an influence on the expected value of the summary of the indicator statistic under the null hypothesis, this is not true when the indicator statistic is applied to the Markov chain rather than the CUSUM chain. The current approach makes no assumption about the shape of the posterior; only that the posterior distribution could be characterized by a cumulative distribution function (CDF). This assumption is justified because the indicator statistic d_i is characterizing the elements by their rank orderings.

CDFs are informative about the likelihoods of particular rank orderings. A CDF represents the percentile ranks of the values of a random variable, and percentile ranks are by definition uniform distributions. Under the condition that the null hypothesis is true, any proper density function that accurately characterizes the posterior would have the same expected value for the summary of the indicator statistic. Thus, any distribution that can be characterized by a CDF should be accurately described by the null thresholds of this indicator statistic, which only takes into consideration the rank order of the Markov chain elements. The probability density function (PDF) describes the absolute relationship that the values of a random variable can take on; however, the CDF is simply concerned with and represents the rank ordering of the values of the random variable. So, if a Markov chain has converged to a stable distribution, the rank orderings of the sequences will be predictable. Thus, the summary of the indicator statistic should be informative about the process producing observed elements in a Markov chain.

The last modification to the method proposed by Brooks (1998c) has to do with the method used for thinning the observed sequence. Recall that the boundaries suggested by Brooks (1998c) are based upon the null hypothesis that we have an independent set of draws. If a posterior chain is to be used as the basis for a parameter estimate, something must be done to ensure that the chain is at least an approximately i.i.d. string of observations before being transformed into the indicator statistic, otherwise the null boundaries are not appropriate and any attempt to diagnose lack of convergence will be misleading. Yu and Mykland (1998) point out that the MCMC procedure builds in a great deal of linear dependence among the observations. To deal with this, the degree of autocorrelation present in the observed chain is used to determine how many elements must be removed to achieve a non-significant autocorrelation at lag 1, and thins the chain accordingly to leave a linearly independent sequence. For example, if at lag 15, the correlation between observations is not significantly different than zero, then by taking only every fifteenth element of the observed posterior for inclusion in the thinned posterior, a linearly independent sample of observations will be obtained. This linearly independent group should also be identically distributed if the sampling is done appropriately and the chain has had enough time to adequately explore the sampling space. The thinning is done using the autocorrelation function in R, (ACF; http://www.r-project.org/). For example, the autocorrelation for an observed chain beyond burn in is shown below in Figure 2, before and after thinning via the ACF. This chain was produced using a MHA within Gibbs sampling procedure. Plots of the autocorrelation functions for stationary Markov chains reveal that they often display exponential behavior

(decelerating curves). The behavior of the autocorrelations is governed by the fact that each element of a Markov chain is dependent only upon the previous state (Hanson and Cunningham, 1998). Thus, when the lag is small (i.e. elements are close together in the chain), the AC tends to be high. The AC is reduced as more intervening elements exist.

Figure 2: Observed autocorrelations in a chain before and after thinning



In Figure 2 ‘Before’, the autocorrelation plot shows the great degree of linear dependence that exists among observations close together in an observed chain from the MCMC sampling procedure using a MH within Gibbs sampler. In this example, the correlation is clearly not significantly different from zero at a lag of 15. So, to obtain a linearly i.i.d. subsample, observations are removed from the chain by choosing a random element that is close to the mean and then taking every fifteenth element after that for selection into the thinned chain. When the autocorrelation is again calculated for the chain after thinning, it can be seen in Figure 2 ‘After’ that the chain is now a linearly independent string of observations (i.e., at a lag of 1, the correlation is not significantly different than zero). The process of thinning is done so that the boundaries proposed by Brooks represent an approximate null hypothesis to which the value of an observed chain can be compared. If the plot of the value of D_t over the length of the observed chain falls within those boundaries, the current chain is behaving similarly to what we would expect a sequence generated from a converged distribution to do. Brooks (1998c) also proposes a method for estimating the thinning parameter; however, this method involves integration of the transition kernel, which may not be feasible. The method for thinning currently provided is simple to implement.

The problem with thinning

However, previous work in a related field has provided evidence that the current technique is at least partially flawed. Particularly, thinning of the chains is undesirable because any sub-sampling of the posterior distribution can be shown to degrade the quality of the estimate obtained (MacEachern and Berliner, 1994). The logic of this

criticism is that the purpose of engaging an MCMC framework is to produce an estimate. Convergence techniques typically address bias of the observed Markov chain. Thus, any activity that introduces bias into the final estimate is contrary to the concept of assessing convergence in this fashion. Additionally, even if the logic of this argument is flawed, it is not desirable to have a biased estimate of a parameter if there is no need to do so.

To address this potential flaw in reasoning, further investigation into the use of the CUSUM as a convergence diagnostic is required. It is reasonable to attempt to modify the technique such that thinning is not necessary. This requires that the relationship between the degree of dependency and the value that the summary of the indicator statistic takes on be clearly explicated. If this relationship is known, it is possible to use the technique without thinning the Markov chain. A slightly different version of the current method for assessing convergence attempts to make characterizations about convergence without thinning the observed posterior distribution. The motivation for taking this approach comes from a criticism of the practice of sub-sampling the MCMC sample. Put simply, any subsample of the sampled chain will produce a poorer estimate of the parameter than the full chain. MacEachern and Berliner (1994) demonstrated this fact by way of a simple proof. Additionally, it is a wasteful practice to generate a sample of *n* elements only to end up using less than all *n* elements. (A potential alternative would entail thinning the chain in order to assess convergence, but then using all elements of a chain to estimate model parameters.)

AC and D_t

In the absence of thinning, an important technical challenge faced by this approach is to understand the influence of autocorrelation (AC) on the value of the summary of the indicator statistic. Understanding the effect of thinning is important because the MCMC procedure results in a linearly dependent sequence of elements To understand the influence of autocorrelation it is necessary to first understand how to describe the probability that the indicator statistic, d_i, takes on a value of 1. When considering 3 elements in a chain (let's call them X_1, X_2, and X_3, respectively), there are two patterns of rank orderings for these three variables that would result in a d_i value of 1. The first of these two patterns is when X_2 is greater than X_1 and X_3 is less than X_2. The second of these two patterns is when X_2 is less than X_1 and X_3 is greater than X_2. We can write these patterns as:

$$P(d_i = 1) = P(X_1, X_2 > X_1, X_3 < X_2) + P(X_1, X_2 < X_1, X_3 > X_2) \qquad (16)$$

The value that the indicator statistic takes on can be described as the integration across particular ranges of the three elements under consideration. Again, to fully describe the probability that the indicator statistic d_i takes on a value of 1, we need to think about the integration for the situation where the second element under consideration is the smallest as well as where it is the largest. In regards to the second of three elements under consideration being the smallest, this involves integrating the first element across the full range of values it can take on, then integrating the second element from the lower

bound up to the value of the first element, and then integrating the third element from the value of the second element up to the upper limit. For the case where the second element is the largest, the bounds for the integration change slightly. When the second element is the largest of the three elements under consideration, this again involves integrating the first element across the full range of values it can take on, then integrating the second element from the value of the first element up to the upper bound of the second element, and then integrating the third element from the lower bound up to the value of the second element. In general this can be expressed as

$$P(d_i = 1) = \int_{lb}^{ub} f(X_1) \int_{lb}^{X1} f(X_2|X_1) \int_{X2}^{ub} f(X_3|X_2) dX_3\, dX_2 dX_1$$
$$+ \int_{lb}^{ub} f(X_1) \int_{X1}^{ub} f(X_2|X_1) \int_{lb}^{X2} f(X_3|X_2) dX_3 dX_2 dX_1 \qquad (17)$$

When a sequence under consideration is i.i.d., it is simple to express the expected value of D_t by way of expected rank orderings of the elements under consideration. However, when the sequence shows dependence of the type commonly encountered in MCMC settings, it is not so easy to express the expected value of the summary of the indicator statistic. To understand the influence that linear dependence will have on the expected value of D_t it is helpful to thoroughly consider the mechanism developed for simulating chains.

Simulating chains

The simulated chains must be realistic representations of those obtained from MCMC samplers. Thus, there must be some degree of dependence among the elements, as this is a common occurrence when using MCMC estimation. To achieve this, each element in the chain will be set equal to the previous value plus some random component generated from known distributions. Additionally, the chains need to be simulated in such a way that allows the mean to move in a predictable fashion.

The formula used to simulate these chains is given by:

$$X_i = c \cdot (X_{i-1}) + U(-(1-d), d) \qquad (18)$$

The value of *c* ranges between 0 and 1. When *c* takes on a value of one, there is a strong degree of autocorrelation present in the chain. When *c* takes on a value of zero, there is no significant autocorrelation in the chain. The value of *d* will be manipulated so as to control whether or not the random component added to the simulated chain values are more or less likely to be positive or negative. This is in essence controlling whether or not the mean of the chain values is moving up or down. As for the fluctuation of the means, there will be two conditions. In one case, the random component of the formula used to simulate chains will be devised so that the mean of the sequence will move randomly up or down with equal probability. This imitates a random walk where the mean is equally likely to move up or down. In the other case, the random component of the formula used to simulate chains will be devised so that the mean will be more likely

to move up (or down) over time. When these random components added to the previous chain element are equally likely to be positive or negative, it will be called a 'balanced' proposal. When these random components added to the previous chain element are more likely to be positive (or negative), it will be called an 'unbalanced' proposal distribution.

With this in mind, we define the probability density function (pdf) for any element in the chain, X_i, as uniform with lower and upper bounds of cX_i–(1-*d*) and cX_i + *d*, respectively. Thus, the pdf is:

$$f(X_i) = \frac{1}{[cX_i + d] - [cX_i - (1-d)]} = 1 \qquad (19)$$

In the case of a continuous uniform distribution with lower and upper bounds of *c*X_i - (1-*d*) and cX_i + *d*, respectively, this integration can be written as (with a slight change in notation to make things more general):

$$P(d_i = 1) = \int_{cX_i-(1-d)}^{cX_i+d} 1 \int_{cX_i}^{cX_{i+1}+d} 1 \int_{cX_{i+2}-(1-d)}^{cX_{i+1}} 1 \; dX_{i+2}dX_{i+1}dX_i$$
$$+ \int_{cX_i-(1-d)}^{cX_i+d} 1 \int_{cX_{i+1}-(1-d)}^{cX_i} 1 \int_{cX_{i+1}}^{cX_{i+2}+d} 1 \; dX_{i+2}dX_{i+1}dX_i \qquad (20)$$

Solving these integrals with various values in the range of *c* and *d*, we can see the influence of autocorrelation and balance on the value of the summary of the indicator statistic. When the autocorrelation is strong (c = 1), a balanced random component leads

to a value of the summary of the indicator statistic of .5. In this case, each new element is strongly influenced by the previous value (in fact it is a random variable centered at the value of the previous element) and is equally likely to be larger or smaller than the previous element. In terms of convergence, the type of chain produced by this arrangement of *c* and *d* is a true 'random walk' in that the mean of the chain is likely to wander up and down and not settle in to a specific location. In other words, each and every segment of the chain is likely to have a different mean. This is the antithesis of 'convergence.' When the autocorrelation is strong (c = 1) and the random component is unbalanced, the value of the summary of the indicator statistic will be less than .5, and will move to 0 when the random component is completely unbalanced (i.e., the random component is always positive (or negative)). In this case, each new element is strongly influenced by the value of the previous element, but the mean of the chain is more than likely to be increasing (or decreasing). In terms of convergence, the chains produced by this arrangement of *c* and *d* are ones which have not settled in to a stable location and are still on the move (i.e. still in the burnin). When there is no autocorrelation present in the simulated chain (i.e., c = 0), then regardless of the balance of the random component, the summary of the indicator statistic will be equal to .67. When making use of an MCMC sampler, this is the ideal case that one would like to see. This indicates a set of completely independent draws with a stable mean and variance.

Research questions revisited

Now that the modified CUSUM convergence diagnostic has been introduced and the goals for the current research have been provided, the specific research questions to

be addressed will be stated. In addition to the research questions, the proposed analyses that will attempt to provide answers to those questions are included.

The first research question that will be addressed is: What is the relationship of the degree of autocorrelation among elements, the balance of the random component in the chain simulator, and the value that the summary of the indicator statistic, $\underline{D_t}$, takes on for the current method? To answer this question, a closed form solution for the value of the summary of the indicator statistic, D_t , for the case of the continuous uniform distribution as the random component of the chain simulator will be presented. This clearly defines the relationship among the degree of autocorrelation (c), the degree of balance of the random component (d), and the value of the summary of the indicator statistic, D_t. It will be shown that the degree of autocorrelation present in the chain has a mediating effect on the influence of balance on the value of the indicator statistic. Again, when there is no thinning of the chain, we need to have a clear understanding of how the indicator statistic is likely to behave.

The first question will also be verified empirically by simulating chains as described previously. In simulation study 1, a 6 by 5 fully crossed factorial design will be used. The first factor is the value of c, which is the proportion of the previous chain element contributing to the following element in simulation. There are six levels of this factor, and they are c = 0, .25, .5, .75, .9, and 1. When c is equal to 0, there is no dependence among chain elements. When c is equal to 1, there is a strong degree of dependence among elements in the chains. The second factor is the degree of balance present in the random component of the chain simulator. There are five levels of this

factor, ranging from 0 to 1 in increments of .25. When d is equal to .5, the random component of the chain simulator is equally likely to be positive or negative. When d is equal to .25 or .75, the random component generates values that are twice as likely to negative or positive, respectively. When d is equal to 0 or 1, the random component is always negative or positive, respectively. The dependent variables will be the value of the summary of the indicator statistic proposed in this paper and the value of the summary of the indicator statistic proposed by Brooks (1998c). The length of the sequence used for each condition will be 10,000, and there will be 25 replications of each condition. The purpose of this simulation is to demonstrate empirically that the value of the summary of the indicator statistic being developed in this paper is affected by the amount of autocorrelation present in the chains and the stability of the location of the chains.

The second research question that will be addressed is: What effect does thinning the Markov chain have on the 'diagnosis' of convergence/non-convergence for the summary of the indicator statistic being developed? Answering this question can be achieved by simulating chains with varying degrees of AC and balance and comparing the value of the summary of the indicator statistic for thinned and un-thinned chains when applied to these chains. In simulation study 2, a 6 by 3 by 2 fully crossed factorial design will be used. The first factor is the degree of autocorrelation present in the generated sequence as controlled by c, and it will have five levels ranging from strong autocorrelation to an independent sequence (*c = 0, .25, .5, .75, .9, 1*). The second factor is the degree of balance in the random component of the generated sequences of values, and there will be three levels; balanced (d = .5, random component added to previous element

in the sequence is equally likely to be positive or negative), imbalanced (d = .75, random component added to previous element in sequence is twice as likely to be positive as it is negative), and completely imbalanced (d = 1, random component added to previous element in the sequence is always positive). The third factor is the method being used to characterize the sequence, and it has two levels. The first level is the method that thins the chains before characterization by way of the current method, and the second level is the method that does not thin the chains before calculating the summary of the indicator statistic. The dependent variable is the value that the summary of the indicator statistic being developed in this research takes on. The sequences used for this simulation will be 10,000 elements long, and there will be 25 replications of each condition. (This research question should also be informed by the first research question.) The second question can also be addressed by applying the new convergence diagnostic to the thinned and un-thinned chains from real MCMC samplers with varying ratios of variances for the proposal and target distributions. Simulation study 3 will be a 3 by 2 fully crossed factorial design. The first factor is the ratio of standard deviations for the proposal and target distribution, and it will have three levels. The first level will have the standard deviation of the proposal distribution as ¼ of the standard deviation of the target distribution. The second level will have the standard deviation of the proposal distribution equal to that of the target distribution. The third level will have the standard deviation of the proposal distribution four times larger than that of the target distribution. The second factor is the method being used to characterize the sequence, and it has two levels. The first level is the method that thins the chains before characterization, and the

second level is the method that does not thin the chains before calculating the summary of the indicator statistic. The dependent variable will be the value that the summary of the indicator statistic takes on. Each Markov chain will be run to 10,000 steps, and there will be ten replications of each condition of this study.

The third research question that will be addressed is: How does the new method compare to the Geweke (1992), Heidelberger and Welch (1983), Raftery and Lewis (1992), and Brooks (1998c) in terms of rates of convergence/non-convergence? This question can be answered in a straightforward way. In simulation study 3, chains of varying AC and balance will be generated and then convergence will be diagnosed by each method. The conditions for this simulation study will include three factors, and it is a 6 x 3, x 2 fully-crossed, factorial design. The first factor is the degree of autocorrelation present in the simulated chain. It will have six levels ranging from no autocorrelation to very strong autocorrelation (c = 0, .25, .5, .75, .9, and 1). The second factor will be the degree of balance in the proposal distribution. This factor will have three levels ranging from completely balanced to completely imbalanced as described previously for the second research question. Controlling the amount of balance will be accomplished by making it such that the balanced condition has a random component that is equally likely to add a positive or negative value to the next element in the chain, the imbalanced condition is twice as likely to add a positive value to the next element in the chain, and the completely imbalanced condition is always going to add a positive value to the next element in the chain. The third factor is the range of values for the random component added to the next element in the chain. This factor will have four levels (range = .1, .5, 1,

5). This fully crossed combination of factor levels results in a 6 x 4 x 3 design meaning there will be 72 conditions. Each sequence of elements will then be characterized by each of the diagnostics. The dependent variables for this study will be the degree of consistency of diagnosis by the different methods (as represented by Cohen's Kappa). Each sequence generated will have 10,000 elements, and there will be 25 replications of each condition.

CHAPTER IV

RESULTS

The results of the five studies addressing the three research questions will be presented. Where appropriate for each study, the relevant data will be portrayed in both graphs and tables. For the sake of economy, graphs for all twenty five replications of each unique combination of factor levels will not be presented. Instead, an example will be provided for each unique combination of factor conditions to portray the trends observed in the data. All data relevant for explaining the trends seen in the studies will be presented and described so as to accurately portray the effects of the manipulations upon the dependent variables of interest. The results will be presented in the order they were proposed.

Findings for research question 1

The first research question attempts to describe the influence that the degree of balance in the random component, d, and the degree of autocorrelation, c, have on the value of the summary of the indicator statistic, D_t. The factor d represents the degree of balance present in the random component of the chain simulator. There are five levels of this factor, ranging from 0 to 1 in increments of .25. When d is equal to .5, the random component of the chain simulator is equally likely to be positive or negative. When d is equal to .25 or .75, the random component generates values that are twice as likely to negative or positive, respectively. When d is equal to 0 or 1, the random component is

always negative or positive, respectively. There are six levels of the factor c, and they are c = 0, .25, .5, .75, .9, and 1. When c is equal to 0, there is no dependence among chain elements. When c is equal to 1, there is a strong degree of dependence among elements in the chains.

The analytical solution based on the integration across ranges of three variables representing the two possible patterns that satisfy the indicator statistic being to one is provided first. The simulated solutions will be presented afterward. After each set of results is presented a description of the important trends will be provided. In particular, the effect of c and d on the value of D_t will be the primary focus.

To begin, a reminder of the levels of the conditions may be helpful. The values of d correspond to the degree and direction of balance for the conditions. When d is equal to 0, it represents a completely imbalanced chain that is always decreasing in the random component because the random component is constrained to always be negative. When d is equal to .25, it represents a partially imbalanced condition that is more likely to decrease rather than increase in the random component because the random component is twice as likely to be negative rather than positive. When d is equal to .5, it represents a balanced condition that is equally likely to increase or decrease in the random component because the random component is equally likely to be positive or negative. When d is equal to .75, it represents a partially imbalanced condition that is more likely to increase rather than decrease in the random component because the random component is twice as likely to be positive rather than negative. When d is equal to 1.0, it represents a completely imbalanced chain that is always increasing in the random component because

the random component is constrained to always be positive. The AC factor, c, represents the proportion of each element in the chain that contributes to the value of the following element. For example, when c is equal to .5, then one half of the current value is added to the random component to determine the next element in the chain. Thus, c is controlling the amount of autocorrelation present in the chains. When c is equal to 0, the chain produced for any level of d will be i.i.d. sequences. As c increases to 1, the autocorrelation increases among the elements in the chains.

The formula for the closed form solution for the value of D_t was presented in the methods. The solution for the integration across ranges of the three random variables under consideration for transformation to the indicator statistic, d_i, is presented in three tables. These findings are presented in three tables for the sake of clarity. An important finding is that the value of D_t differs depending on the degree of balance, the amount of autocorrelation present, and the particular pattern satisfying the indicator statistic being set equal to one. In Table 1, the closed form solution is presented for the pattern where the second element under consideration is the largest of the three elements (LHL). In Table 2, the closed form solution is presented for the pattern where the second element under consideration is the smallest of the three elements (HLH). In Table 3, the closed form solution for the sum of both patterns is presented (BOTH). Each table has the levels of factor c represented as rows, and levels of factor d represent the columns. After each table is presented it is described in detail.

Table 1: Values of *Dt* for combinations of balance (d) and AC factor (c) for HLH

	D				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.333	.333	.333	.333	.333
.25	.352	.335	.314	.289	.258
.50	.313	.316	.297	.254	.188
.75	.201	.269	.277	.222	.107
.90	.092	.225	.262	.202	.047
1.0	.000	.188	.25	.188	.000

Table 1 provides evidence for the fact that the degree and direction (increasing versus decreasing) of the imbalance, as well as the degree of autocorrelation have an influence on the value of the summary of the indicator statistic, D_t. When c is equal to zero (i.e. converged chain), all conditions of d are equal to one another for the value of D_t. The value that D_t takes on (.33) is also what would be expected for an i.i.d. sequence when considering this pattern (HLH). As c increases to 1, the value that D_t takes on for a given level of c differs depending on the level of d and the direction of the imbalance. For the HLH pattern in general, D_t decreases in value as c goes from 0 to 1 and does so to a greater degree when imbalance is present. Additionally, the direction of the imbalance (positive, d > .5, or negative, d < 0) influences the rate of decrease for D_t. The complete imbalance conditions (d = 0 and 1) decrease at the greatest rate as c increases, but when the pattern HLH is being considered the decrease is greater across levels of c for positive imbalance than it is for negative imbalance. When the imbalance is negative (d = 0), the

pattern is a bit different, and will be described shortly. The partial imbalance conditions (d =.25 and .75) decrease at a lesser rate than the complete imbalance conditions as c increases. However, the trend still holds that when the pattern HLH is being considered the decrease is greater across levels of c for positive imbalance than it is for negative imbalance. When the random component is balanced, the rate of decrease of D_t is the smallest. There are two cases that do not follow the pattern as described here. When d is equal to 0 or .25 (negative imbalance) and c is equal to .25, D_t increases in value compared to the other conditions. When c is equal to .25 and d is equal to 0 and .25, D_t is equal to .352 and .335, respectively. These two exceptions imply that as the analytical solution is stated, it would be expected to code more elements as ones according to the indicator statistic when there is as mild a degree of autocorrelation for a chain that is imbalanced in a decreasing direction when considering the pattern HLH.

Table 2 presents the results of the analytical solution for the pattern LHL. Table 2 is similar to the results presented in Table 1.

Table 2: Values of D_t for combinations of balance (d) and AC factor (c) for LHL

	D				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.333	.333	.333	.333	.333
.25	.258	.289	.314	.335	.352
.50	.188	.254	.297	.316	.313
.75	.107	.222	.277	.269	.201
.90	.047	.202	.262	.225	.092
1.0	.000	.188	.25	.188	.000

The results in Table 2 are identical to those of Table 1 except for the fact that the pattern of decreases in the value of D_t are greater now when the imbalance is positive rather than negative (d > .5). Table 2 again provides evidence for the fact that the degree and direction (increasing versus decreasing) of the imbalance, as well as the degree of autocorrelation have an influence on the value of the summary of the indicator statistic, D_t. When c is equal to zero all conditions of d are equal to one another for the value of D_t. The value that D_t takes on (.33) is also what would be expected for an i.i.d. sequence when considering this pattern (LHL). As c increases to 1, the value that D_t takes on for a given level of c differs depending on the level of d and the direction of the imbalance. For the LHL pattern in general, D_t decreases in value as c goes from 0 to 1 and does so to a greater degree when imbalance is present. Additionally, the direction of the imbalance (positive, d > .5, or negative, d < 0) influences the rate of decrease for D_t. The complete imbalance conditions (d = 0 and 1) decrease at the greatest rate as c increases, but when

the pattern LHL is being considered the decrease is greater across levels of c for negative imbalance than it is for positive imbalance. The partial imbalance conditions (d =.25 and .75) decrease at a lesser rate than the complete imbalance conditions as c increases. However, the trend still holds that when the pattern LHL is being considered the decrease is greater across levels of c for negative imbalance than it is for positive imbalance. When the random component is balanced, the rate of decrease of D_t is the smallest and is identical to the HLH pattern presented in Table 1. There are two cases that do not follow the pattern as described here. When the imbalance is positive (d is equal to 1 or .75) and c is equal to .25, D_t increases in value compared to the other conditions. When c is equal to .25 and d is equal to 1 and .75, D_t is equal to .352 and .335, respectively. These two exceptions imply that as the analytical solution is stated, it would be expected to code more elements as ones according to the indicator statistic when there is as mild a degree of autocorrelation for a chain that is imbalanced in an increasing direction when considering the pattern LHL. Taken together, the findings presented in Tables 1 and 2 are informative in that it was not anticipated that the particular patterns satisfying the indicator statistic shared a relationship with the boundaries of the continuous uniform distribution.

Table 3 presents the solution for both patterns. It represents the expected value of D_t for the combinations of c and d summed for the HLH and LHL patterns.

Table 3: Values of D_t for combinations of balance (d) and AC factor (c) for BOTH

	D				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.667	.667	.667	.667	.667
.25	.610	.624	.628	.624	.610
.50	.500	.570	.594	.570	.500
.75	.308	.491	.554	.491	.308
.90	.139	.427	.524	.427	.139
1.0	.000	.375	.500	.375	.000

When both patterns are combined together, the relationship between the direction of imbalance and the pattern under consideration is no longer evident. When c is equal to 0, all levels of imbalance have D_t equal to .667. This value is what would be expected when characterizing three i.i.d. variables with the indicator statistic. The value of D_t decreases as c goes from 0 to 1, and this decrease is larger as the degree of imbalance increases. The complete imbalance conditions (d = 0 and 1) decrease at the greatest rate as c increases. The partial imbalance conditions (d =.25 and .75) decrease at a lesser rate than the complete imbalance conditions as c increases. When the random component is balanced, the rate of decrease of D_t is the smallest.

In summary, the analytical solutions for D_t produced somewhat unexpected results. It was not anticipated that the value of D_t would depend on the relationship between the direction of the imbalance and the particular pattern satisfying the indicator

statistic being equal to one. When d is such that the imbalance is negative, D_t shows a greater change across levels of c for the pattern LHL and a lesser change across levels of c for the pattern HLH. When d is such that the imbalance is positive, D_t shows a greater change across levels of c for the pattern HLH and a lesser change across levels of c for the pattern LHL. Also in respect to the relationship between the direction of imbalance and the particular pattern being satisfied, it was not expected that any combination of experimental conditions would produce a value of D_t greater than .33 for a single pattern satisfying the indicator statistic. When c is equal to .25, positive imbalance is associated with larger than expected values of D_t for the pattern LHL. When c is equal to .25, negative imbalance is associated with larger than expected values of D_t for the pattern HLH. These findings will be revisited in the Discussion.

The simulated solutions for the value of D_t as described in simulation study 1 will now be presented. The simulated solutions to the value of D_t were obtained as detailed in the Methods section when the first research question was presented. Table 4 presents the D_t values for all levels of d and c. The value in each cell of the table was obtained by averaging the final value of D_t for the 25 replications of each condition. The standard deviations are presented in parentheses.

Table 4: Mean D_t values (SD) for all levels of d and c

	d				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.668(.004)	.666(.004)	.667(.004)	.668(.003)	.667(.004)
.25	.627(.005)	.625(.003)	.625(.004)	.624(.004)	.627(.005)
.50	.584(.004)	.584(.005)	.585(.003)	.585(.004)	.584(.004)
.75	.541(.005)	.542(.005)	.541(.004)	.542(.005)	.542(.004)
.90	.520(.005)	.517(.006)	.516(.005)	.512(.005)	.517(.005)
1.0	.000(.000)	.374(.005)	.503(.005)	.375(.005)	.000(.000)

Table 4 shows that the simulated solutions differ somewhat from the analytical solutions. This disparity is indicative of a potential mistake in one of the solutions, and will be further explained in the discussion. Something that stands out in Table 4 is the consistency of the simulated mean values across levels of balance. For each level of c between 0 and .9, the mean values of D_t for each level of d are very similar to one another, and none are greater than one standard deviation away from any other. When c is equal to 0 this outcome is expected because the chains generated are i.i.d. samples from a stable distribution. However, as c increases from .25 to 1 it was expected that the degree of imbalance would influence the value of the summary of the indicator statistic. The anticipated effect of d is only evident when c is equal to 1. When c is equal to one, the mean values of D_t observed depend on the level of imbalance present and are clearly different than one another considering the standard deviations Thus, for the simulated

chains, the balance factor did not affect the chains as anticipated across all levels of c. The anticipated pattern of results, which is essentially that seen in Table 3 for the analytical solutions, is only partially seen in Table 4 for the simulated solutions. This result was unexpected and will receive more attention briefly when viewing plots of the chains.

Table 5 presents the D_t values for all levels of d and c for the CUSUM chains. The value in each cell of the table was obtained by averaging the final value of D_t for the 25 replications of each condition. The standard deviations are presented in parentheses. Table 5 is presented to demonstrate the difference between the new method and that proposed by Brooks (1998c).

Table 5: Mean D_t values (SD) for all levels of d and c for the CUSUM chains

	d				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.497(.005)	.497(.005)	.501(.006)	.498(.005)	.499(.005)
.25	.436(.005)	.435(.005)	.435(.005)	.434(.006)	.435(.005)
.50	.360(.006)	.360(.006)	.360(.004)	.361(.004)	.359(.006)
.75	.251(.006)	.248(.005)	.250(.005)	.249(.005)	.248(.004)
.90	.156(.004)	.156(.006)	.155(.006)	.155(.005)	.156(.005)
1.0	.0001(1.4e-20)	.0001(6.6e-05)	.008(.005)	.0001(5.7e-05)	.0001(1.1e-19)

The value of D_t when applied to the CUSUM chains is what would be expected according to Brooks' method (1998c). Similar to the findings in Table 4, the effect of c is

evident, but the effect of d is not. As c increases, the value of D_t decreases. Also, when c is equal to 1 the summary of the indicator statistic goes to zero for all levels of d. The value of D_t does not depend on the value of d.

When comparing the two sets of solutions for the current method, the simulated values are similar to the analytical solutions for some of the conditions, but there are some marked differences. When c is equal to zero or one, or if d is equal to .5, the simulated values match the analytical solutions very closely. When c is equal to 0, both sets of solutions indicate that the value of D_t tends towards .67 (considering both HLH and LHL) regardless of the levels of d. This finding was anticipated and is expected. When c is equal to one, the simulated mean values of D_t are similar to the analytical solutions, and the degree of imbalance is clearly related to the value of D_t. The agreement between the analytical and simulated solutions for this set of conditions is expected. When c is equal to one and d represents complete imbalance in the chain simulator (d= 0 or 1), then none of the elements in the chain were coded as 1 by the indicator statistic. This result is to be expected based on the way these chains are simulated. Each and every element is equal to the previous element plus a random component that is constrained to be in the range of -1 to 0 (d = 0) or in the range of 0 to 1 (d = 1). This means that chains simulated in this way cannot decrease (or increase) in value from element to element. Thus, every element in the chain is coded as zero by the indicator statistic. However, as c goes from .25 to .9 and there is imbalance present, the differences between the analytical and simulated solutions grow in disagreement, especially when completely imbalanced. The unanticipated discrepancy between the analytical and simulated solutions invited

greater scrutiny of the results. An explanation will be provided through investigation of the time-series plots to help guide understanding of the results. The time series plots are informative about the behavior of the elements across the length of the chain. There is not enough information presently to compare the values of D_t for the CUSUM chains.

To help guide the understanding of the difference between the analytical and simulated solutions, investigation of the path plots are informative. Extensive investigation of the path plots across all twenty five replications for a given condition led to the conclusion that the interesting trends could be elucidated without the provision of a plot for every replication of every condition Also, for the sake of economy, only three levels of d will be presented visually. These three levels are the balanced condition along with one each of the imbalanced and complete imbalanced condition. The reason for not showing the other two imbalance conditions is that they are the same as the ones presented, except the imbalance is in the opposite direction. As there is no information concerning the particular pattern satisfying the indicator statistic being equal to 1, presenting both directions of imbalance would be redundant. As such, there is nothing to be gained by presenting both directions of imbalance. The findings of interest can clearly be seen in the conditions presented.

First, summary tables will provide descriptive statistics about the chains. For the descriptive statistics there will be three tables. Each table will represent one level of d, and it will contain the averages of the mean chain values (with standard errors), as well as the mean chain minimums, maximums, and ranges (with standard deviations) across the twenty-five replications. After the descriptive statistics, the autocorrelation plots will be

presented. These plots are informative about manipulations of the simulated chains to achieve characteristics similar to real MCMC samplers. The autocorrelation plots will be followed by the path plots. The path plots will provide information concerning behavior of the chains over time. For both the autocorrelation and path plots, the three levels of balance will be plotted together in one figure. Each figure produced will correspond to one level of c. The result will be 6 figures, with 3 graphs in each for both the path plots and plots of the autocorrelations. These descriptive statistics and graphs will help provide an explanation for the pattern of results seen in the analytical and simulated solutions for the value of D_t.

Table 6 is presented below. It contains the descriptive statistics for the chains for all levels of c when d is equal to one (complete imbalance).

Table 6: Descriptive statistics for simulated chains when d=1 (Complete Imbalance)

c	**Average Mean**	**Average Minimum**	**Average Maximum**	**Average Range**
0.0	.501(.002)	.0001(.0001)	.999(.0001)	.999(.0001)
.25	.667(.004)	.028(.009)	1.30(.009)	1.28(.012)
.50	.999(.006)	.128(.032)	1.87(.025)	1.74(.039)
.75	2.00(.013)	.430(.176)	3.37(.072)	2.94(.205)
.90	5.00(.022)	.489(.269)	7.38(.251)	6.89(.321)
1.0	2497(13.9)	.451(.278)	4997(25.1)	4996(25.1)

Table 6 shows that for the condition of complete imbalance (positive, in this case) as c increases then each descriptive statistic increases in both the mean and variability. The trend is true of the average mean, average minimum, the average maximum, and the average range. When c is equal to 0, the chains are i.i.d. sequences that fall within the boundaries defined by d. As c increases, the values of the chain elements tend to increase and become more variable. Thus, minimums, maximums, and ranges are affected. These findings are an indication that these chains have at least some of the desired characteristics that they were intended to have.

Table 7 contains the descriptive statistics for the case when d is equal to .75. This level of d represents partial imbalance in a positive direction.

Table 7: Descriptive statistics for simulated chains when d=.75 (Partial Imbalance)

c	**Average Mean**	**Average Minimum**	**Average Maximum**	**Average Range**
0.0	.251(.003)	-.250(.0001)	.750(.0001)	.999(.0002)
.25	.334(.003)	-.304(.007)	.971(.008)	1.28(.010)
.50	.498(.005)	-.374(.018)	1.38(.022)	1.75(.031)
.75	1.00(.011)	-.386(.065)	2.39(.070)	2.78(.096)
.90	2.49(.026)	.021(.198)	4.76(.182)	4.74(.232)
1.0	1257(14.0)	.139(.252)	2505(27.8)	2505(27.8)

For Table 7, the pattern is slightly different than seen previously in Table 6. It is still the case that as c increases the average mean, maximum, and range increases. It is

also still the case that the variability of these statistics increases as c increases. However, when d represents partial imbalance (positive) the average minimum first decreases and then increases over the range of c. These findings are an indication that these chains have at least some of the desired characteristics that they were intended to have.

Table 8 contains the descriptive statistics for the case when d is equal to .5. It is presented below.

Table 8: Descriptive statistics for simulated chains when d=.5 (Balance)

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.000(.002)	-.500(.0001)	.500(.0001)	.999(.0001)
.25	.000(.005)	-.641(.007)	.637(.008)	1.28(.011)
.50	.003(.007)	-.874(.024)	.871(.019)	1.75(.027)
.75	-.004(.013)	-1.38(.063)	1.40(.048)	2.78(.079)
.90	.001(.030)	-2.33(.224)	2.29(.171)	4.62(.302)
1.0	-3.79(17.8)	-27.3(19.5)	20.6(14.5)	47.9(16.1)

As can be seen in Table 8, as c increases the average minimum gets smaller and more variable, the average maximum gets larger and more variable, and the average range gets larger and more variable. The average mean has no discernible pattern, but its variability increases as c increases, and it generally tends to stay near zero, which is to be expected. These findings are an indication that these chains have at least some of the desired characteristics that they were intended to have.

In summary, the descriptive statistics for the chains indicate that the chains have the characteristics expected based on the combinations of factors associated with each condition. However, they do not provide insight into the discrepancy between the analytical and simulated solutions. The descriptive statistics in Tables 6, 7, and 8 help guide investigation of the path plots to be presented shortly.

The autocorrelation plots are presented next in Figures 3 through 9. Again, a plot for each of the three levels of d representing complete imbalance, partial imbalance, and balance will be presented in each figure, and there will be a separate figure for each of the six levels of c. Each plot will be presented and then followed by an immediate description. Investigation of the autocorrelation plots is informative when assessing the chains produced in MCMC, so these plots give the reader a sense of how the chains behave in this regard. As a reminder, each of these autocorrelation plots represents a single chain. There is such a high degree of similarity among all of the 25 replications of each condition that this economy is deemed acceptable. Again, the plots are clear representations of the trends seen in all 25 replications of each unique factorial combination.

Figure 3 contains the autocorrelation plots for the three levels of d when c is equal to 0. Figure 3 is presented immediately below.

Figure 3: Autocorrelation plot for all levels of d when c = 0

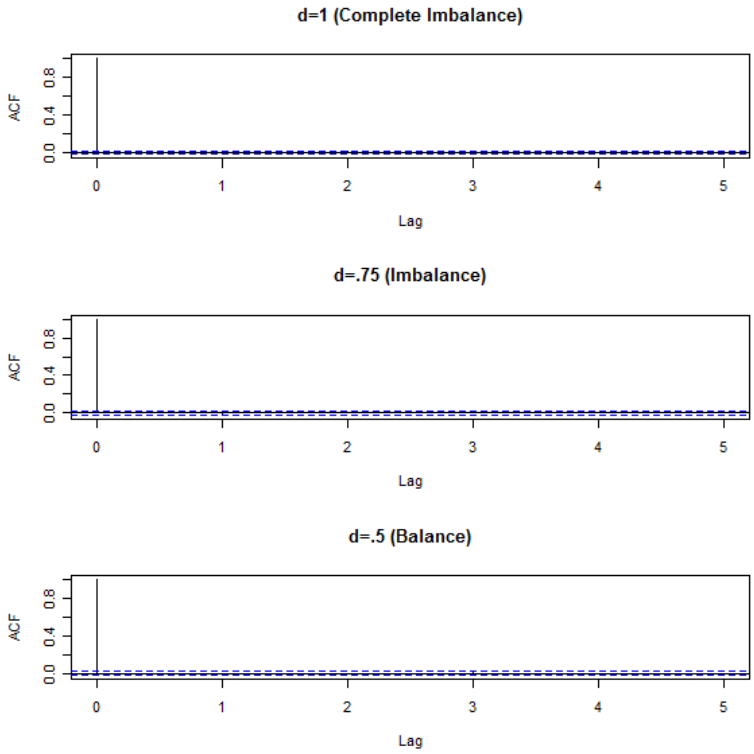


Figure 3 shows that there is no autocorrelation present in the chains when c is equal to zero. This result is expected because the chains for all conditions where c is equal to 0 are i.i.d. sequences by definition. Figure 4 is presented next.

Figure 4: Autocorrelation plot for all levels of d when c =.25

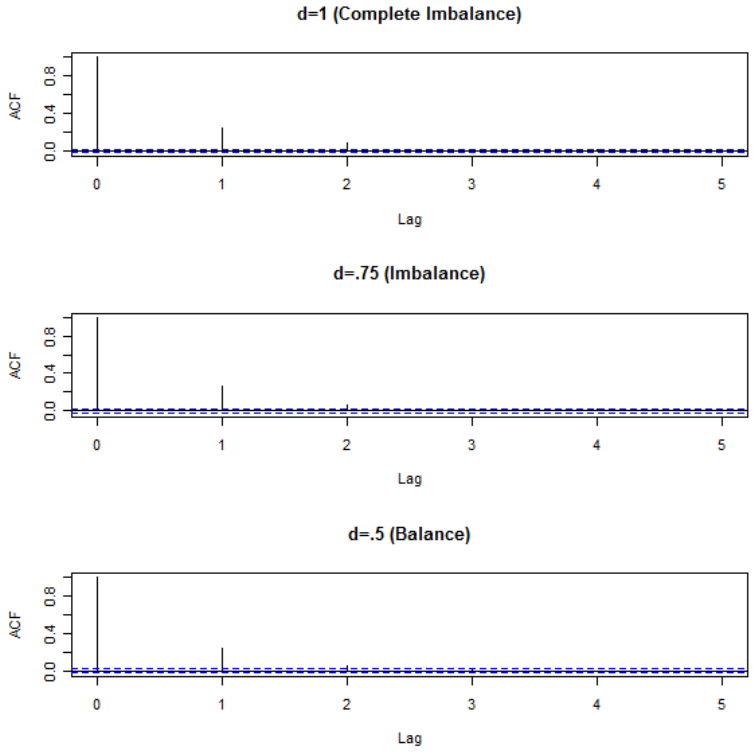


Figure 4 shows that when c is equal to .25, there is a very small amount of autocorrelation present in the chains. The autocorrelations tend to zero within a lag of two or three elements. The mean autocorrelations are roughly .25 and .06 at lags of 1 and 2, respectively. The presence of an association among elements indicates that the factor c is having its desired effect.

Figure 5: Autocorrelation plot for all levels of d when c = .5

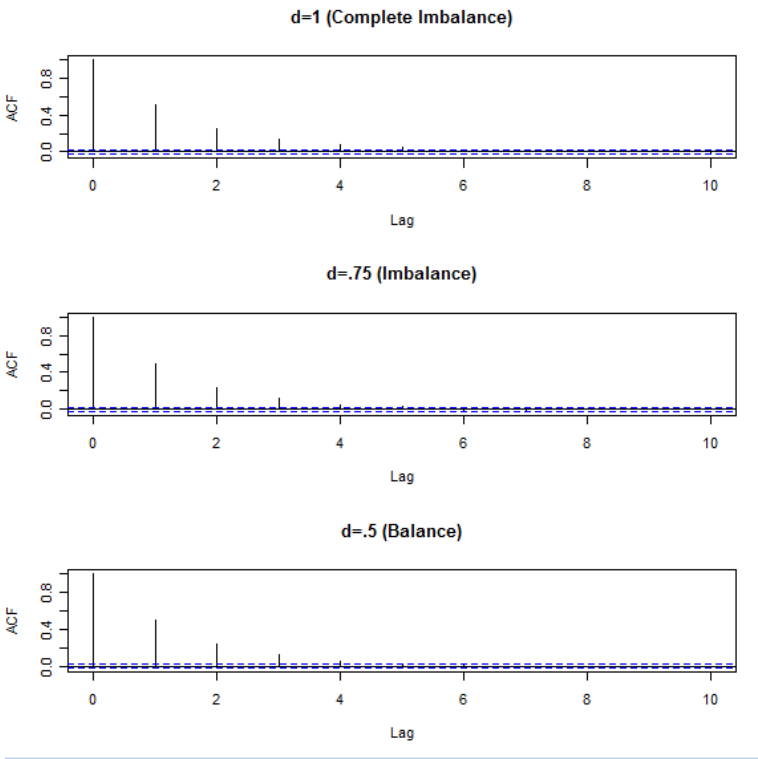
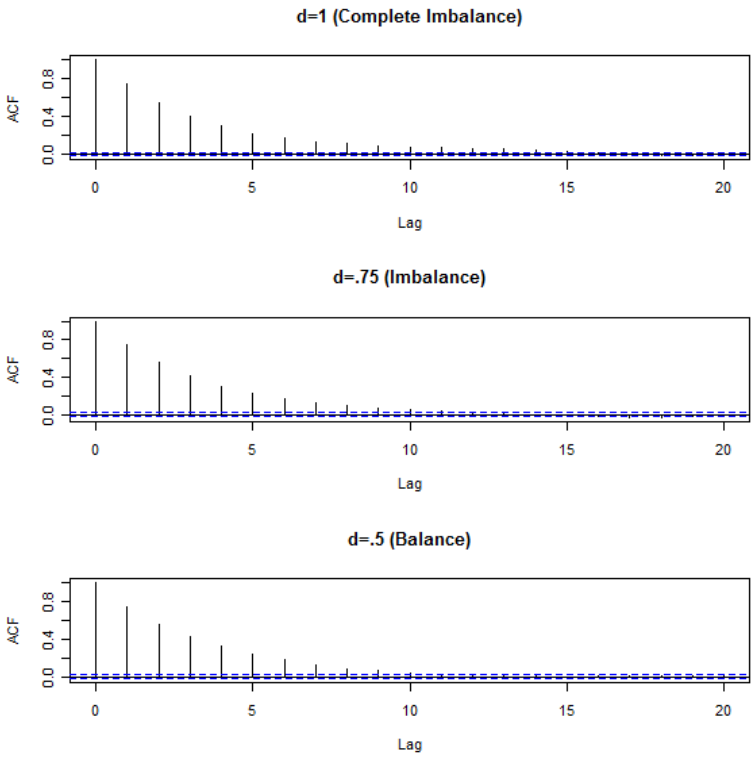


Figure 5 shows that when c is equal to .5, there is again a small amount of autocorrelation present in the chains, slightly larger than in the previously presented condition. The autocorrelations tend to zero within a lag of four to five elements. The autocorrelations at lag 1, 2, 3, and 4 are roughly .51, .26, .12, and .06, respectively. The presence of an association among elements indicates that the factor c is having its desired effect.

Figure 6: Autocorrelation plot for all levels of d when c = .75



In Figure 6, the autocorrelations are slightly larger again. There is evidence that the level of imbalance is influencing the amount of autocorrelation present in the chains. The autocorrelations tend towards zero by a lag of 15 for complete imbalance and by a lag of 10 for the other two levels of balance. The autocorrelations for a lag of 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 are roughly .76, .58, .44, .34, .26, .20, .15, .10, .08, and .05, respectively. The presence of an association among elements indicates that the factor c is having its desired effect.

Figure 7 is presented immediately below. In Figure 7, the autocorrelations are larger and extend over a longer lag. At lag 1, the correlation is roughly .9 and decreases at

a rate of roughly .05 at each successive increase in lag until the autocorrelation reaches 0. The autocorrelations tend to zero at a lag of roughly 35 to 40 and decrease very slowly.

Figure 7: Autocorrelation plot for all levels of d when c = .9

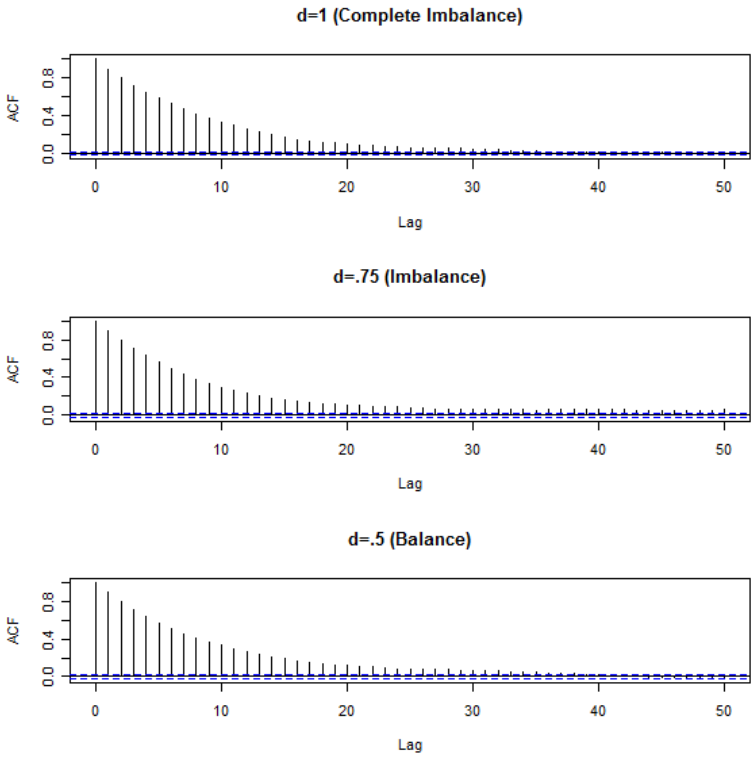
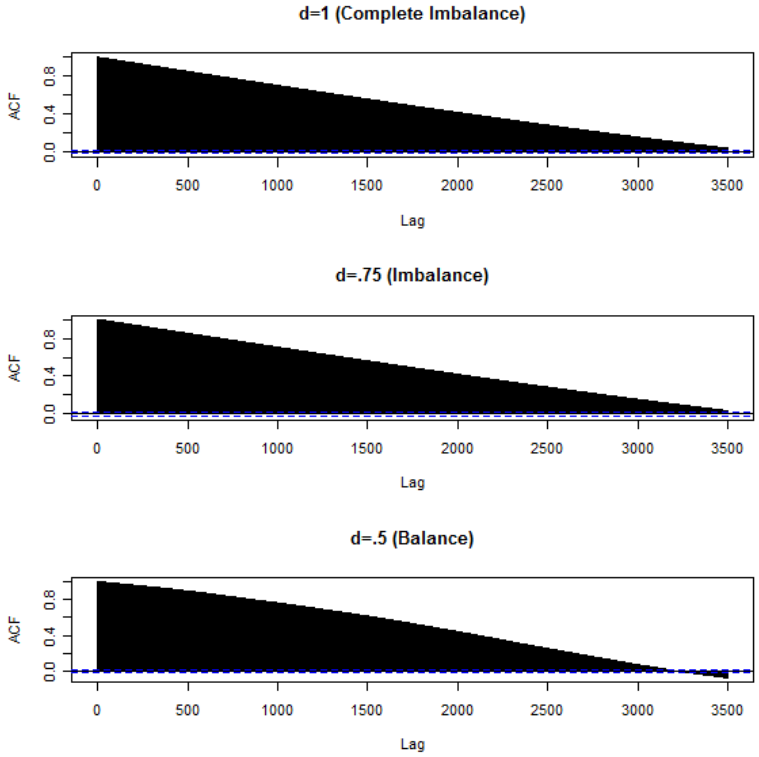


Figure 8 shows the autocorrelation plots for the case where c is equal to 1. It is presented immediately below. There is a great deal of autocorrelation present in these chains. The autocorrelations are large and extend to a lag of roughly 3000 to 3500.

Figure 8: Autocorrelation plot for all levels of d when c = 1.0



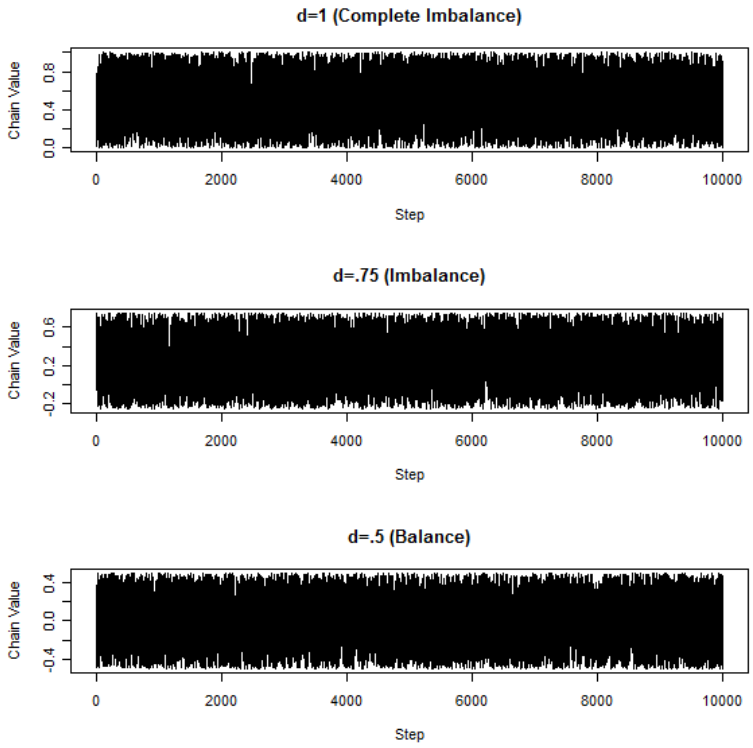
In summary, the degree of autocorrelation and the lag over which it persists is clearly dependent on the level of c. Similar to what was expected, as c increases so does the amount of autocorrelation present in the chains. Also, the autocorrelation plots do not differ over the lower levels of d, but begin to show some effect of as c increases and is clear when c is equal to 1. When c is equal to 1, the lag at which the autocorrelation goes to zero is slightly lower when there is balance present than when imbalance is present. However, for the purpose of MCMC estimation, the amount of autocorrelation present in the chains when c is equal to 1 is indicative of a problematic chain regardless of the balance. Also, the values of the autocorrelations presented in the previous six figures are

representative of all conditions in all studies similar to these throughout the rest of this document. The focus of the research questions is on the value of D_t, not the particular value of the autocorrelations. For the purposes of this research, it is sufficient to show that as autocorrelation increases the expected value of the summary of the indicator statistic decreases. As such, any further reference to the values of the autocorrelation plots will reference the descriptions of Figures 3-8 and Tables 9 through 14.

The path plots will be presented in Figures 9-14. Similar to the figures used to present the autocorrelation plots, each figure will contain the path plots for the three levels of d, and there will be an individual figure for each level of c. These plots provide information that helps explain the pattern of results seen in the simulated solutions presented in Table 4.

Figure 9 contains path plots of the chains for the conditions of complete imbalance (top), imbalance (middle), and balance (bottom) when c is equal to zero.

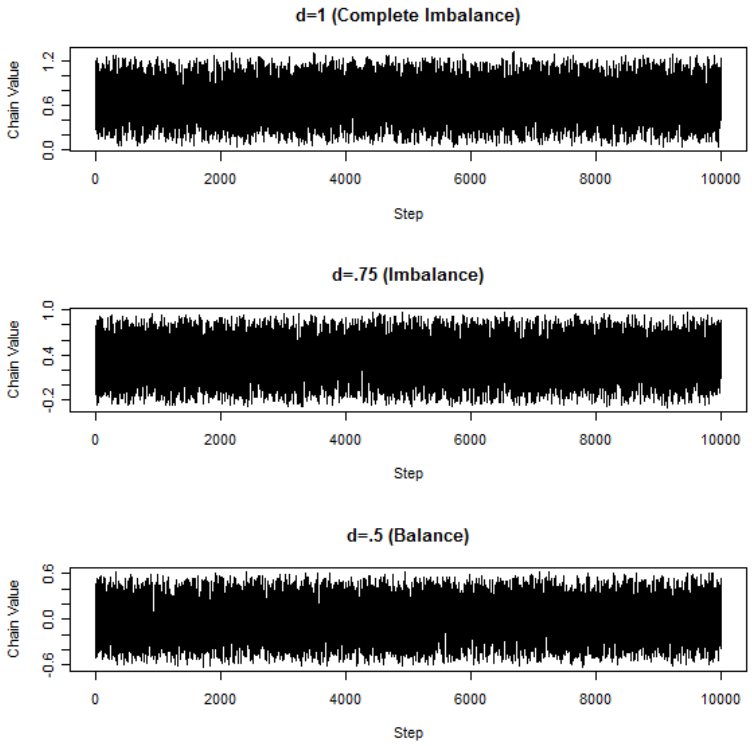
Figure 9: Path plots for all levels of balance when c = 0



When c is equal to zero; the only discernible difference between these three path plots is the bounds within which they fall. Because the previous element has no influence on the successive element, these chains represent random draws from a stable distribution with the respective bounds. When there is complete imbalance (in this case, positive), all values fall between 0 and 1. When there is partial imbalance, all of the values fall within the lower and upper bounds of -.25 and .75, respectively. When there is balance, the plotted chain values always remain between the lower and upper bounds of -.5 and .5, respectively. These results are expected for this particular set of conditions. These path plots indicate that these chains have the appearance of i.i.d. sequences.

Figure 10 contains the path plots for the three levels of d when c is equal to .25. It is presented below.

Figure 10: Path plots for all levels of balance when c = .25



As can be seen in Figure 10, when c is equal to .25, the path plots still look very similar to one another and to the case where c is equal to zero. However, the influence of c can be seen in that the y axis has expanded due to a greater range present in the chain values. The range in the chain values is slightly larger than the range between the upper and lower bounds for each of these three conditions. When there is complete imbalance present, the bounds now range from the lower bound of roughly 0 up to 1.2. When partial

imbalance is present, the values in the path plot range from roughly -.2 to 1. When there is balance present, the values in the chain range from roughly -.6 to .6. In all cases, the range of values is similar, but the location is shifted (in these cases positively) depending on the degree of imbalance present.

Figure 11: Path plots for all levels of balance when c = .5

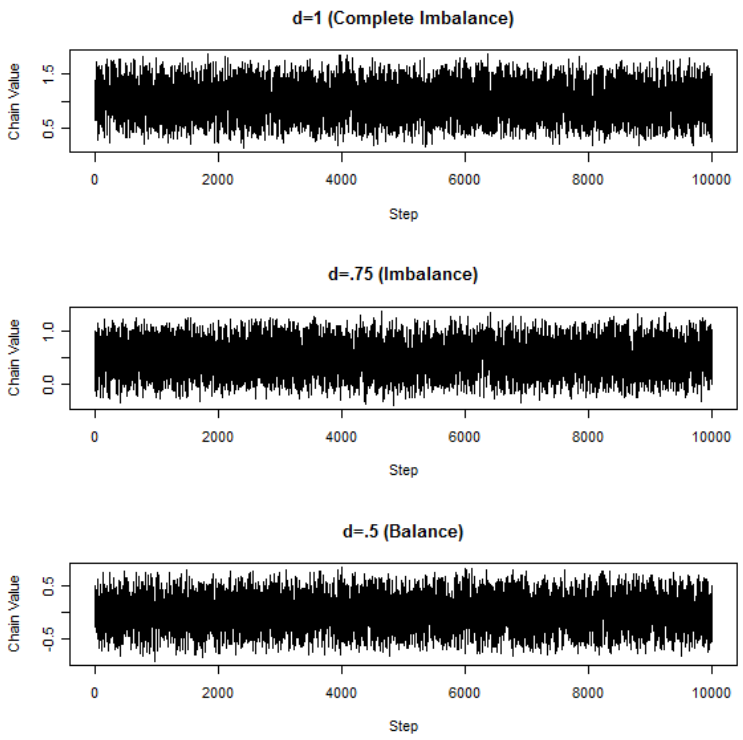
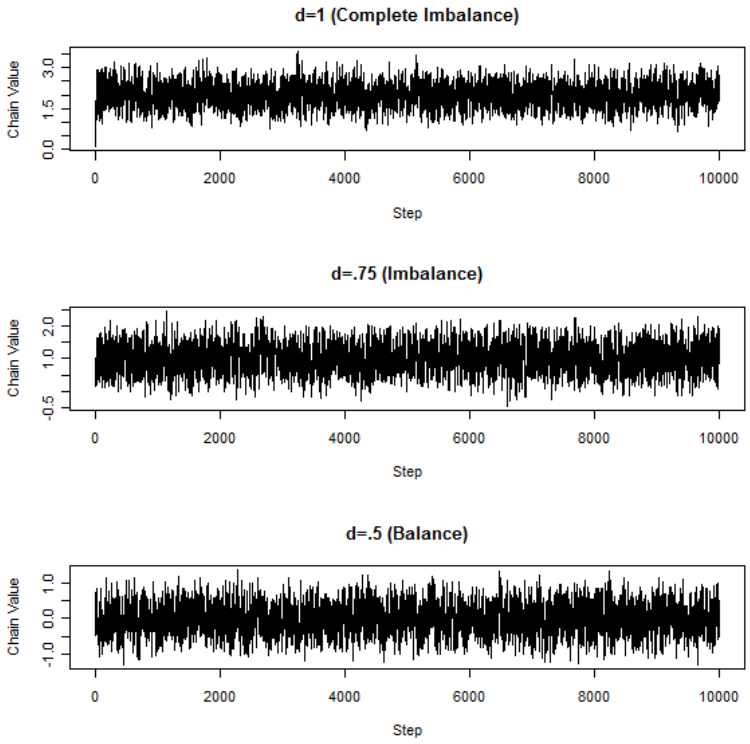


Figure 11 follows the same trend as the previous figure, where the range of chain values is slightly expanded. The range of chain values for the complete imbalance condition ranges from roughly 0 to 2. The range of chain values for the partial imbalance condition ranges from roughly -.5 to 1.5. The range of chain values for the balanced

condition ranges from roughly -1 to 1.

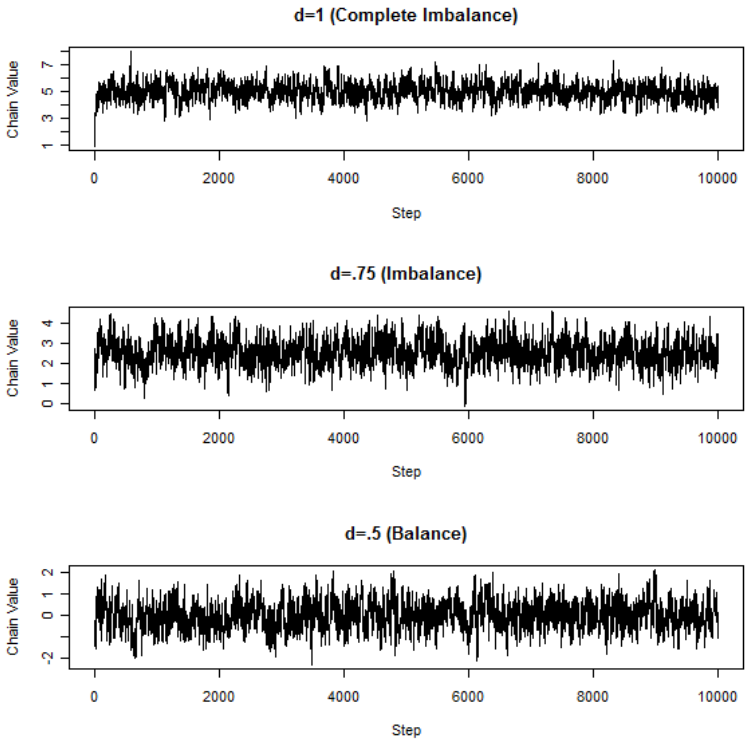
Figure 12 presents the path plots for the case where, c is equal to .75. It is presented below.

Figure 12: Path plots for all levels of balance when c = .75



The pattern of results in Figure 12 shows the same general trend as the previous figures (Figure 10 and Figure 11, specifically). The ranges of chain values are expanded for all conditions. The range of chain values for the complete imbalance condition ranges from roughly 0 to 3.5. The range of chain values for the partial imbalance condition ranges from roughly -.5 to 2.5. The range of chain values for the balanced condition ranges from roughly -1.5 to 1.5.

Figure 13: Path plots for all levels of balance when c = .9



In Figure 13, the path plot indicates a slightly greater degree of the pattern that has become more and more evident over the figures representing the last few conditions (i.e., d = .25, .5, and .75). The range of chain values is expanded, and the partial and complete imbalance conditions show that the chain has moved in the direction desired.

Figure 14: Path plots for all levels of balance when c = 1.0

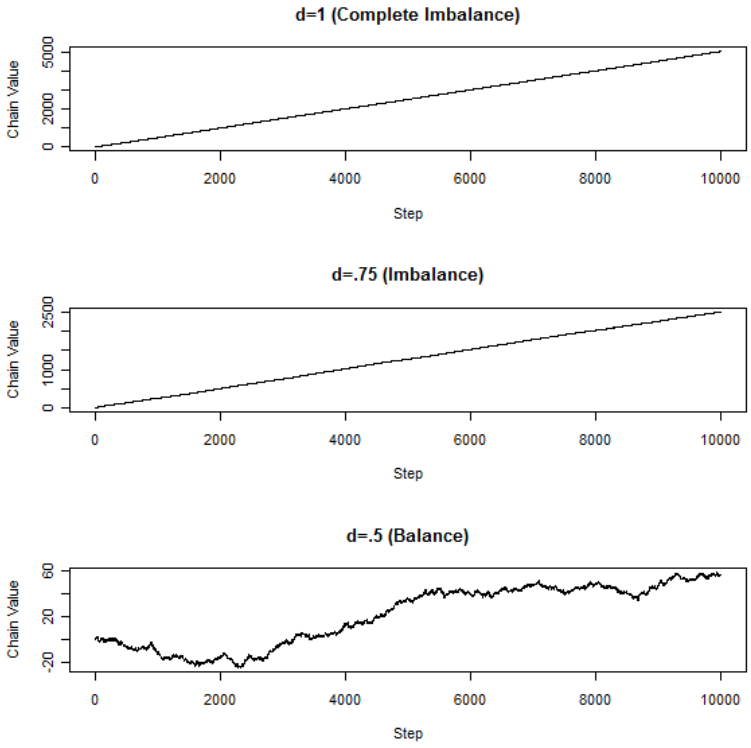


Figure 14 exaggerates the pattern that has become more evident as c moves from .25 to .9. When there is complete imbalance, the chain moves in the direction imposed by d. It is not possible for any element to be less than the previous element (because the imbalance is positive in this case). When there is partial imbalance, the chain moves reliably in one direction, although there are a handful of cases where the ordering of elements in the chain satisfies the indicator statistic being equal to one. This pattern was seen in Table 4. When there is balance present, the each new element in the chain is equally likely to be larger or smaller than the previous element. When this is the case, the chains take on the appearance of a Brownian walk, and randomly move up and down. In

Table 4, this pattern is associated with D_t being equal to .5.

A brief explanation at this point is warranted. The simulated solutions produced values of D_t that were equal across all level of balance. This finding was unanticipated at the outset of this study. It was thought that the simulated solutions should match the analytical solutions. Closer investigation of the chains by way of the path plots provides some insight into why the simulated D_t values were equal across levels of balance. The path plots indicate that the chains seem to settle to a relatively stable range for the vast majority of the conditions in this study. The only exception is when c is equal to 1. When c is equal to 1, the chains continue to move in the direction anticipated over the course of all the elements. For the cases where c is not equal to 1, the chains look more stable than was anticipated. The similarity in appearance of the chains produced across balance conditions can also be seen in the pattern of the simulated solutions. When there is balance present, it was expected that the chains would appear to remain stable unless there was a very large amount of autocorrelation present (e.g., as is the case when c = 1). When there is imbalance present, the chains still appear very stable, only exhibiting a slight shift in the anticipated direction. This stability appears to be a result of a cancellation of the factors c and d. It seems that for many levels of c (i.e., c = .25, .5, .75, and .9), the influence of c may counteract the influence of d for the imbalanced levels. For example, when there is complete imbalance (positive) in the chains and c is equal to .5, then even though every new random component generated is constrained to be positive, it will be added to half of the value of the previous element. This still allows for a chance to observe patterns that satisfy the indicator statistic. The imbalance present in

the random component will tend to move the chain in its given direction, but it can only move the value so far before the reduction in value due to the factor c moves the following element further than the range of the random component being added. Thus, the chain can only wander so far in one direction before it has to move in the other direction. Thus, these chains are essentially converged. The first simulation study has provided evidence that the method of simulating chains proposed in this paper has characteristics different than originally anticipated. The findings of the analytical study and simulation study 1 will be redressed in the Discussion.

Findings for research question 2

The second research question attempts to determine the influence of thinning on the behavior of the indicator statistic. Simulation study 2 attempts to answer the second research question by generating chains in a fashion similar to the first simulation study, but the chains are thinned to achieve approximately i.i.d. sequences before applying the indicator statistic. Simulation study 2 had some conditions identical to those presented in the first simulation study. These results are omitted because they are virtually identical to findings already presented. Everything true of the results already presented for those conditions is true for those being omitted. The emphasis will be placed on the results where thinning was performed on the chains. The format is similar to that of the previous studies. First, the simulated values of D_t will be presented to show the effect of thinning. The results will be briefly described and an explanation for the trends will be provided. Following the values of D_t, descriptive statistics will be presented and graphs of the autocorrelations and the path plots will be presented. Each table and figure will be

described and a brief summary will be provided.

Table 9 contains the D_t values for the thinned chains. Thinning for the chains was done by taking every n^{th} element, where n was determined by inspection of the autocorrelation plots. For each condition, a value of n was chosen such that thinning all 25 chains produced for that condition by that n produced a thinned chain that was approximately i.i.d. to be characterized by the indicator statistic. The only exception was the case where c is equal to one. The autocorrelations were so strong over such a long lag (roughly 3000 to 3500) that there would only be 3 or 4 values left in the thinned chain. It was decided to use n= 250 in this case to provide enough values in the chain to estimate D_t reasonably well. The values chosen for thinning are 1, 4, 6, 10, 30, and 250 for c being equal to 0, .25, .5, .75, .9, and 1, respectively. Again, the reason that thinning was done in this fashion was because of the uniformity of results across all replications within a unique combination of factor levels. After thinning, each and every chain was inspected to determine if the thinning worked to produce i.i.d. sequences uniformly across repetitions. The thinned chains were all deemed to be linearly independent sequences for the sake of this research.

Table 9: D_t values for thinned chains

	D				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.668(.004)	.666(.004)	.668(.004)	.667(.004)	.666(.004)
.25	.663(.008)	.664(.009)	.665(.010)	.664(.007)	.666(.008)
.50	.660(.011)	.659(.011)	.662(.011)	.658(.010)	.661(.012)
.75	.652(.005)	.656(.013)	.659(.014)	.657(.015)	.651(.013)
.90	.654(.019)	.656(.020)	.653(.017)	.657(.021)	.655(.022)
1.0	.000(.000)	.000(.000)	.467(.057)	.000(.000)	.000(.000)

As can be clearly seen in Table 8, when the effect of thinning serves to provide an at least approximately i.i.d. sequence of values, then the indicator statistic tends towards the expected value of .67. These findings must be interpreted in light of the fact that the method for simulating chains essentially produced mostly converged chains. Based on the interval about D_t that was originally presented by Brooks (1998), the ranges of D_t values that would be considered 'converged' are .658 to .676, .649 to .685, .644 to .690, .638 to .696, .616 to .718, and .519 to .815 for the chain lengths associated with the thinning used for values of c equal to 0, .25, .5, .75, .9 and 1, respectively. The ranges are different for the levels of c due to the differing thinning values used for the conditions. According to these intervals, all of these chains would be considered to be 'converged' except for the case where c is equal to one. When c is equal to 1 and there is balance present, 5 out of the 25 chains produced values of D_t within the bounds specified previously. The fact that

most of the chains for this condition do not produce values of D_t within the ranges provided above is due to the fact that the chains after thinning did not produce completely i.i.d. sequences (which will be shown below). The strong degree of autocorrelation left in the chains for the conditions where c is equal to one shows the effect of reducing the value of D_t. Another finding of interest is that when c is equal to one and there is partial imbalance present, the value of D_t is .000, rather than .375 when there is not thinning. Thinning the chains by taking every n^{th} element seems to eliminate the cases that satisfy the indicator statistics being equal to one. This finding was not anticipated.

The descriptive statistics for the thinned chains are presented in the three following tables. Again, descriptive statistics for the chains before thinning are virtually identical to those presented in Tables 6, 7, and 8 so they are omitted. Tables 10, 11, and 12 will present the descriptive statistics for all levels of c for the cases where d is equal to 1 (complete imbalance), .75 (partial imbalance), .5 (balance), respectively. Each table will be presented and immediately followed by a brief description of the data presented therein.

Table 10 is presented below. It contains the descriptive statistics for all levels of c when d is equal to one (complete imbalance) for the thinned chains.

Table 10: Descriptive statistics for thinned chains when d=1 (Complete Imbalance)

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.499(.004)	.0001(.0001)	.999(.0001)	1.00(.0001)
.25	.664(.007)	.040(.012)	1.29(.012)	1.25(.019)
.50	.998(.008)	.157(.030)	1.83(.027)	1.68(.039)
.75	2.00(.015)	.763(.102)	3.23(.093)	2.47(.149)
.90	4.99(.034)	3.14(.245)	6.83(.211)	3.70(.375)
1.0	2557(17.0)	248.5(6.65)	4865(27.0)	4616(24.7)

It is important to mention that when c is equal to zero, there is no thinning of the chains because they are already i.i.d. sequences. The results for these conditions are very similar to those in Table 6. Table 10 is similar to Table 5 in that it shows that when c increases then each descriptive statistic increases in both the mean and variability. Again, these findings are an indication that these chains have at least some of the desired characteristics that they were intended to have. For example, the chains were intended to increase in the mean, and they do. Again it can be seen that as c increases, the range gets larger. Compared to the chains that are not thinned, these chains tend to produce descriptive statistics that are more variable. For the cases where c is greater than zero, this increased variability is due to the fact that the descriptive statistics are based on chains with fewer observations.

Table 11 contains the descriptive statistics for the case when d is equal to .75, representing partial imbalance, for the thinned chains.

Table 11: Descriptive statistics for thinned chains when d=.75 (Partial Imbalance)

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.250(.003)	-.250(.0001)	.750(.0001)	1.00(.0001)
.25	.334(.006)	-.294(.012)	.960(.012)	1.25(.017)
.50	.498(.008)	-.340(.025)	1.33(.028)	1.67(.042)
.75	1.00(.016)	-.224(.065)	2.29(.091)	2.51(.140)
.90	2.50(.041)	.602(.096)	4.41(.247)	3.80(.304)
1.0	1281(20.9)	123.6(7.89)	2439(32.4)	2315(30.6)

For Table 11, the pattern is slightly different than for the previous table. It is still the case that as c increases the average mean, maximum, and range increases. It is also still the case that variability of all statistics increases as c increases. However, when d represents partial imbalance (positive) the average minimum first decreases and then increases over the range of c. Compared to Table 7, the descriptive for all cases where c is greater than zero produce statistics that are more variable. Again, this increased variability for the statistics associated with the thinned chains is likely due to the decreased number of observations on which the statistics are based.

Table 12 contains the descriptive statistics for the case when d is equal to .5 for the thinned chains. It is presented below.

Table 12: Descriptive statistics for thinned chains when d=.5 (Balance)

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.000(.002)	-.500(.0002)	.500(.0001)	1.00(.0002)
.25	-.002(.007)	-.624(.011)	.629(.011)	1.25(.016)
.50	.003(.009)	-.831(.040)	.836(.034)	1.67(.055)
.75	.003(.017)	-1.24(.095)	1.25(.105)	2.49(.148)
.90	-.003(.039)	-1.94(.264)	1.89(.293)	3.83(.381)
1.0	4.49(14.3)	-13.7(18.5)	24.1(16.8)	37.8(16.3)

As can be seen in Table 12, as c increases the average minimum gets smaller and more variable, the average maximum gets larger and more variable, and the average range gets larger and more variable. The average mean has no discernible pattern, but it stays close to zero and its variability increases as c increases. When compared to Table 8, these estimates tend to be more variable. Again, the fact that the increased variability observed as c increases compared to the case where there is no thinning is due to the fact that the estimates are based on fewer observations.

In summary, the descriptive statistics for the thinned chains are very similar to those for the chains that are not thinned. The primary difference is that the thinning of the chains means that the descriptive statistics are based on fewer observations for all cases where c is greater than 0. Overall, the thinned chains look similar to the full chains in terms of the descriptive statistics, even though the value of the summary statistic is

clearly influenced by the thinning, as was seen in Table 8.

The autocorrelation plots for these chains before thinning are virtually the same as the plots presented in Figures 3 through 8. As such, these autocorrelation plots will not be presented again. The autocorrelation plots for the thinned chains will be provided in Figures 15 through 20. For all conditions except c = 1, the autocorrelation plots on the thinned chains reveal that the thinning had the desired effect of producing linearly i.i.d. sequences. Again, the autocorrelation plots for the thinned chains will be presented such that the plots for completely imbalanced, partially imbalanced, and balanced chains will be presented together in one figure for each of the six levels of c.

Figure 15 presents the autocorrelation plots for all levels of balance for the case where c is equal to 0.

Figure 15: Autocorrelation plot for all levels of d when c = 0

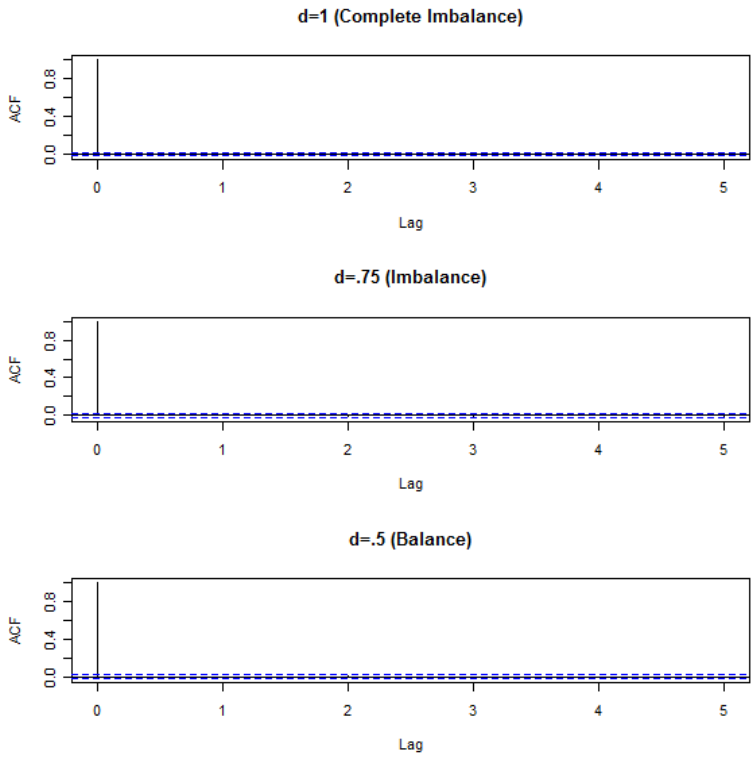


Figure 15 shows that the chains produced for this set of conditions are i.i.d. sequences, and the interpretation is identical to Figure 8. Because these are i.i.d. sequences, no thinning is necessary.

Figure 16: Autocorrelation plot for all levels of d when c = .25

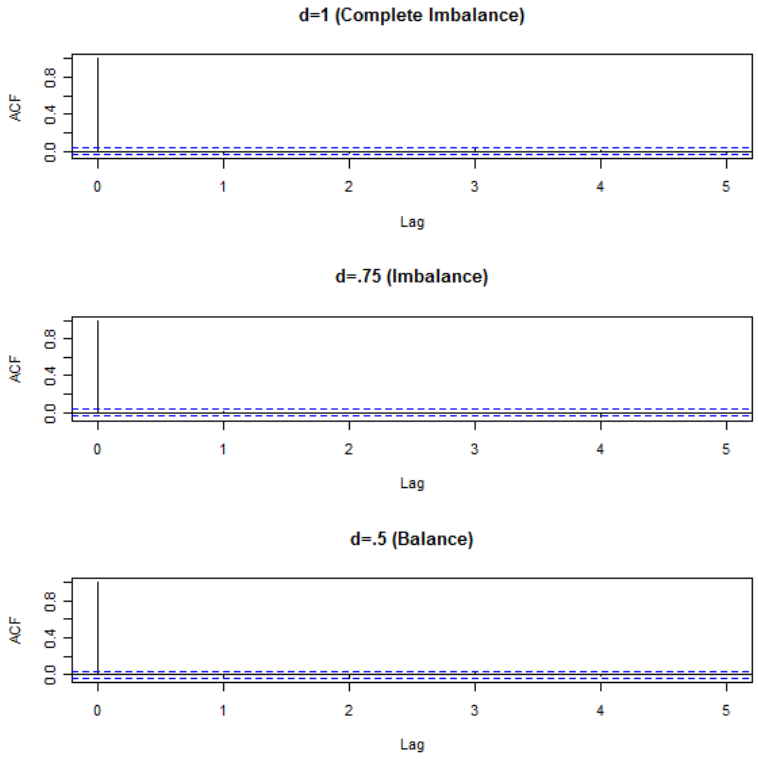


Figure 16 shows the chains produced for all levels of d when c is equal to .25, after thinning. This set of plots looks like those for the previous figure. All chains in this set of conditions are at least linearly i.i.d.

Figure 17: Autocorrelation plot for all levels of d when c = .5

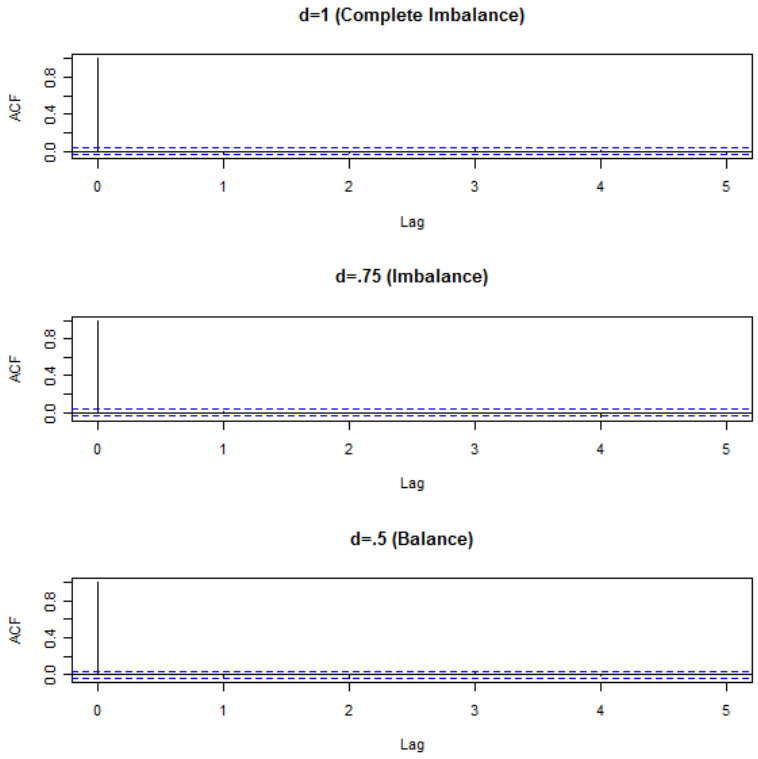


Figure 17 shows the chains produced for all levels of d when c is equal to .5, after thinning. This set of plots looks like those for the previous figure. All chains in this set of conditions are at least linearly i.i.d.

Figure 18 presents the autocorrelation plots for all levels of d when c is equal to .75. It is presented below.

Figure 18: Autocorrelation plot for all levels of d when c = .75

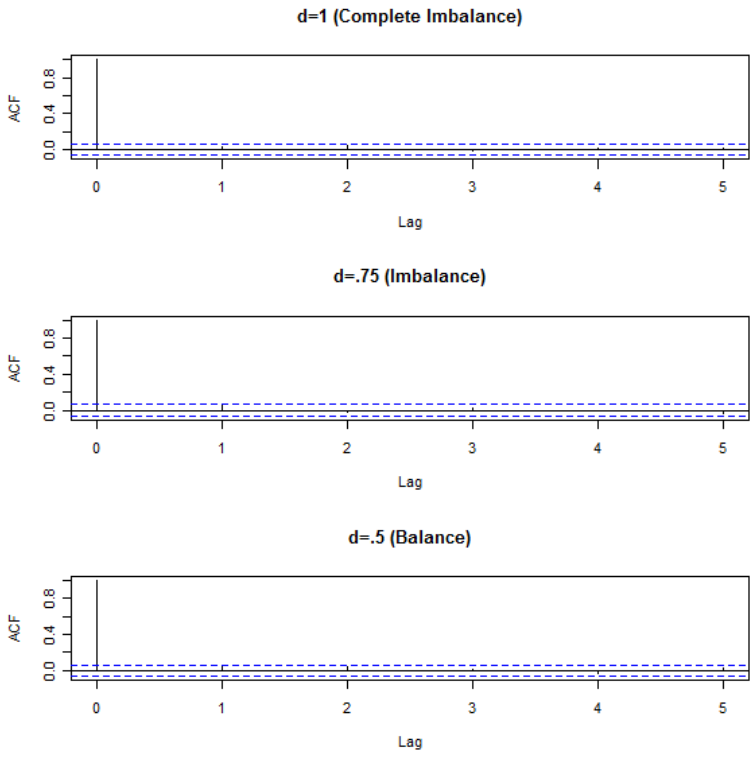


Figure 18 shows the chains produced for all levels of d when c is equal to .5, after thinning. This set of plots looks like those for the previous figure. All chains in this set of conditions are at least linearly i.i.d.

Figure 19 presents the autocorrelation plots for all levels of d when c is equal to .90. It is presented below.

Figure 19: Autocorrelation plot for all levels of d when c = .9

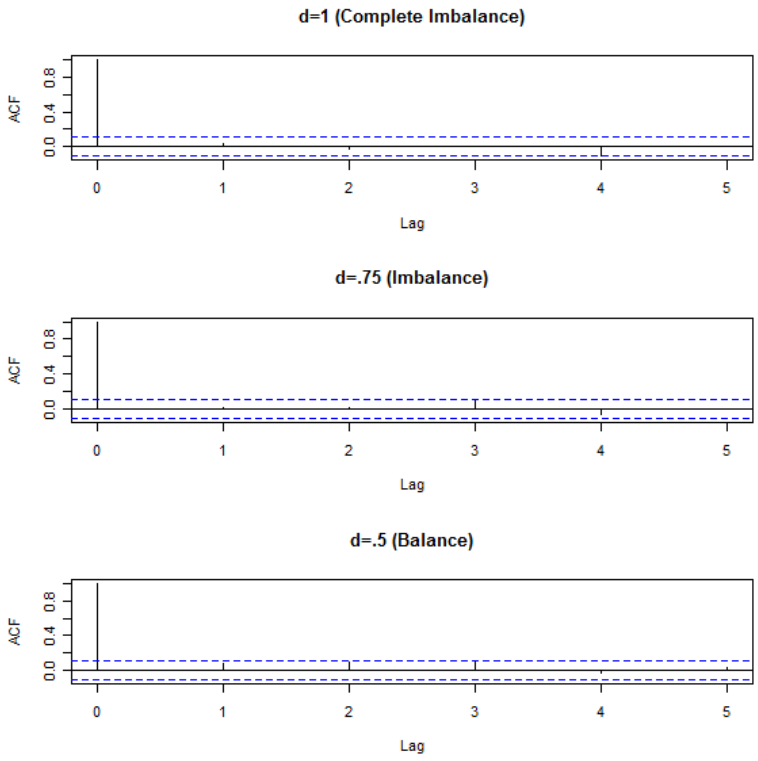


Figure 19 shows the chains produced for all levels of d when c is equal to .9, after thinning. This set of plots looks like those for the previous figure.All chains in this set of conditions are at least linearly i.i.d.

Figure 20 presents the autocorrelation plots for all levels of d when c is equal to 1. It is presented below.

Figure 20: Autocorrelation plot for all levels of d when c = 1.0

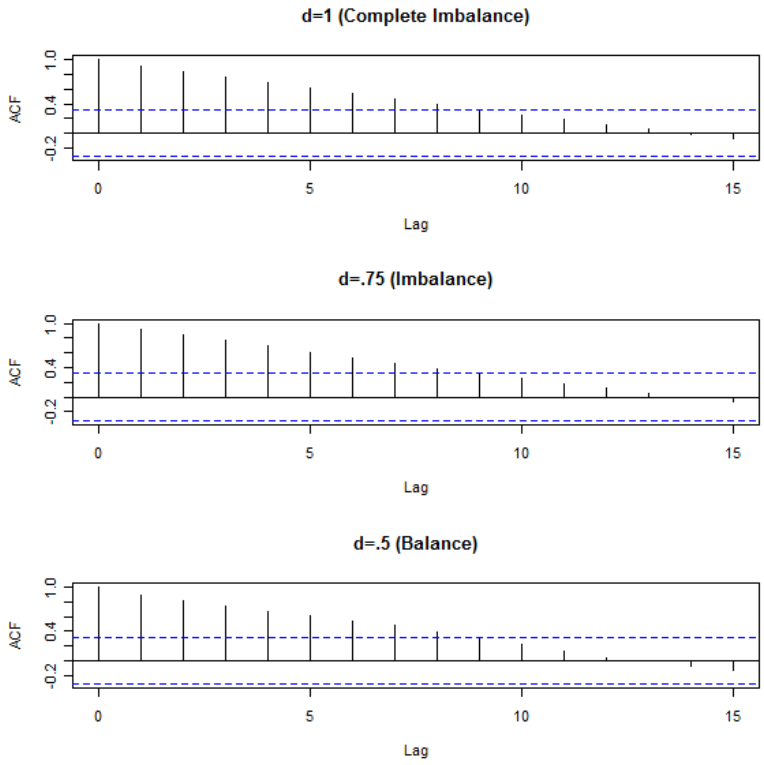


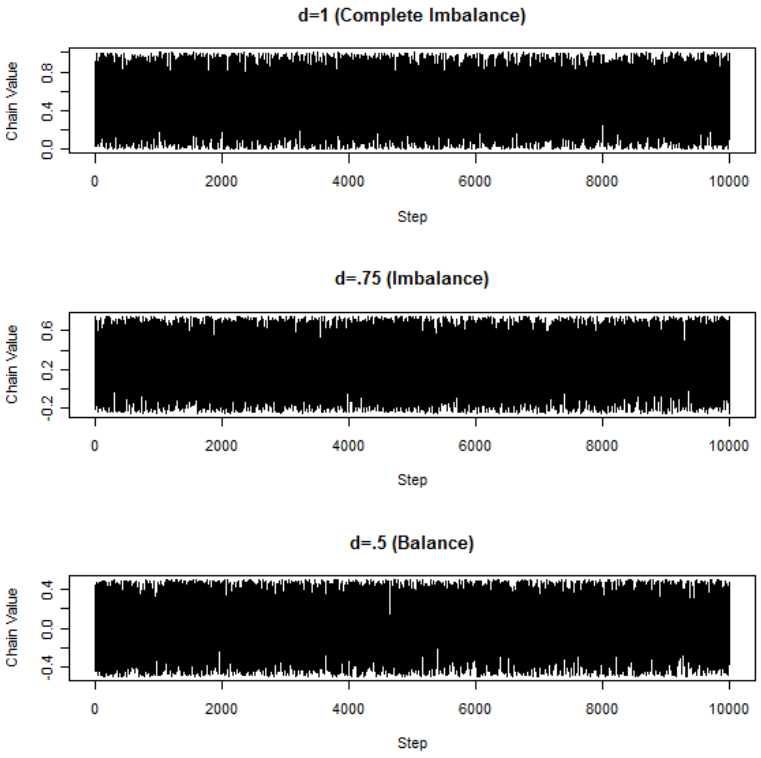
Figure 20 shows the chains produced for all levels of d when c is equal to 1, after thinning. All chains in this set of conditions show that there is still some degree of linear dependence remaining among the elements of the thinned chains.

In summary, the thinning done for all levels of c (except for the case where c is equal to 1), achieved the desired effect of a linearly independent sequence of elements. When these thinned sequences are characterized by way of the indicator statistic, it is found that none of the chains in any condition would be deemed non-converged. Thus, the effect of thinning on the value of D_t is to bring it close to the expected value of .67.

The path plots for the thinned chains are presented in Figures 21 through 26. Each figure contains the path plots for three levels of d, and there is an individual figure for each level of c. Again, see Figures 9-14 for characteristics of un-thinned chains.

Figure 21 presents the path plots for the three levels of d for the case where c is equal to zero for the thinned chains. It is presented and then described immediately below.

Figure 21: Path plots for all levels of d when c = 0



As can be seen in Figure 21, each of these chains traverses the space between the bounds of the respective distribution. As these are i.i.d. sequences, all values stay within

the bounds as specified by the levels of d.

Figure 22 presents the path plots for the three levels of d for the case where c is equal to .25 for the thinned chains. It is presented and then described immediately below.

Figure 22: Path plots for all levels of d when c = .25

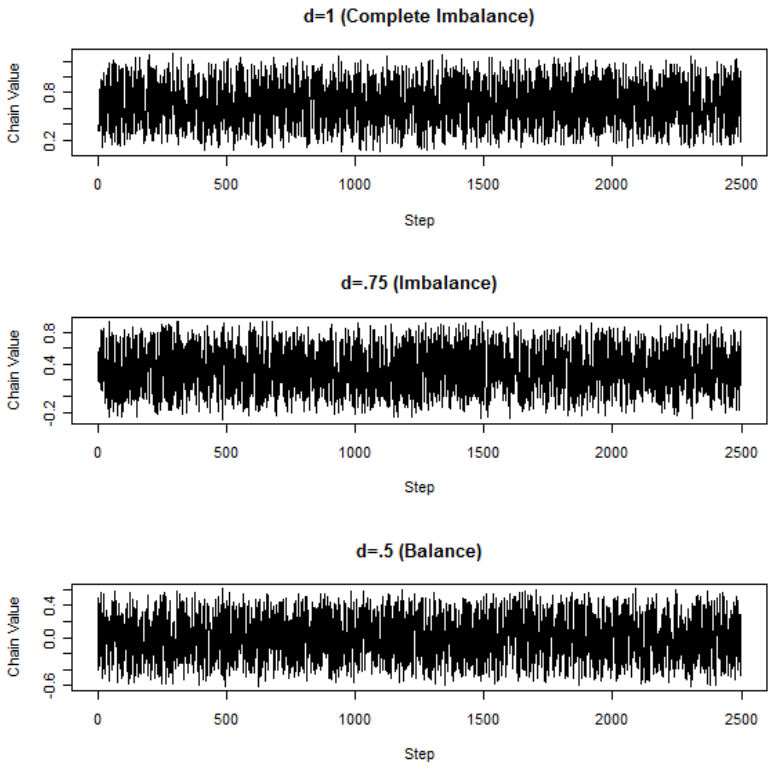


Figure 22 shows the same expansion of range that was seen in the descriptive statistics. Each of these chains has been reduced to 2500 elements by the thinning. The chains are virtually indistinguishable from one another in their behavior except for the bounds within which they traverse. The expanded range can be seen in the plots.

Figure 23 is presented below. It contains the path plots for all levels of d when c is equal to .5.

Figure 23: Path plots for all levels of d when c = .5

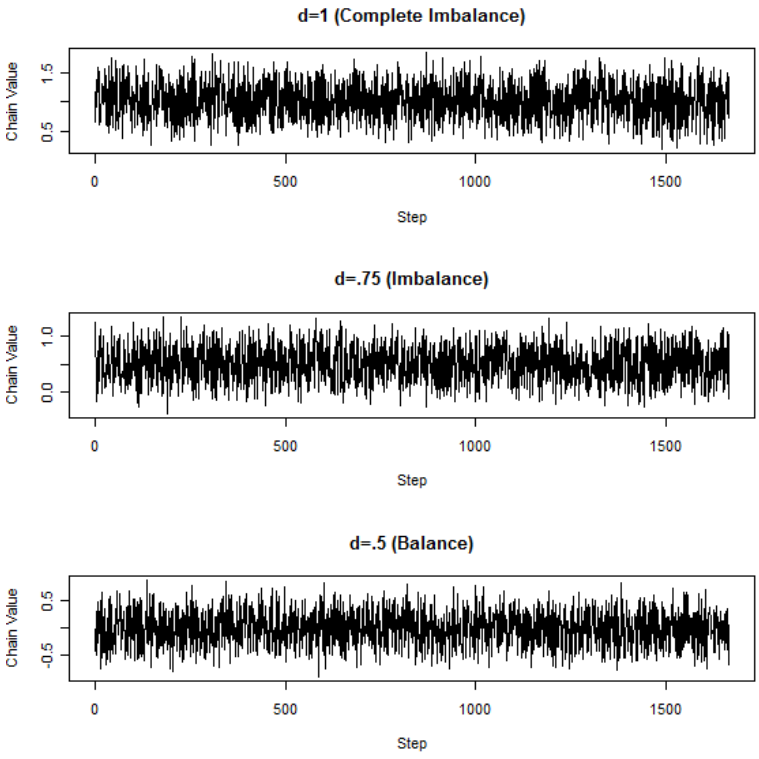


Figure 23 shows the same expansion of range that was seen in the descriptive statistics. Each of these chains has been reduced to 1667 elements by the thinning. The chains are virtually indistinguishable from one another.

Figure 24 presents the path plots for all levels of d when c is equal to .75. It is presented below.

Figure 24: Path plots for all levels of d when c = .75

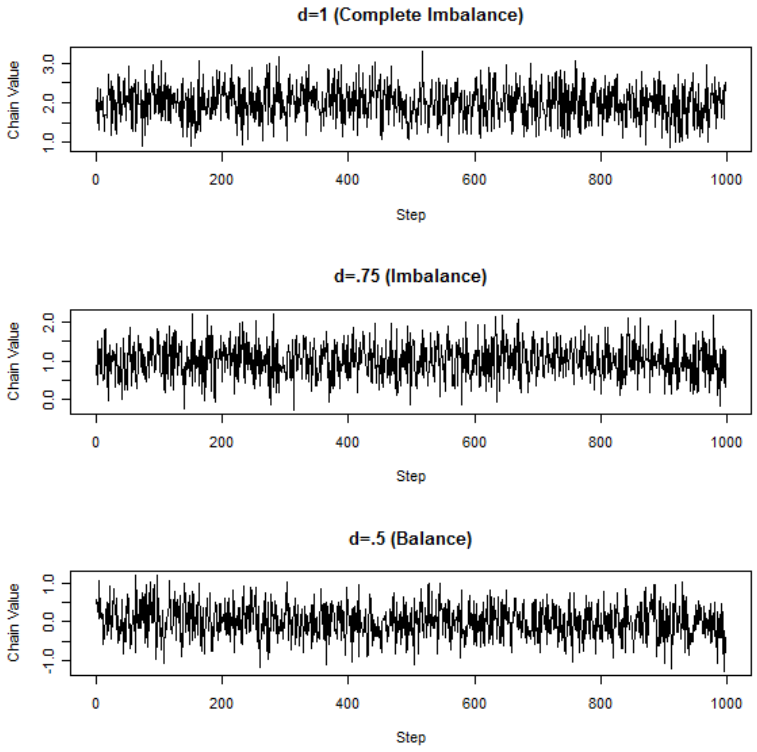


Figure 24 shows the same expansion of range that was seen in the descriptive statistics. Each of these chains has been reduced to 1000 elements by the thinning. The chains are virtually indistinguishable from one another.

Figure 25 shows the path plots for all levels of d when c is equal to .9. It is presented below.

Figure 25: Path plots for all levels of d when c = .9

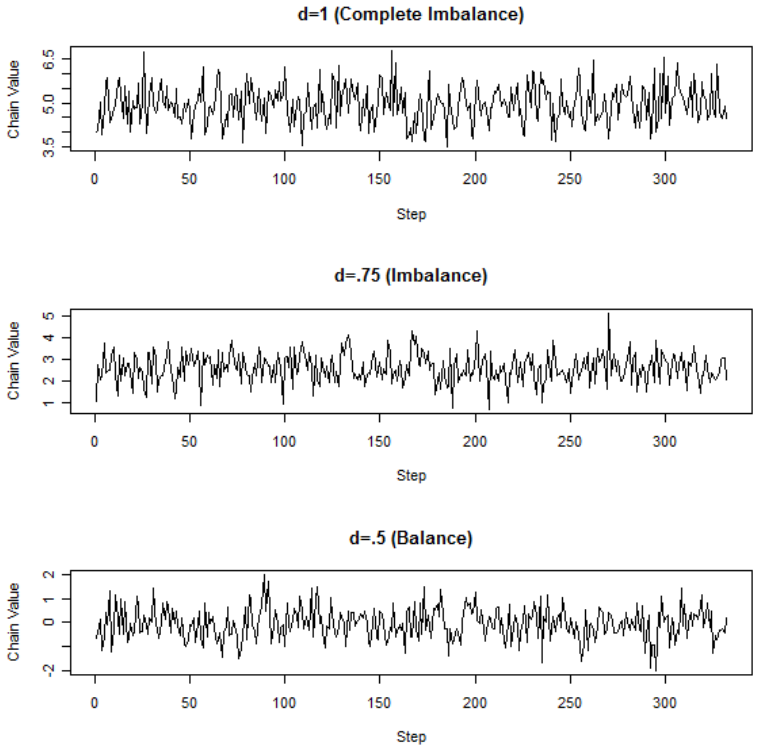


Figure 25 shows the same expansion of range that was seen in the descriptive statistics. Each of these chains has been reduced to 333 elements by the thinning. The chains are virtually indistinguishable from one another.

Figure 26 presents the path plots for the three levels of d for the case where c is equal to one for the thinned chains. It is presented and then described immediately below.

Figure 26: Path plots for all levels of d when c = 1

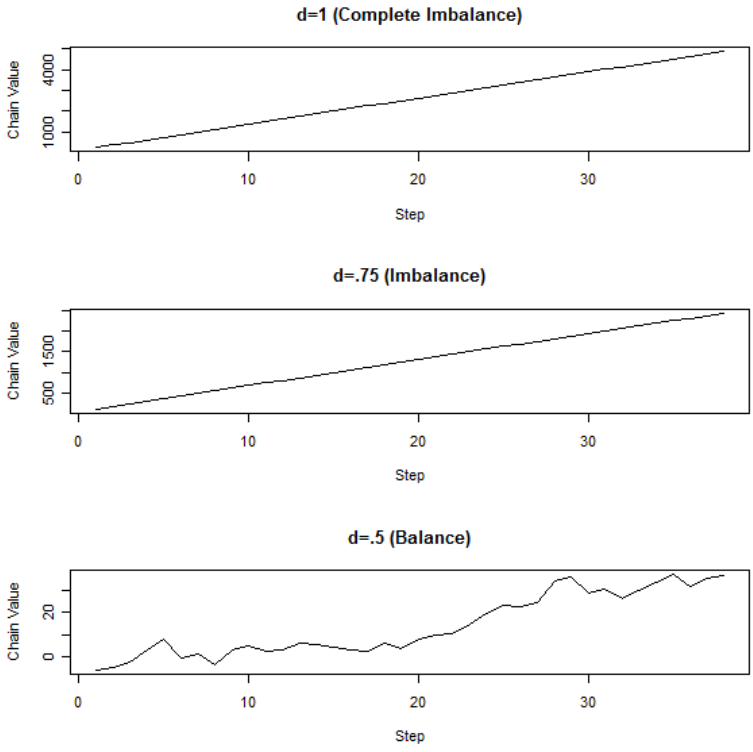


Figure 26 shows the chains for the case where c is equal to one. As was seen in the autocorrelation plots for these conditions, Figure 20, the strong degree of autocorrelation was not removed by the thinning. This was due to the fact that thinning until an i.i.d. sequence is achieved would leave these chains with only three elements left. As the indicator statistic looks at three elements simultaneously, nothing can be gained from thinning the chains to this point. Also, because the criticism of thinning is that it reduces the quality of the estimates obtained, reducing these chains any more than was done here would leave a chain that was so reduced so as not to be useful.

In summary, when thinning is performed on the chain and an i.i.d. sequence is obtained, then the value of the summary of the indicator statistic tends towards .67, and all conditions produce values within the bounds as specified in the methods section. These chains behave as if converged when characterized by the indicator statistic after thinning. The amount of autocorrelation present has an influence on the amount of thinning necessary to obtain an i.i.d. sequence. As autocorrelation increases, then more thinning is necessary to achieve linear independence. The question of whether or not thinning is artificially making the chains look 'converged' will be addressed in the discussion.

The second research question was also addressed in Simulation study 3. Simulation study 3 involved creating chains with a real MCMC sampler under conditions that would influence the autocorrelation present in the chains. Code was written to implement a Metropolis Hastings within Gibbs sampler in R. The code is presented in Appendix A. The factor of interest in this study was the relationship between the variability of the target and proposal distributions (RATIO). The ratio of variabilities is known to influence the behavior of the sampler, and was described previously. There are three levels of the RATIO factor. The first level is the case where the variability of the proposal distribution is one quarter the size of the variability of the target distribution (e.g., sd = .25 and 1, respectively). The second level is the case where the variability of the proposal distribution and the target distribution is equal (e.g., sd = 1 and 1, respectively. The third level is the case where the variability of the proposal distribution is four times larger than the variability of the target distribution (e.g., sd =4 and 1,

respectively). These levels were controlled by ensuring that the proposal distribution was constrained to have the appropriate variability compared to the conditions used to generate the true parameter values on which the simulated data were based. These parameters are discussed shortly.

Data was generated following the 2-parameter logistic model (2PL; Hambleton and Swaminathan, 1985) for 1000 persons and twenty items. The item response function defining the probability of a correct response for the 2PL is given as:

$$P_i(\theta|a_i, b_i) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \qquad (16)$$

where θ refers to an individual examinee's ability, a_i is an item discrimination parameter, b_i is an item difficulty parameter and D is a scaling constant that is equal to 1 for logistic scaling and 1.7 for normal ogive scaling. Item parameters and person ability parameters were generated at random. The "*a*" parameters were simulated from a normal distribution with a mean of 0 and a standard deviation of .2, and then the exponent of the values was taken to ensure that all values were positive. For the first level of the factor RATIO, the standard deviation of the proposal distribution was set equal to .05, or one quarter of the true distribution. For the second level of the factor RATIO, the standard deviation of the proposal distribution for "a" parameters was set equal to .2. For the third level of the factor RATIO, the standard deviation of the proposal distribution for "a" parameters was set equal to .8. The "b" and θ parameters were simulated from a standard normal distribution. For the first, second, and third levels of the factor RATIO, the standard

deviation of the proposal distribution for these parameters was set equal to .25, 1, and 4, respectively. Data will be presented for the item parameters, but not the person parameters. The results concerning the person parameters are more than enough to illustrate the influence of RATIO on the chains produced from each type of sampler.

The probability of a correct response for each simulated examinee to each item was calculated based on the item and person parameters randomly generated as described above. Each of these probabilities was then compared to a random value generated from a continuous uniform distribution with lower and upper bounds of 0 and 1, respectively. When the probability of a correct response was larger than the corresponding random value, then it was coded as 1 to indicate a correct response; otherwise it was coded a 0 to indicate an incorrect response. A MCMC sampler corresponding to each of the three conditions of RATIO was applied to each dataset. In this way, the three levels of the factor are applied to the same dataset, and this process is repeated 25 times. Each chain was run for 10,000 steps, and the first 5000 iterations are removed as burn-in. While this amount of thinning may be more than is necessary, it is commonly done to ensure that the resulting chains have settled to a location. All of the results presented for this simulation study will be based on the final chain of 5000 elements.

The findings for simulation study 3 will be presented next. First, the average autocorrelation at each lag will be presented for the three levels of RATIO. The autocorrelations for the a parameters and b parameters will be presented in separate tables. The average autocorrelation is informative about the intended effect of the factor RATIO. Following the autocorrelations, examples of chains from the three levels of

RATIO will be presented. Again, plots of all chains will not be presented for the sake of economy. After inspection of the chains from all replications of all conditions, it is sufficient to show a few chains to exemplify the trends present in the data. Following the path plots, the mean D_t values for the chains will be presented.

Table 13 presents the average value of the autocorrelation at each lag from 1 through 25 across all chains for all replications for the "a" parameters for each of the levels of RATIO. It is presented below

Table 13: Average autocorrelations for ‘a’ parameters for all levels of RATIO

	‘a’ Parameters		
Lag	**¼**	**1**	**4**
1	.965(.1e-15)	.973(.000)	.996(.000)
2	.931(.006))	.947(.009)	.992(.005)
3	.899(.011)	.922(.017)	.988(.010)
4	.869(.017)	.898(.025)	.985(.015)
5	.839(.021)	.874(.032)	.980(.020)
6	.811(.026)	.851(.038)	.977(.025)
7	.784(.030)	.829(.044)	.973(.030)
8	.758(.034)	.807(.050)	.970(.035)
9	.733(.037)	.787(.055)	.966(.040)
10	.708(.041)	.766(.060)	.962(.045)
11	.685(.044)	.747(.064)	.959(.046)
12	.663(.047)	.727(.069)	.955(.047)
13	.641(.049)	.709(.073)	.952(.048)
14	.620(.052)	.691(.076)	.948(.050)
15	.600(.054)	.673(.080)	.945(.051)
16	.581(.056)	.656(.083)	.941(.052)
17	.562(.058)	.640(.086)	.938(.053)
18	.545(.060)	.624(.088)	.934(.055)
19	.527(.062)	.608(.091)	.931(.056)
20	.511(.064)	.593(.093)	.927(.057)
21	.495(.065)	.578(.096)	.924(.059)
22	.479(.066)	.564(.098)	.921(.060)
23	.464(.068)	.550(.100)	.917(.061)
24	.450(.069)	.536(.102)	.914(.063)
25	.435(.070)	.523(.103)	.910(.064)

In Table 13 it can be seen that as the ratio of the proposal distribution variability to the target distribution variability decreases, there is less autocorrelation present at each lag of 1 through 25. When the proposal distribution is less variable than the target distribution, there is less dependence among draws. These findings are evidence that variations in the sampling mechanism affect the characteristics of the resulting Markov

chains. These findings are slightly different than what was expected based on previous literature. These findings will be revisited in the discussion.

Table 14 presents the average value of the autocorrelation at each lag from 1 through 25 across all chains for all replications for the b parameters for each of the levels of RATIO. It is presented below.

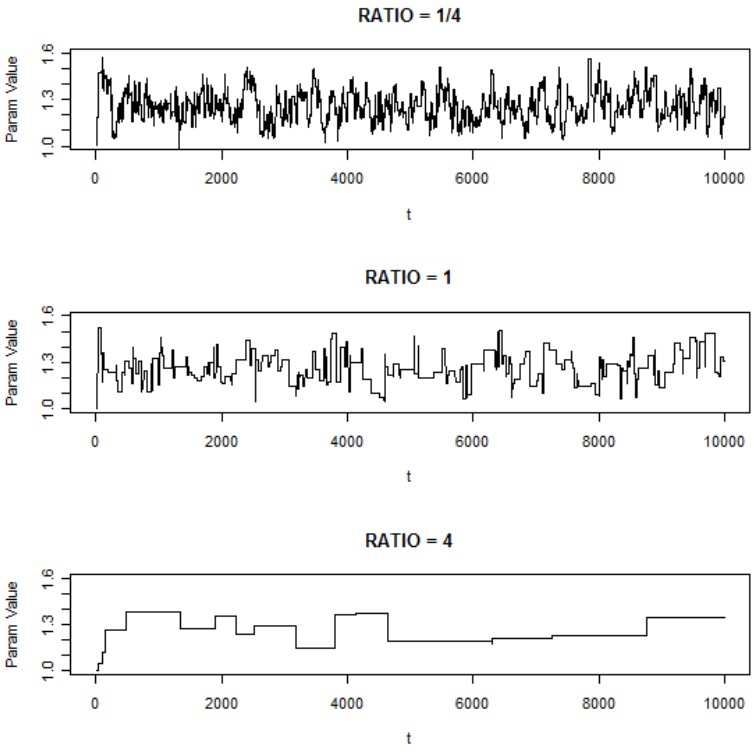
Table 14: Average autocorrelations for 'b' parameters for all levels of RATIO

	'b' Parameters		
Lag	**¼**	**1**	**4**
1	.849(.9e-16)	.967(.000)	.996(.000)
2	.738(.031)	.936(.012)	.992(.004)
3	.656(.051)	.906(.022)	.988(.009)
4	.593(.065)	.878(.032)	.984(.013)
5	.544(.075)	.852(.040)	.980(.018)
6	.506(.081)	.826(.047)	.976(.022)
7	.476(.085)	.802(.054)	.972(.026)
8	.451(.088)	.779(.060)	.968(.031)
9	.430(.089)	.757(.066)	.964(.035)
10	.413(.090)	.736(.071)	.961(.039)
11	.397(.090)	.716(.076)	.957(.041)
12	.384(.089)	.696(.080)	.953(.042)
13	.372(.088)	.678(.084)	.949(.044)
14	.362(.087)	.660(.087)	.946(.046)
15	.352(.086)	.643(.091)	.942(.047)
16	.343(.085)	.627(.094)	.939(.049)
17	.335(.083)	.612(.097)	.935(.051)
18	.327(.082)	.597(.099)	.931(.053)
19	.320(.081)	.583(.102)	.928(.055)
20	.313(.079)	.569(.104)	.924(.056)
21	.306(.078)	.555(.106)	.921(.058)
22	.299(.077)	.543(.108)	.917(.060)
23	.293(.076)	.531(.110)	.914(.062)
24	.287(.075)	.519(.112)	.910(.064)
25	.281(.074)	.507(.113)	.907(.065)

In Table 14 it can again be seen that as the ratio of the proposal distribution variability to the target distribution variability decreases, there is less autocorrelation present at each lag of 1 through 25. When the proposal distribution is less variable than the target distribution, there is less dependence among chain elements. These findings again indicate that there is an effect of RATIO on the characteristics of the resulting Markov chains.

Examples of the path plots for chains produced by the MCMC samplers corresponding to the three levels of RATIO will now be presented. The path plots provide information regarding the behavior of the chains over time. Also, the particular shape of the plots can provide feedback concerning whether or not the manipulations of RATIO created chains with differing appearances. There will be a figure for the a parameters and a separate figure for the b parameters. In each figure, all levels of RATIO will be plotted to allow for direct comparison.

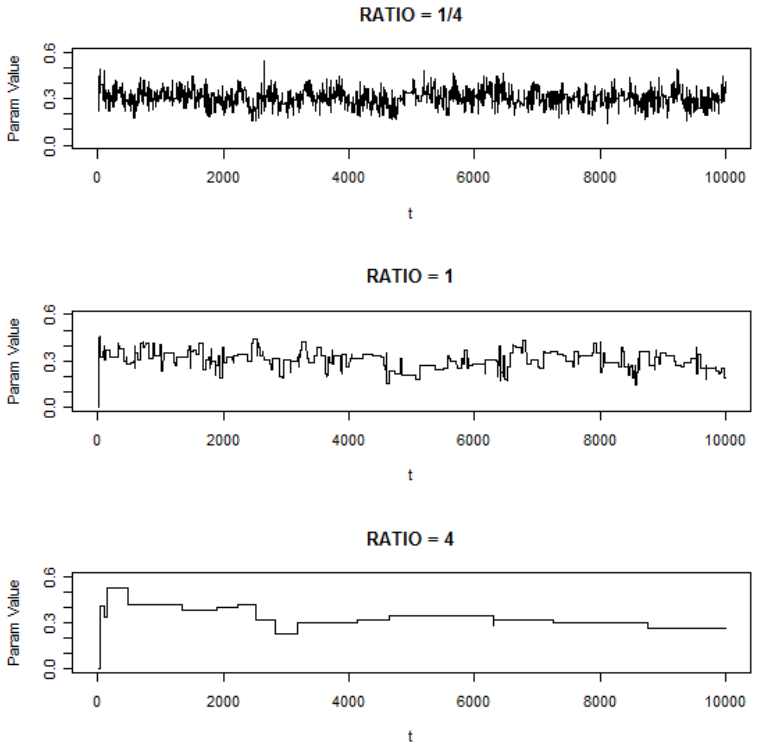
Figure 27: Path plots for all levels of RATIO for the ‘a’ parameters



In Figure 27, the effect of RATIO can be seen in the behavior of the chains for the a parameters. When RATIO is equal to 1/4, there are the fewest ties of the experimental conditions considered here. More unique values were accepted into these chains, and the chains explore more of the parameter space than those produced for the cases where RATIO is equal to 1 and 4. When RATIO increases, the increased variability of the proposed values means that fewer unique values are accepted into these chains. In fact, when RATIO is equal to 4, there are such a large number of rejections that in the example presented here there are only 15 unique values in this chain of 10,000.

Figure 28 will present the path plots for the three levels of RATIO for the a parameters. It is presented below.

Figure 28: Path plots for all levels of RATIO for the 'b' parameters



In Figure 28, the effect of RATIO can be seen in the behavior of the chains for the b parameters. The results are similar to those for the a parameters. When RATIO is equal to 1/4, there are the fewest ties of the experimental conditions considered here. More unique values were accepted into these chains, and the chains explore more of the parameter space than those produced for the cases where RATIO is equal to 1 and 4. When RATIO increases, the increased variability of the proposed values means that

fewer unique values are accepted into these chains.

To briefly summarize the path plots, it is important to note that while the levels of RATIO used in this study produced samplers that were not expected on the basis of Hanson and Cunningham (1998), the ratio of standard deviations of the proposal and target distributions has an effect on the acceptance ratios and consequently the chains. These path plots show that the characteristics of the samplers affect the behavior of the chains. Next, the values of D_t for these chains will be presented.

Table 15 presents the descriptive statistics for the values of D_t for the three levels of RATIO for both the a parameters and the b parameters. The mean value of D_t, as well as the standard deviation, maximum and minimum will be presented for each of the levels of RATIO. These descriptive statistics are based on all D_t values for all item parameters across all 25 replications. The purpose of this table is to demonstrate the general effect of RATIO on the behavior of the summary of the indicator statistic.

Table 15: Summary of D_t across all chains and replications for each level of RATIO

		RATIO		
		1/4	**1**	**4**
Parameter	**Statistic**			
a	Mean	.031	.0010	.00002
	SD	.017	.0012	.00005
	Max	.012	.0000	.00000
	Min	.183	.0170	.00040
b	Mean	.038	.0010	.00002
	SD	.019	.0013	.00005
	Max	.015	.0000	.00000
	Min	.200	.0190	.00040

The first thing noticeable in Table 15 is that the mean value of D_t is very small for all levels of RATIO. The reason for the small values of D_t is that the chains contained a great deal of ties. Ties are commonly encountered when constructing samplers using the Metropolis-Hastings algorithm. The sampler essentially compares each new candidate to the previous entry. Thus, when the variability of the proposal distribution is relatively large, it is more often the case that new candidates are proposed that are rejected. The value of D_t is directly related the amount of ties present in the chains produced. The particular samplers set up in this simulation study resulted in chains having an increasing amount of ties as RATIO increased. Without thinning, none of these ties are removed. The indicator statistic is currently defined in terms of strict inequalities. Thus, the ties are coded as zeroes, and the value of D_t is decreased. The issue of ties and how to deal with them will be revisited in the Discussion.

There is an inverse relationship between the amount of autocorrelation present in the chains and the mean value of D_t. It is expected that as the amount of autocorrelation increases, the value of D_t decreases. For the conditions presented in simulation study 3, the smallest degree of autocorrelation is present in the chains generated for the case where RATIO is equal to 1/4. For these chains D_t takes on the largest value on average. As RATIO increases, we see stronger degrees of autocorrelation as well as the increase in ties. Correspondingly, the mean values of D_t for these conditions show a decrement compared to the case where RATIO is equal to 1/4. When RATIO is equal to 4, very few of the chain elements satisfy the indicator statistic being equal to 1. On average, only 4 elements in these chains are coded as ones according to the indicator statistic. The

relationship between the autocorrelation and the number of ties in the chain both are influencing the value of D_t.

It is important to show at this point the quality of the estimates provided by these chains. The convergence diagnostic D_t must be linked to the quality of the estimates. The most important criterion when determining the quality of chains is how close the estimates are to the true values of the parameters. Because these data are simulated and truth is known, it is important to show how close the estimates were to the true values for the parameters. Table 16 presents the mean absolute deviations (MADs) for the estimates of the 'a' parameters for each level of RATIO for each of the 25 replications. In this table, the MAD and the variability of the absolute deviations across the 20 items in each replication will be reported. The average MAD across all replications for each level of RATIO and its standard error will also be included. The purpose of this table is to show the quality of the estimates provided by the samplers representing each level of RATIO.

Table 16: MADs for all levels of RATIO for the ‘a’ parameters

MADs for ‘a’ Parameters			
	1/4	**1**	**4**
Replications			
1	.063(.043)	.058(.040)	.071(.043)
2	.064(.052)	.069(.063)	.065(.050)
3	.065(.056)	.064(.053)	.081(.063)
4	.086(.092)	.084(.093)	.092(.090)
5	.086(.059)	.084(.064)	.073(.063)
6	.061(.045)	.063(.043)	.061(.047)
7	.099(.052)	.099(.054)	.097(.074)
8	.073(.056)	.070(.056)	.084(.060)
9	.109(.109)	.110(.113)	.126(.124)
10	.066(.041)	.063(.039)	.067(.050)
11	.058(.039)	.060(.038)	.066(.046)
12	.073(.046)	.072(.048)	.077(.055)
13	.100(.082)	.095(.078)	.098(.077)
14	.072(.055)	.078(.054)	.071(.054)
15	.060(.054)	.058(.054)	.059(.057)
16	.071(.080)	.072(.080)	.079(.085)
17	.058(.045)	.049(.048)	.065(.044)
18	.099(.075)	.100(.077)	.106(.080)
19	.072(.050)	.071(.049)	.072(.057)
20	.085(.070)	.085(.077)	.083(.070)
21	.072(.070)	.075(.076)	.076(.066)
22	.062(.048)	.063(.046)	.072(.056)
23	.062(.041)	.063(.041)	.065(.048)
24	.073(.061)	.065(.055)	.073(.067)
25	.057(.055)	.062(.056)	.051(.057)
Mean MAD(SE)	.073(.015)	.073(.015)	.077(.016)

Table 16 shows that chains from each level of RATIO do an equally good job of recovering the true parameters. Overall, for all levels of RATIO, estimates were quite close to truth. These findings are important in that they show that the quality of the estimate is not immediately revealed by autocorrelations or the value of any particular

convergence diagnostic. Although D_t differed across experimental conditions, the quality of the estimates did not.

Table 17: MADs for all levels of RATIO for the 'b' parameters

	MADs for 'b' Parameters		
Replications	**¼**	**1**	**4**
1	.047(.041)	.048(.044)	.050(.037)
2	.050(.048)	.061(.049)	.068(.060)
3	.070(.102)	.066(.093)	.072(.094)
4	.079(.086)	.077(.087)	.073(.086)
5	.060(.043)	.054(.042)	.055(.037)
6	.085(.067)	.076(.059)	.087(.072)
7	.083(.062)	.087(.061)	.086(.073)
8	.059(.041)	.057(.041)	.054(.032)
9	.117(.225)	.120(.238)	.139(.228)
10	.064(.045)	.059(.043)	.056(.041)
11	.059(.046)	.071(.048)	.071(.049)
12	.046(.033)	.046(.032)	.055(.027)
13	.107(.087)	.097(.082)	.120(.092)
14	.072(.084)	.074(.090)	.076(.097)
15	.051(.036)	.052(.036)	.048(.034)
16	.066(.041)	.063(.040)	.071(.041)
17	.055(.045)	.052(.048)	.057(.051)
18	.085(.076)	.086(.075)	.087(.083)
19	.052(.044)	.053(.043)	.065(.047)
20	.048(.051)	.047(.050)	.056(.052)
21	.060(.067)	.059(.071)	.055(.054)
22	.047(.044)	.050(.045)	.068(.051)
23	.058(.060)	.060(.059)	.062(.068)
24	.058(.054)	.046(.046)	.060(.054)
25	.059(.036)	.056(.033)	.056(.044)
Mean MAD(SE)	.065(.018)	.065(.018)	.070(.021)

Table 17 is very similar to Table 16. Table 17 also shows that chains from each level of RATIO do an equally good job of recovering the true b parameters. Overall, for

all levels of RATIO, estimates were quite close to truth. These findings are important in that they show that the quality of the estimate is not immediately revealed by autocorrelations or the value of any particular convergence diagnostic. Although D_t differed across experimental conditions, the quality of the estimates did not.

It is appropriate to briefly summarize the information about the quality of the estimates in the chains for this study. D_t showed a great deal of sensitivity to ties and autocorrelation, but all chains for a parameters for all conditions produced very good estimates. The indicator statistic under development is influenced by RATIO, but the quality of the estimates for the chains when no thinning is done is good as indicated by small MADs across items and replications. These findings will be revisited in the discussion.

The results for the thinned chains will now be presented. First, the average autocorrelation at each lag will be presented for the three levels of RATIO after thinning has been done. The autocorrelations for the a parameters and b parameters will be presented in separate tables. The average autocorrelations presented in these tables will provide feedback concerning the effect of thinning the chains. Following the autocorrelations, examples of chains from the three levels of RATIO after thinning will be presented. Again, plots of all chains will not be presented for the sake of economy. After inspection of the chains from all replications of all conditions, it is deemed sufficient to show a few chains to exemplify the trends present in the data. Following the path plots, the average D_t values for the chains after thinning will be presented.

Table 18: Average value of AC at lag 1 through 25 for thinned chains for ‘a’ parameters

	‘a’ Parameters		
	¼	1	4
Lag			
1	.116(.125)	.078(.148)	.255(..277)
2	.005(.124)	-.005(.138)	-.002(.250)
3	-.014(.119)	-.025(.141)	-.078(.209)
4	-.011(.109)	-.020(.134)	-.098(.182)
5	-.022(.110)	-.024(.124)	-.108(.184)
6	-.013(.109)	-.020(.136)	-.100(.195)
7	-.013(.108)	-.019(.125)	-.104(.178)
8	-.025(.105)	-.014(.128)	-.082(.158)
9	-.021(.110)	-.028(.126)	-.078(.149)
10	-.012(.109)	-.027(.122)	-.057(.128)
11	-.015(.112)	-.022(.129)	-.035(.111)
12	-.018(.098)	-.016(.118)	-.014(.072)
13	-.015(.106)	-.024(.120)	
14	-.010(.100)	-.016(.120)	
15	-.013(.104)	-.019(.118)	
16	-.011(.104)	-.024(.117)	
17	-.016(.101)	-.019(.112)	
18	-.019(.104)	-.010(.113)	
19	-.012(.101)	-.014(.109)	
20	-.008(.099)	-.014(.106)	
21	-.012(.104)	-.015(.107)	
22	-.016(.101)	-.013(.105)	
23	-.013(.098)	-.016(.102)	
24	-.019(.095)	-.014(.102)	
25	-.013(.096)	-.015(.102)	

In Table 18, the effect of thinning can be seen in that the values of the autocorrelations are much smaller than when thinning is not done (see Table 13). At lag 1, there is a small positive association for all levels of RATIO for the “a” parameters. At lag 2, the association is even smaller, practically 0. These findings indicate that the

thinning is having its desired effect, and the remaining sequence is at least linearly independent.

Table 19: Average value of AC at lag 1 through 25 for thinned chains for 'b' parameters

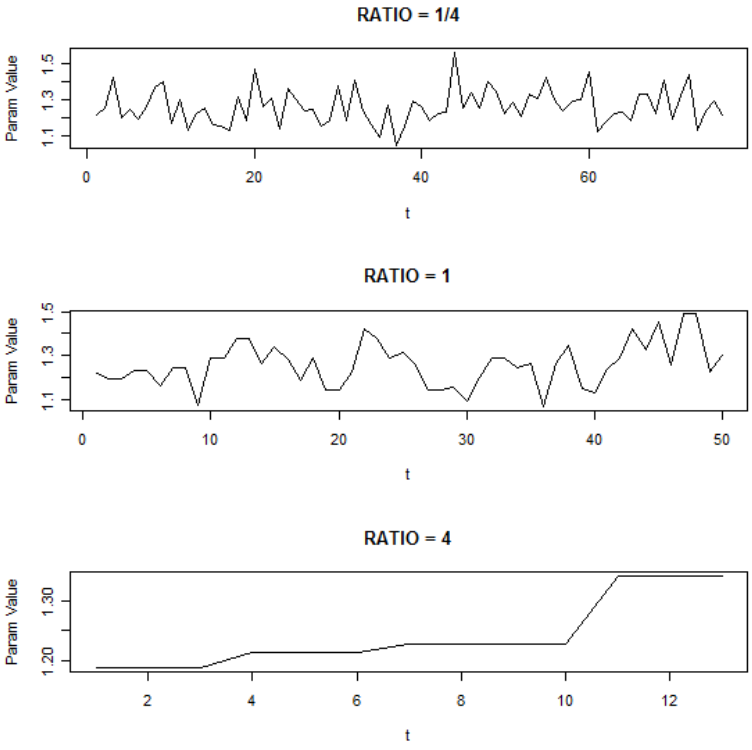
	'b' Parameters		
	¼	1	4
Lag			
1	.122(.122)	.178(.164)	.390(.254)
2	.032(.112)	.082(.150)	.142(.251)
3	-.009(.113)	.032(.143)	.026(.225)
4	-.012(.115)	-.003(.140)	-.021(.200)
5	-.013(.115)	-.022(.131)	-.063(.177)
6	-.016(.116)	-.022(.131)	-.084(.160)
7	-.017(.114)	-.026(.131)	-.091(.171)
8	-.013(.110)	-.024(.132)	-.099(.166)
9	-.022(.109)	-.037(.132)	-.101(.158)
10	-.012(.107)	-.042(.125)	-.104(.159)
11	-.008(.108)	-.030(.123)	-.099(.151)
12	-.011(.106)	-.018(.120)	-.090(.145)
13	-.024(.110)	-.025(.115)	-.081(.145)
14	-.018(.106)	-.018(.116)	-.062(.131)
15	-.007(.106)	-.026(.120)	-.049(.123)
16	-.009(.105)	-.024(.113)	-.042(.113)
17	-.020(.095)	-.016(.115)	-.034(.094)
18	-.028(.102)	-.011(.117)	-.025(.076)
19	-.020(.103)	-.017(.114)	-.012(.051)
20	-.016(.099)	-.024(.104)	
21	-.016(.099)	-.036(.105)	
22	-.010(.092)	-.032(.106)	
23	-.015(.097)	-.029(.099)	
24	-.014(.092)	-.031(.106)	
25	-.015(.091)	-.031(.100)	

In Table 19, the values of the average autocorrelations at lags 1 through 25 are presented for the 'b' parameters. It can be seen that the same general pattern holds as does for the 'a' parameters, except that the autocorrelations are somewhat larger at lag 1

and 2. The largest average autocorrelation is observed at lag 1 for the case where RATIO is equal to 4. Again, it should be stated that the thinning was done on the basis of autocorrelation plots. When visually investigating the autocorrelation plots for the chains for the 'b' parameters when RATIO equals 4, none of the plots showed a correlation deemed significant (alpha = .05) at a lag greater than 250.

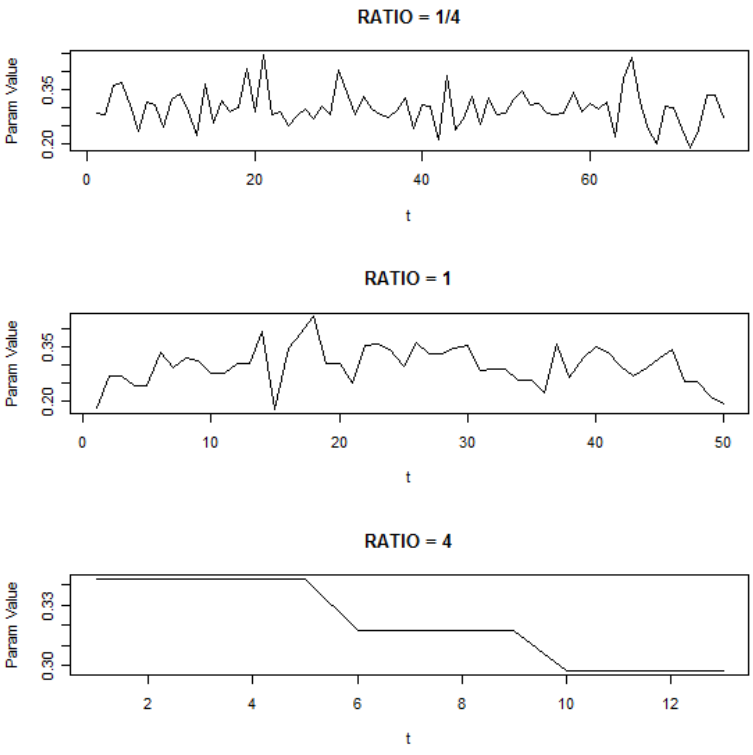
Next, the path plots for the thinned chains will be presented. Similar to the path plots for the chains that aren't thinned, each level of RATIO will be presented in each figure, and there will be a figure for both the "a" parameters and the b parameters. Figure 29 presents the path plots for the "a" parameters.

Figure 29: Path plots for all levels of RATIO for the 'a' parameters



It can be seen that the chains have been greatly reduced in length due to the thinning. However, the same general pattern holds for the thinned chains as was seen for the path plots for the chains without thinning. When RATIO is equal to 1/4, there are the fewest ties of the experimental conditions considered here. More unique values were accepted into these chains, and the chains explore more of the parameter space than those produced for the cases where RATIO is equal to 1 and 4. When RATIO increases, the increased variability of the proposed values means that fewer unique values are accepted into these chains.

Figure 30: Path plots for all levels of RATIO for the 'b' parameters



In Figure 30, the pattern of results for the b parameters can be seen to be very similar to those for the a parameters in Figure 29. When RATIO is equal to 1/4, there are the fewest ties of the experimental conditions considered here. More unique values were accepted into these chains, and the chains explore more of the parameter space than those produced for the cases where RATIO is equal to 1 and 4. When RATIO increases, the increased variability of the proposed values means that fewer unique values are accepted into these chains.

To briefly summarize, even after thinning, the characteristics of the samplers as influenced by RATIO demonstrate the same general pattern as the cases where thinning was not done. When RATIO is equal to 1/4, the chains take on more unique values than for the other levels, but the comparison is less clear due to unequal number of elements after thinning. The thinning has removed a great deal of the observed chain values, and the path plots have become a great deal shorter, which is problematic for practical reasons. This aspect of the issue of thinning will be revisited in the discussion.

Table 20 presents the descriptive statistics for the values of D_t for the three levels of RATIO for both the a parameters and the b parameters after the chains have been thinned. The mean value of D_t, as well as the standard deviation, maximum and minimum will be presented for each of the levels of RATIO. These descriptive statistics are based on all D_t values for all item parameters across all 25 replications. The purpose of this table is to demonstrate the general effect of thinning on the behavior of the summary of the indicator statistic.

Table 20: Summary of D_t across all thinned chains and replications for each level of RATIO

		RATIO		
		¼	1	4
Parameter	**Statistic**			
'a'	Mean	.687	.674	.673
	SD	.136	.141	.179
	Max	.960	.918	1.00
	Min	.347	.265	.250
'b'	Mean	.672	.666	.668
	SD	.157	.161	.163
	Max	.960	.959	.947
	Min	.267	.265	.263

Table 20 shows the effect that thinning has on the value of D_t. Thinning is removing both the large number of ties as well as the autocorrelation among chain elements, both of which influence the summary of the indicator statistic. The result is that the average value of D_t increases compared to the same chains without thinning. For all levels of RATIO for both 'a' and 'b' parameters, the average value of D_t goes the value that would be expected for an i.i.d. sequence. This finding is not to be interpreted as saying that all chains move towards 2/3. Rather, it is simply the case that the thinning is removing some qualities of the chains that are known to influence the value that D_t takes on.

Again, it is important to show at this point the quality of the estimates provided by these chains. The convergence diagnostic D_t must be linked to the quality of the estimates. The most important criterion when determining the quality of chains is how

close the estimates are to the true values of the parameters. Because these data are simulated and truth is known, it is important to show how close the estimates were to the true values for the parameters. Table 21 presents the mean absolute deviations (MADs) for the estimates of the 20 ‘a’ parameters for each of the 25 replications for each level of RATIO for the thinned chains. In this table, the MAD and the variability of the absolute deviations across the 20 items in each replication will be reported. The average MAD across all replications for each level of RATIO and its standard error will also be included. The purpose of this table is to show the quality of the estimates provided by the samplers representing each level of RATIO when thinning is done.

Table 21: MADs for all levels of RATIO for the 'a' parameters for thinned chains

MADs for 'a' Parameters (thinned chains)			
	¼	1	4
Replications			
1	.064(.042)	.057(.043)	.069(.045)
2	.063(.051)	.067(.063)	.068(.051)
3	.066(.057)	.067(.054)	.082(.059)
4	.085(.091)	.089(.092)	.096(.097)
5	.087(.061)	.086(.064)	.072(.060)
6	.059(.047)	.065(.041)	.060(.049)
7	.099(.153)	.100(.055)	.097(.074)
8	.072(.056)	.069(.055)	.086(.057)
9	.111(.109)	.109(.112)	.127(.127)
10	.064(.042)	.061(.038)	.065(.050)
11	.057(.039)	.059(.038)	.064(.047)
12	.074(.046)	.074(.048)	.077(.056)
13	.101(.084)	.096(.079)	.097(.078)
14	.072(.053)	.079(.056)	.065(.057)
15	.062(.053)	.058(.055)	.059(.061)
16	.071(.081)	.072(.078)	.080(.086)
17	.059(.045)	.051(.049)	.066(.050)
18	.100(.075)	.099(.076)	.108(.076)
19	.075(.052)	.073(.048)	.071(.061)
20	.083(.071)	.084(.079)	.083(.071)
21	.072(.070)	.079(.077)	.079(.067)
22	.063(.050)	.065(.044)	.070(.055)
23	.062(.042)	.061(.043)	.067(.043)
24	.073(.062)	.068(.054)	.077(.067)
25	.058(.056)	.061(.057)	.050(.058)
Mean MAD(SE)	.074(.015)	.074(.015)	.078(.017)

Table 21 shows that chains from each level of RATIO do an equally good job of recovering the true parameters, and these results are very similar to those for the un-thinned chains. Overall, for all levels of RATIO, estimates were quite close to truth. These findings are important in that they show that the quality of the estimate is not

affected by the thinning. The quality of the estimates did not suffer even when there was a great deal of thinning.

Table 22 contains the MADs for the 25 replications of the 'b' parameters. The mean MAD across replications and its standard error are also presented.

Table 22: MADs for all levels of RATIO for the 'b' parameters for thinned chains

MADs for 'b' Parameters (thinned chains)			
	¼	1	4
Replications			
1	.048(.041)	.049(.044)	.052(.041)
2	.051(.049)	.060(.048)	.067(.058)
3	.069(.103)	.068(.094)	.071(.099)
4	.078(.089)	.080(.092)	.076(.089)
5	.059(.049)	.056(.044)	.057(.036)
6	.085(.070)	.075(.059)	.085.069)
7	.082(.060)	.089(.062)	.085(.072)
8	.060(.040)	.056(.043)	.055(.033)
9	.117(.224)	.122(.234)	.138(.226)
10	.062(.046)	.055(.042)	.061(.043)
11	.062(.046)	.073(.049)	.069(.048)
12	.047(.030)	.050(.030)	.053(.028)
13	.107(.089)	.098(.084)	.121(.091)
14	.070(.085)	.076(.105)	.075(.088)
15	.052(.037)	.054(.038)	.048(.035)
16	.066(.042)	.062(.043)	.071(.041)
17	.053(.046)	.056(.053)	.059(.052)
18	.085(.078)	.085(.074)	.091(.088)
19	.051(.046)	.054(.042)	.057(.052)
20	.047(.051)	.048(.051)	.056(.054)
21	.060(.068)	.061(.072)	.070(.052)
22	.048(.044)	.052(.046)	.062(.065)
23	.060(.061)	.059(.058)	.056(.050)
24	.061(.056)	.046(.047)	.053(.040)
25	.060(.037)	.055(.035)	.061(.039)
Mean MAD(SE)	.066(.018)	.066(.018)	.070(.022)

Table 22 is very similar to Table 21. Table 22 also shows that chains from each level of RATIO do an equally good job of recovering the true b parameters. Overall, for all levels of RATIO, estimates were quite close to truth. These findings are important in that they show that the quality of the estimate is not necessarily affected by the thinning.

It is appropriate to briefly summarize the information about the quality of the estimates in the chains for this study. D_t showed a great deal of sensitivity to ties and autocorrelation, but all chains for 'a' parameters for all conditions produced very good estimates. The thinning of the chains had a great deal of influence on the value of the summary of the indicator statistic, but not on the MADs for the estimates produced by the chains. The indicator statistic under development is influenced by thinning, but the quality of the estimates is good as indicated by small MADs across items and replications. These findings will be revisited in the discussion.

Findings for research question 3

Research question 3 compares the diagnostic currently being investigated to three existing diagnostics. Simulation study 4 is similar to simulation studies 1 and 2 in that it also has chains created using the same levels of c and d. In addition, this study also varies the range of the random component of the chain simulator. The D_t values for these new conditions (i.e., where the range of the random component is equal to .1, .5, and 5) will be presented along with descriptive statistics, autocorrelation plots, and path plots. The D_t values, descriptive statistics, autocorrelation plots and path plots for the case where the range of the random component is equal to 1 will be omitted because they are virtually

the same as results presented in the previous studies, but this condition will be included in the comparison to the three existing diagnostics.

Presented below are the tables containing the D_t values for the three different ranges of the random component of the chain simulator. Each table is followed by a brief description of the results.

Table 23: Mean D_t values (SD) for combinations of balance (d) and AC factor (c) when range is equal to .1

	d				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.668(.004)	.666(.004)	.666(.004)	.668(.004)	.667(.004)
.25	.627(.005)	.626(.004)	.625(.004)	.625(.004)	.624(.004)
.50	.584(.004)	.584(.005)	.583(.003)	.584(.004)	.582(.004)
.75	.541(.005)	.542(.004)	.542(.005)	.542(.005)	.541(.005)
.90	.520(.004)	.517(.006)	.515(.005)	.514(.006)	.516(.004)
1.0	.000(.000)	.376(.005)	.500(.006)	.375(.006)	.000(.000)

Table 23 shows the same patterns of results for values of D_t based on simulated chains that were apparent in Table 4 for simulation study 1. The effect that the balance factor was intended to have is not apparent in this table either. Also, it is clear that setting the range of the random component equal to .1 did not have an influence on the value of D_t.

Table 24 presents the values of the summary of the indicator statistic for the case where the range is equal to .5. It is presented below.

Table 24: Mean D_t values (SD) for combinations of balance (d) and AC factor (c) when range is .5

	d				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.668(.004)	.666(.004)	.667(.004)	.667(.003)	.666(.004)
.25	.627(.005)	.626(.003)	.625(.004)	.624(.004)	.625(.003)
.50	.584(.004)	.585(.005)	.584(.003)	.583(.004)	.584(.003)
.75	.541(.003)	.542(.005)	.542(.004)	.541(.005)	.542(.005)
.90	.520(.005)	.516(.006)	.517(.005)	.512(.005)	.516(.005)
1.0	.000(.000)	.376(.005)	.499(.005)	.375(.005)	.000(.000)

Table 24 shows the same patterns of results for values of D_t based on simulated chains that were apparent in Tables 4 and 12. The effect that the balance factor was intended to have is not apparent in this table either. Also, it is clear that setting the range of the random component equal to .5 did not have an influence on the value of D_t.

Table 25 presents the values of D_t for the case where the range of the random component of the chain simulator is equal to 5. It is presented below.

Table 25: Mean D_t values (SD) for combinations of balance (d) and AC factor (c) when range is 5

	d				
c	**0**	**.25**	**.5**	**.75**	**1.0**
0.0	.668(.004)	.667(.003)	.668(.004)	.668(.004)	.666(.004)
.25	.628(.005)	.625(.003)	.624(.005)	.624(.004)	.626(.005)
.50	.583(.004)	.585(.004)	.582(.005)	.585(.005)	.585(.004)
.75	.541(.004)	.542(.005)	.542(.004)	.541(.006)	.542(.005)
.90	.521(.006)	.518(.005)	.517(.004)	.517(.005)	.518(.005)
1.0	.000(.000)	.374(.006)	.501(.005)	.377(.007)	.000(.000)

Table 25 shows the same patterns of results for values of D_t based on simulated chains that were apparent in Tables 4, 17, and 18. The effect that the balance factor was intended to have is not apparent in this table either. Also, it is clear that setting the range of the random component equal to 5 did not influence the value of D_t.

Taken together; these results indicate that the value of the indicator statistic is not influenced by the range of the random component of the chain simulator. This was expected for D_t given that the diagnostic is only sensitive to the rank orderings of the elements produced by the chain simulator. The relative rank orderings of values generated from a continuous uniform distribution with differing boundaries would still be expected to produce the same pattern.

Now the descriptive statistics will be provided for the three new conditions of RANGE. Each table will be presented and then followed by a brief description of the results. Table 26 (below) presents the descriptive statistics for all levels of c when there is complete imbalance and the range of the random component is .1. It is presented below.

Table 26: Descriptive statistics for simulated chains when d=1 (Complete Imbalance) and range = .1

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.050(.0003)	.00001(.00001)	.100(.00001)	.100(.00001)
.25	.067(.0004)	.003(.001)	.131(.001)	.128(.001)
.50	.100(.001)	.012(.004)	.187(.003)	.175(.004)
.75	.200(.001)	.04(.018)	.335(.006)	.295(.019)
.90	.499(.003)	.060(.030)	.729(.017)	.670(.035)
1.0	250(1.38)	.049(.030)	500(2.22)	500(2.21)

As can be seen in Table 26, the general pattern for all descriptive statistics that the mean and variability increases as c goes from 0 to 1. These values in the table reflect the range of the random component of the chain simulator. In general, the trend seen in this table reflects the trend seen in all similar descriptive statistic tables for the cases where there is complete imbalance.

Table 27 presents the descriptive statistics for all levels of c when there is partial imbalance and the random component is equal to .1. It is presented below.

Table 27: Descriptive statistics for simulated chains when d=.75 (Partial Imbalance) and range = .1

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.025(.003)	-.025(.00001)	.075(.00001)	.100(.00001)
.25	.003(.0003)	-.030(.001)	.097(.001)	.128(.001)
.50	.050(.001)	-.037(.005)	.138(.001)	.174(.002)
.75	.100(.001)	-.039(.005)	.239(.007)	.278(.007)
.90	.250(.003)	.007(.017)	.483(.017)	.477(.022)
1.0	125(1.46)	.022(.028)	250(2.52)	250(2.53)

For Table 27, the pattern is slightly different than for the previous table. It is still the case that as c increases the average mean, maximum, and range increases. It is also still the case that variability increases as c increases. However, when d represents partial imbalance (positive) the average minimum first decreases and then increases over the range of c. The descriptive statistics also reflect the influence of the range of the random component of the chain simulator. The pattern in Table 27 is similar to the trend seen in each other table of descriptive statistics for the case where partial imbalance is present in the chain simulator.

Table 28 presents the descriptive statistics for all levels of c when there is balance and the random component is equal to .1. It is presented below.

Table 28: Descriptive statistics for simulated chains when d=.5 (Balance) and range = .1

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	-.0001(.0002)	-.050(.00001)	.050(.00001)	.100(.00001)
.25	-.0001(.0003)	-.064(.001)	.064(.001)	.128(.001)
.50	.0001(.0006)	-.088(.002)	.088(.0022)	.175(.003)
.75	.00004(.001)	-.139(.007)	.140(.007)	.280(.009)
.90	-.0002(.004)	-.231(.018)	.226(.020)	.457(.026)
1.0	.033(1.58)	-2.01(1.43)	2.11(1.61)	4.11(1.04)

As can be seen in Table 28, as c increases the average minimum gets smaller and more variable, the average maximum gets larger and more variable, and the average range gets larger and more variable. The average mean has no discernible pattern, but its variability increases as c increases. The descriptive statistics in the table reflect the range of the random component of the chain simulator. The pattern of results in Table 22 is similar to all other conditions where there is balance present in the random component of the chain simulator.

Table 29 presents the descriptive statistics for all levels of c when there is complete imbalance and the random component is equal to .5. It is presented below.

Table 29: Descriptive statistics for simulated chains when d=1 (Complete Imbalance) and range = .5

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.250(.001)	.0001(.0001)	.500(.0001)	.500(.0001)
.25	.333(.002)	.013(.003)	.651(.004)	.638(.005)
.50	.499(.003)	.056(.012)	.937(.012)	.882(.014)
.75	1.00(.005)	.208(.100)	1.69(.033)	1.49(.099)
.90	2.49(.013)	.216(.160)	3.63(.062)	3.42(.166)
1.0	1251(10.4)	.234(.160)	2501(16.3)	2501(16.3)

As can be seen in Table 29, the general pattern for all descriptive statistics that the mean and variability increases as c goes from 0 to 1. These values in the table reflect the range of the random component of the chain simulator. In general, the trend seen in this table reflects the trend seen in all similar descriptive statistic tables for the cases where there is complete imbalance.

Table 30 presents the descriptive statistics for all levels of c when there is partial imbalance and the random component is equal to .5. It is presented below.

Table 30: Descriptive statistics for simulated chains when d=.75 (Partial Imbalance) and range = .5

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.125(.001)	-.125(.00004)	.375(.00004)	.500(.0001)
.25	.167(.002)	-.152(.005)	.486(.004)	.638(.010)
.50	.250(.003)	-.188(.011)	.686(.011)	.874(.015)
.75	.499(.005)	-.197(.034)	1.20(.043)	1.39(.058)
.90	1.25(.013)	.023(.093)	2.41(.107)	2.39(.140)
1.0	625.3(8.34)	.125(.175)	1249(14.9)	1249(14.9)

For Table 30, the pattern is slightly different than for the previous table. It is still the case that as c increases the average mean, maximum, and range increases. It is also still the case that variability increases as c increases. However, when d represents partial imbalance (positive) the average minimum first decreases and then increases over the range of c. The descriptive statistics also reflect the influence of the range of the random component of the chain simulator. This pattern is similar to the trend seen in each other table of descriptive statistics for the case where partial imbalance is present in the chain simulator.

Table 31 presents the descriptive statistics for all levels of c when there is balance and the random component is equal to .5. It is presented below.

Table 31: Descriptive statistics for simulated chains when d=.5 (Balance) and range = .5

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	.0001(.002)	-.250(.0001)	.250(.0001)	.500(.0001)
.25	.0005(.002)	-.318(.004)	.318(.003)	.636(.005)
.50	.0003(.002)	-.437(.010)	.436(.012)	.873(.014)
.75	-.0001(.006)	-.706(.039)	.684(.030)	1.39(.054)
.90	.004(.017)	-1.16(.107)	1.15(.088)	2.31(.147)
1.0	1.72(8.29)	-9.35(7.74)	12.39(8.1)	21.74(5.66)

As can be seen in Table 31, as c increases the average minimum gets smaller and more variable, the average maximum gets larger and more variable, and the average range gets larger and more variable. The average mean has no discernible pattern, but its variability increases as c increases. The descriptive statistics in the table reflect the range of the random component of the chain simulator. This pattern is similar to the trend seen in each other table of descriptive statistics for the case where balance is present in the chain simulator.

Table 32 presents the descriptive statistics for all levels of c when there is complete imbalance and the random component is equal to 5. It is presented below.

Table 32: Descriptive statistics for simulated chains when d=1 (Complete Imbalance) and range = 5

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	2.50(.016)	.0003(.0004)	5.00(.0004)	5.00(.001)
.25	3.33(.015)	.146(.044)	6.54(.037)	6.39(.058)
.50	4.99(.028)	.553(.188)	9.41(.125)	8.85(.225)
.75	9.99(.046)	1.87(.963)	16.9(.288)	15.0(1.07)
.90	25.0(.167)	2.40(1.23)	36.7(.932)	34.2(1.59)
1.0	12488(63.2)	2.17(1.34)	24966(103.4)	24964(103.5)

As can be seen in Table 32, the general pattern for all descriptive statistics that the mean and variability increases as c goes from 0 to 1. These values in the table reflect the range of the random component of the chain simulator. In general, the trend seen in this table reflects the trend seen in all similar descriptive statistic tables for the cases where there is complete imbalance.

Table 33 presents the descriptive statistics for all levels of c when there is partial imbalance and the random component is equal to 5. It is presented below.

Table 33: Descriptive statistics for simulated chains when d=.75 (Partial Imbalance) and range = 5

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	1.25(.017)	-1.25(.0002)	3.75(.001)	5.00(.001)
.25	1.67(.019)	-1.52(.032)	4.86(.043)	6.38(.052)
.50	2.51(.024)	-1.84(.118)	6.89(.127)	8.73(.175)
.75	4.98(.041)	-1.99(.416)	11.9(.401)	13.9(.493)
.90	12.45(.128)	.381(.933)	23.7(.753)	23.3(1.28)
1.0	6223(64.5)	.947(1.35)	12455(140.4)	12454(140.3)

For Table 33, the pattern is slightly different than for the previous table. It is still the case that as c increases the average mean, maximum, and range increases. It is also still the case that variability increases as c increases. However, when d represents partial imbalance (positive) the average minimum first decreases and then increases over the range of c. The descriptive statistics also reflect the influence of the range of the random component of the chain simulator. This pattern is similar to the trend seen in each other table of descriptive statistics for the case where partial imbalance is present in the chain simulator.

Table 34 presents the descriptive statistics for all levels of c when there is balance and the random component is equal to 5. It is presented below.

Table 34: Descriptive statistics for simulated chains when d=.5 (Balance) and range = 5

c	Average Mean	Average Minimum	Average Maximum	Average Range
0.0	-.001(.017)	-2.50(.0004)	2.50(.0004)	5.00(.001)
.25	-.011(.022)	-3.19(.041)	3.19(.043)	6.38(.059)
.50	-.003(.030)	-4.40(.102)	4.39(.090)	8.79(.161)
.75	.010(.061)	-6.88(.347)	6.97(.252)	13.8(.430)
.90	.039(.127)	-11.7(.543)	11.4(.585)	23.1(.791)
1.0	22.9(85.1)	-95.5(79.8)	148.2(105.1)	243.7(84.2)

As can be seen in Table 34, as c increases the average minimum gets smaller and more variable, the average maximum gets larger and more variable, and the average range gets larger and more variable. The average mean has no discernible pattern, but its variability increases as c increases. The descriptive statistics in the table reflect the range of the random component of the chain simulator.

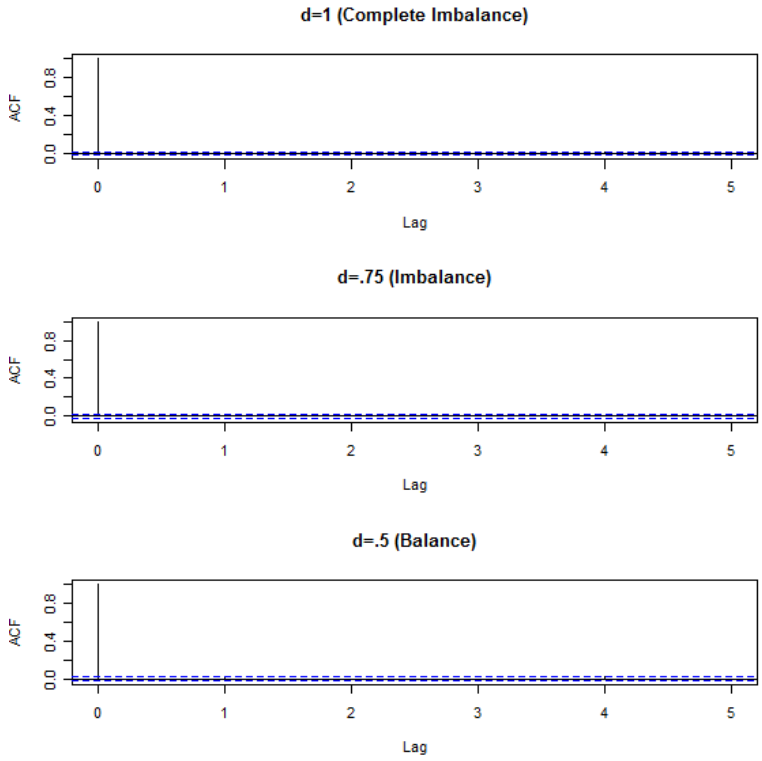
Overall, the range of the random component has an influence on the values present in the chains. For example, as the range of the random component increases, the values observed in the chains become more extreme. It was expected that the range of the random component of the chain simulator would have this effect. Also, the degree of autocorrelation influences the values of the chain elements. Generally speaking, as the degree of autocorrelation present among elements in the chain increases, the variability of the descriptive statistics increases. Finally, the general trend for each level of balance

holds across all ranges of the random component of the chain simulator, indicating that the range of the random component doesn't necessarily have an influence on the chain other than to set bounds for how far the values will wander for a given amount of iterations.

Now, the autocorrelation plots will be presented for the levels of range new to Simulation study 4 (range = .1, .5 and 5 respectively). Again, the autocorrelation plots for the condition where the range is equal to one are omitted because they are virtually identical to the autocorrelation plots already presented for this condition in other simulation studies. Each autocorrelation plot will represent the three levels of the balance factor, d, and a plot for each level of c will be represented in an individual figure.

Figure 31 presents the autocorrelation plots for all levels of d when c is equal to 0 and range is equal to .1. It is presented below.

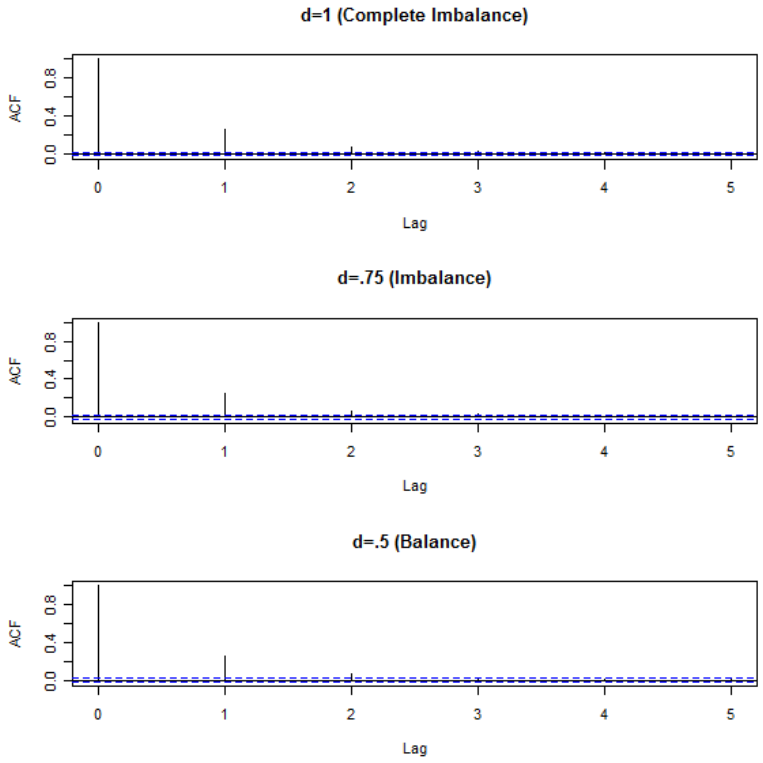
Figure 31: Autocorrelation plots for all levels of d when c is equal to 0 and range = to .1



In Figure 31, it can again be seen that when c is equal to zero, the chains are an i.i.d. sequence of elements. This pattern is similar to all cases where c is equal to 0.

Figure 32 presents the autocorrelation plots for all levels of d when c is equal to .25 and range is equal to .1. It is presented below.

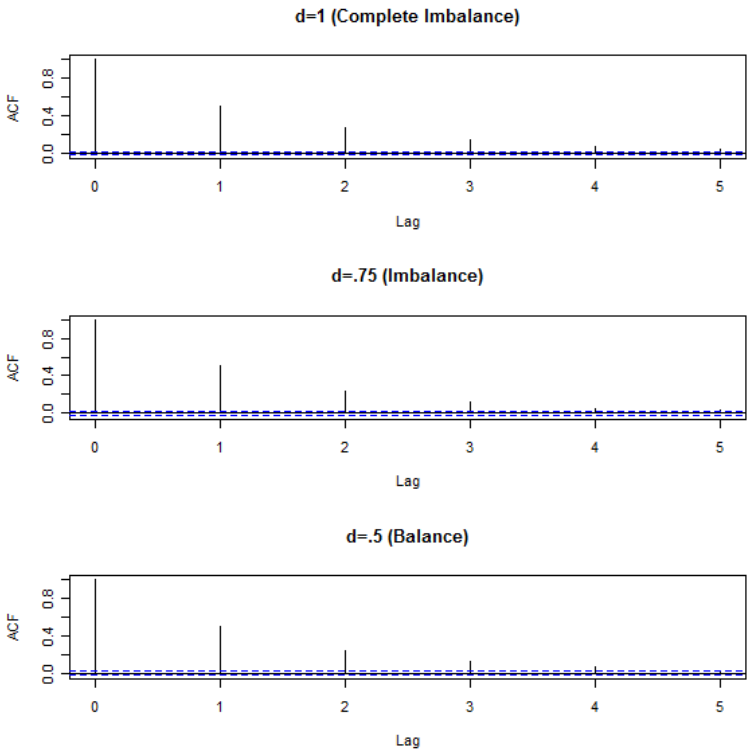
Figure 32: Autocorrelation plots for all levels of d when c is equal to .25 and range = to .1



In Figure 32, it can again be seen that when c is equal to .25, there is a small amount of autocorrelation that exists among elements in the chain. For the chains in this condition, the autocorrelation is no longer observable by a lag of three or four. This pattern is similar to all the cases where the value of c is equal to .25.

Figure 33 presents the autocorrelation plots for all levels of d when c is equal to .5 and range is equal to .1. It is presented below.

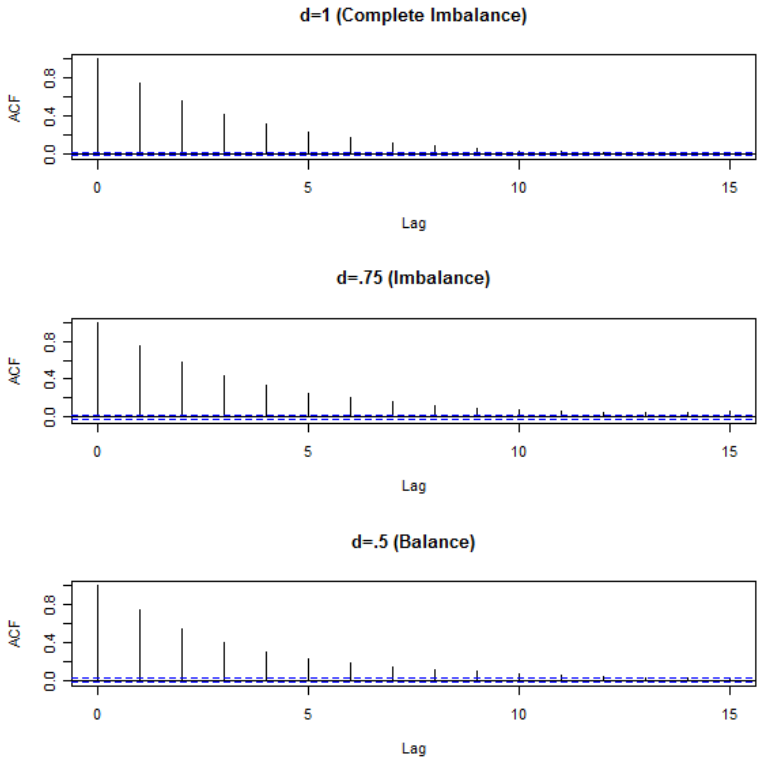
Figure 33: Autocorrelation plots for all levels of d when c is equal to .5 and range = to .1



In Figure 33, there is again an increase in the degree of autocorrelation present in the chains when c is set equal to .5. In the chains simulated for this condition it can be seen that the autocorrelation tends towards zero by a lag of roughly five. This pattern is similar for all conditions where c is equal to .5.

Figure 34 presents the autocorrelation plots for all levels of d when c is equal to .75 and range is equal to .1. It is presented below.

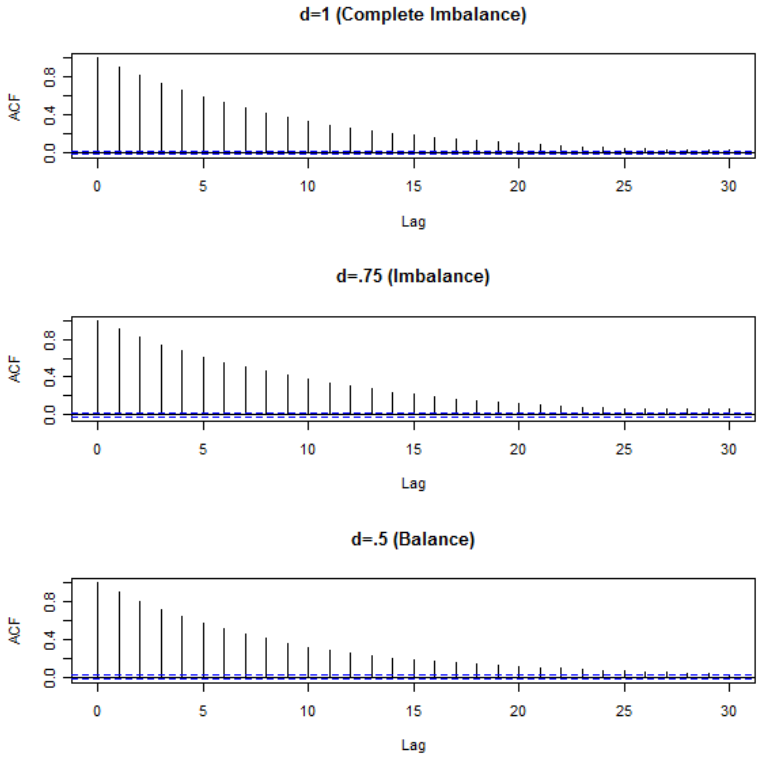
Figure 34: Autocorrelation plots for all levels of d when c is equal to .75 and range = to .1



In Figure 34, it can again be seen that when c is equal to .75 there is dependency among elements for a lag of up to 15. This pattern is similar to all other case where c is equal to .75.

Figure 35 presents the autocorrelation plots for all levels of d when c is equal to .9 and range is equal to .1. It is presented below.

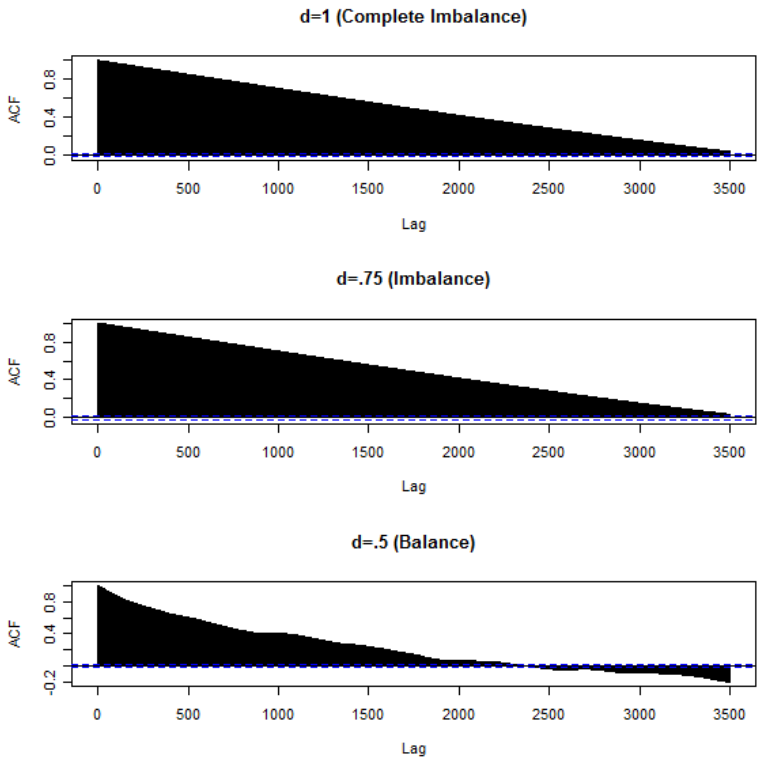
Figure 35: Autocorrelation plots for all levels of d when c is equal to .9 and range = to .1



In Figure 35, it can again be seen that when c is equal to .9 there is a relationship among elements separated by a lag of up to roughly 25 to 30 elements. This pattern is similar to all other cases where c is equal to .9.

Figure 36 presents the autocorrelation plots for all levels of d when c is equal to 1 and range is equal to .1. It is presented below.

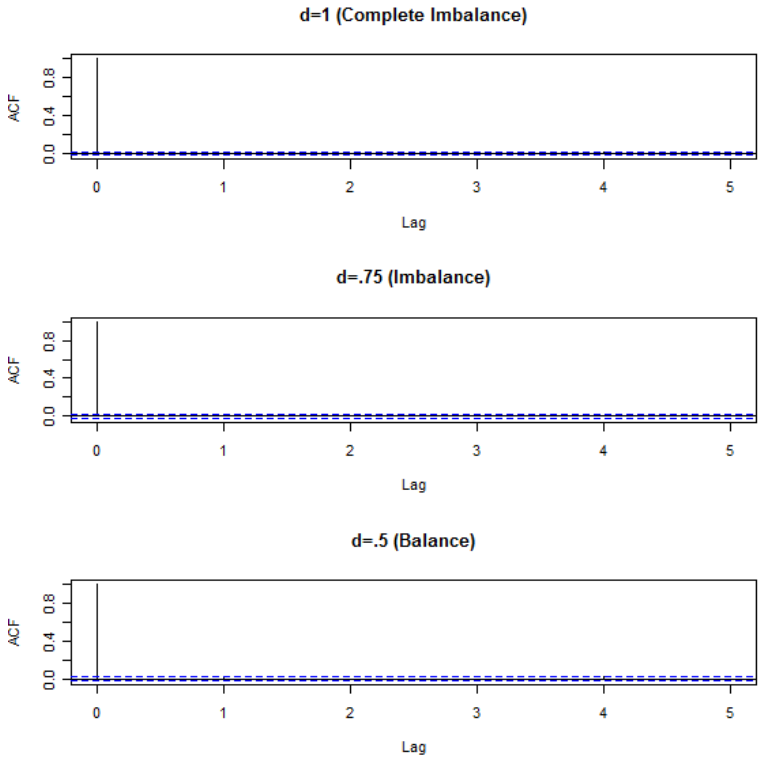
Figure 36: Autocorrelation plots for all levels of d when c is equal to 1 and range = to .1



In Figure 36, it can again be seen that when c is equal to 1, there is a strong degree of autocorrelation present among elements. When there is any type of imbalance present, this autocorrelation persists to a lag of up to 3500 elements. When there is balance present (and the chain is equally likely to move up or down, rather than in one direction only), the autocorrelation exists among elements for a smaller lag. The lag over which the autocorrelation persists in the case of balance is roughly 2000 to 2500.

Figure 37 presents the autocorrelation plots for all levels of d when c is equal to 0 and range is equal to .5. It is presented below.

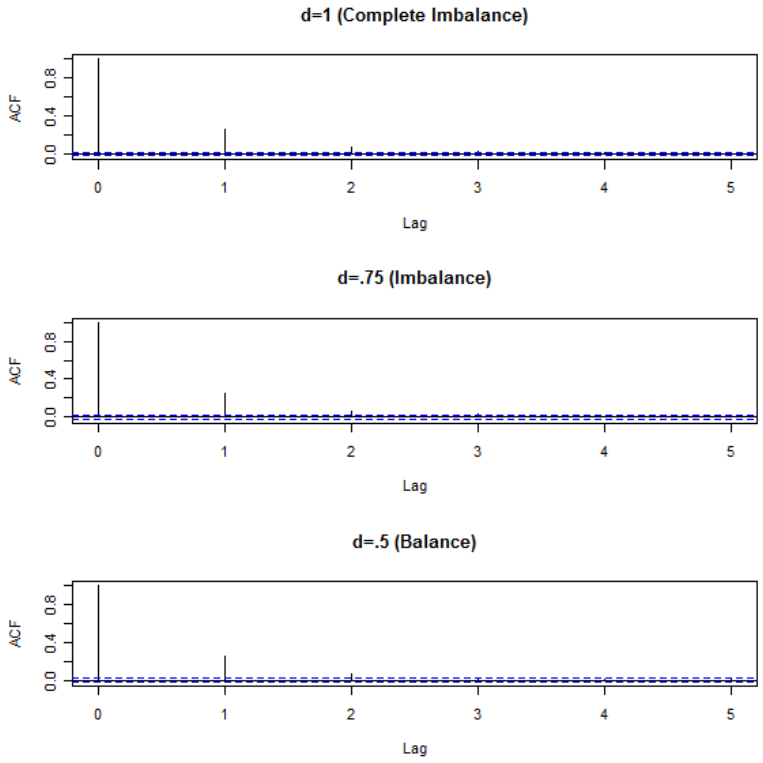
Figure 37: Autocorrelation plots for all levels of d when c is equal to 0 and range = to .5



In Figure 37, it can again be seen that when c is equal to zero, the chains are an i.i.d. sequence of elements. This pattern is similar to all cases where c is equal to 0.

Figure 38 presents the autocorrelation plots for all levels of d when c is equal to .25 and range is equal to .5. It is presented below.

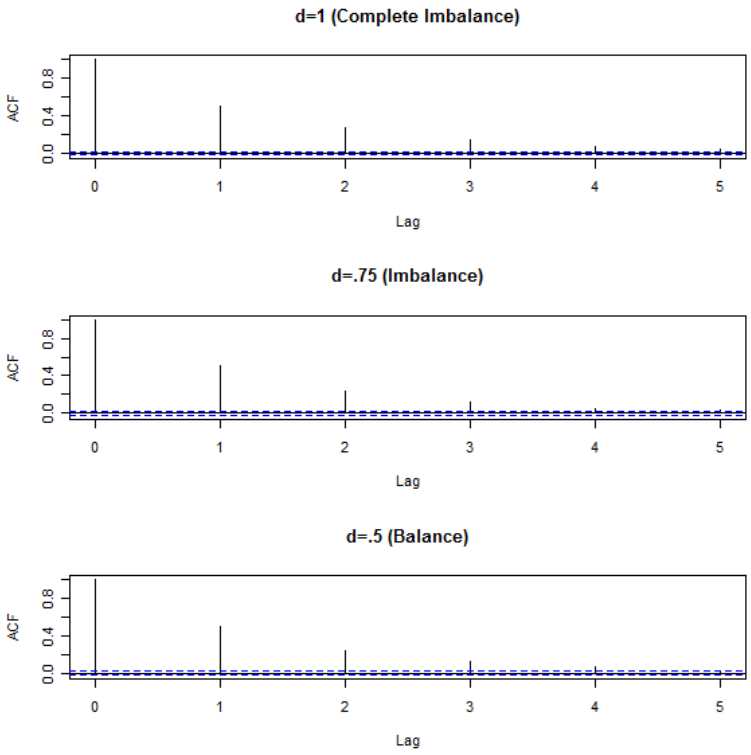
Figure 38: Autocorrelation plots for all levels of d when c is equal to .25 and range = to .5



In Figure 38, it can again be seen that when c is equal to .25, there is a small amount of autocorrelation that exists among elements in the chain. For the chains in this condition, the autocorrelation is no longer observable by a lag of three or four. This pattern is similar to all the cases where the value of c is equal to .25.

Figure 39 presents the autocorrelation plots for all levels of d when c is equal to .5 and range is equal to .5. It is presented below.

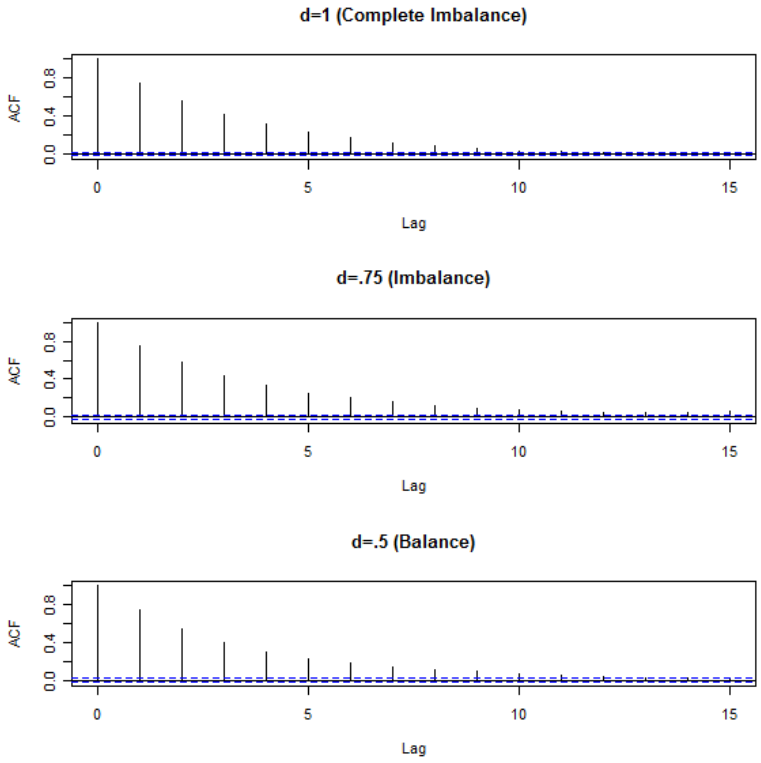
Figure 39: Autocorrelation plots for all levels of d when c is equal to .5 and range = to .5



In Figure 39, there is again an increase in the degree of autocorrelation present in the chains when c is set equal to .5. In the chains simulated for this condition it can be seen that the autocorrelation tends towards zero by a lag of roughly five. This pattern is similar for all conditions where c is equal to .5.

Figure 40 presents the autocorrelation plots for all levels of d when c is equal to .75 and range is equal to .5. It is presented below.

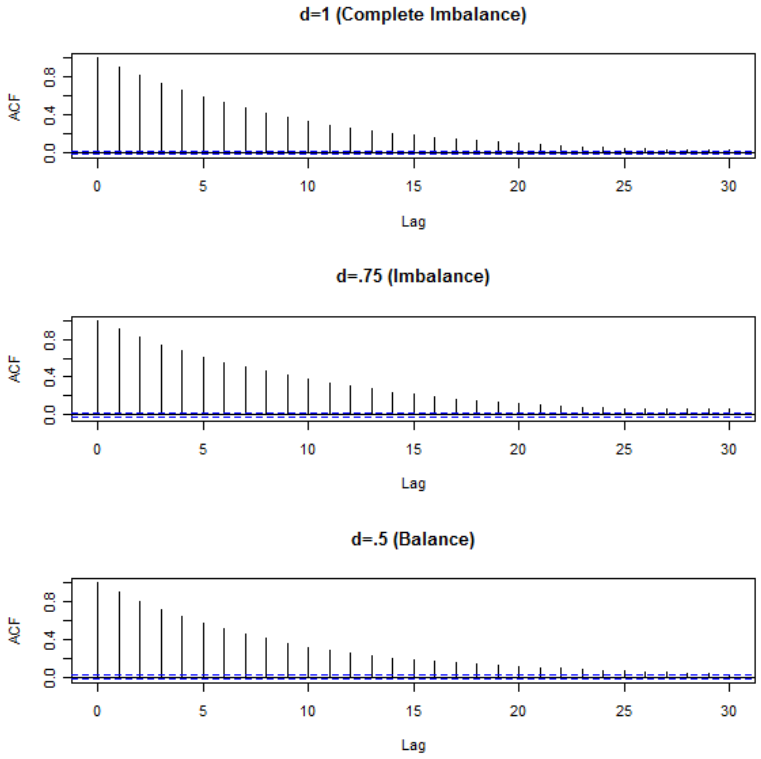
Figure 40: Autocorrelation plots for all levels of d when c is equal to .75 and range = to .5



In Figure 40, it can again be seen that when c is equal to .75 there is dependency among elements for a lag of up to 15. This pattern is similar to all other case where c is equal to .75.

Figure 41 presents the autocorrelation plots for all levels of d when c is equal to .9 and range is equal to .5. It is presented below.

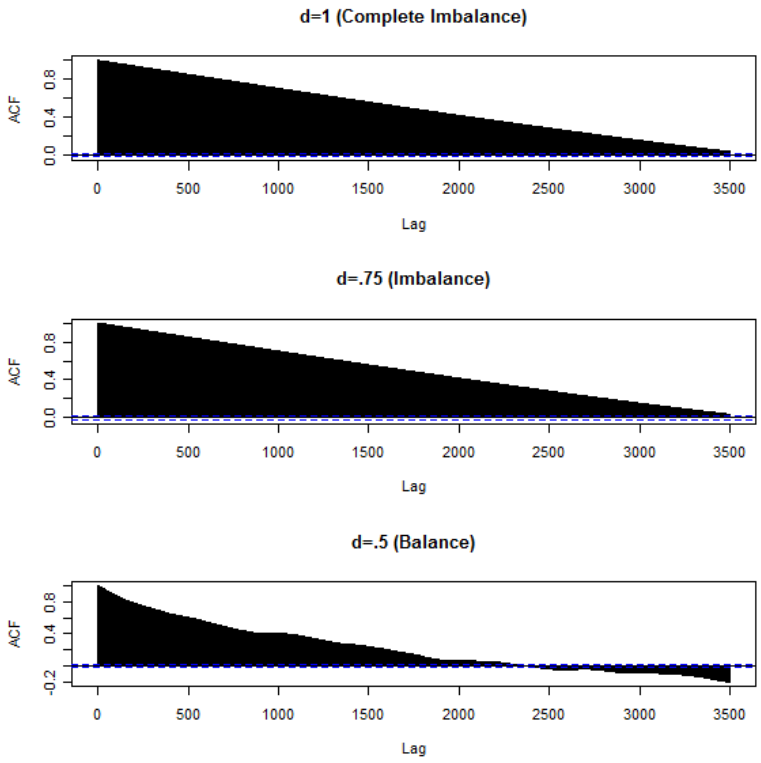
Figure 41: Autocorrelation plots for all levels of d when c is equal to .9 and range = to .5



In Figure 41, it can again be seen that when c is equal to .9 there is a relationship among elements separated by a lag of up to roughly 25 to 30 elements. This pattern is similar to all other cases where c is equal to .9.

Figure 42 presents the autocorrelation plots for all levels of d when c is equal to 1 and range is equal to .5. It is presented below.

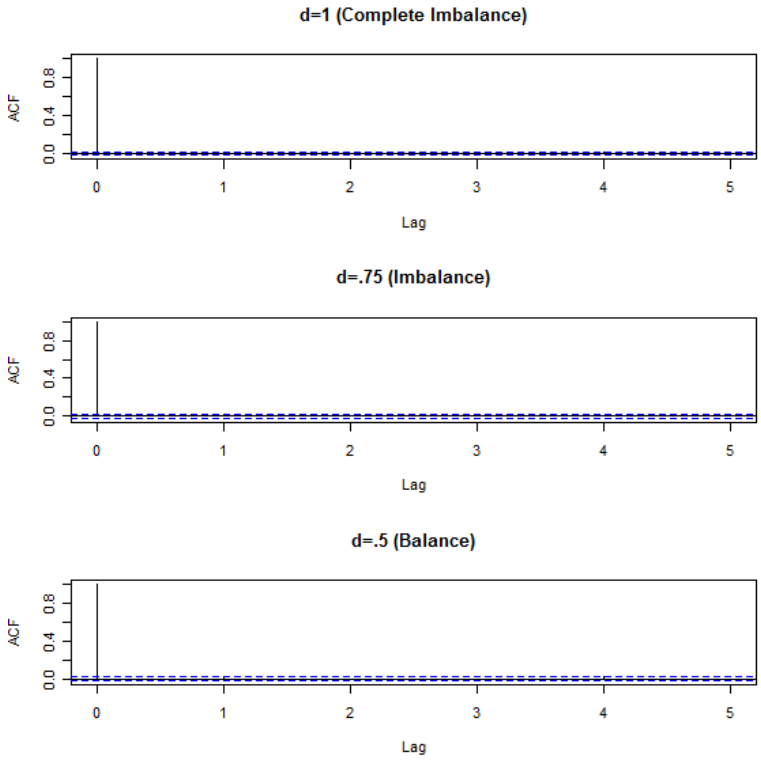
Figure 42: Autocorrelation plots for all levels of d when c is equal to 1 and range = to .5



In Figure 42, it can again be seen that when c is equal to 1, there is a strong degree of autocorrelation present among elements. When there is any type of imbalance present, this autocorrelation persists to a lag of up to 3500 elements. When there is balance present (and the chain is equally likely to move up or down, rather than in one direction only), the autocorrelation exists among elements for a smaller lag. The lag over which the autocorrelation persists in the case of balance is roughly 2000 to 2500.

Figure 43 presents the autocorrelation plots for all levels of d when c is equal to 0 and range is equal to 5. It is presented below.

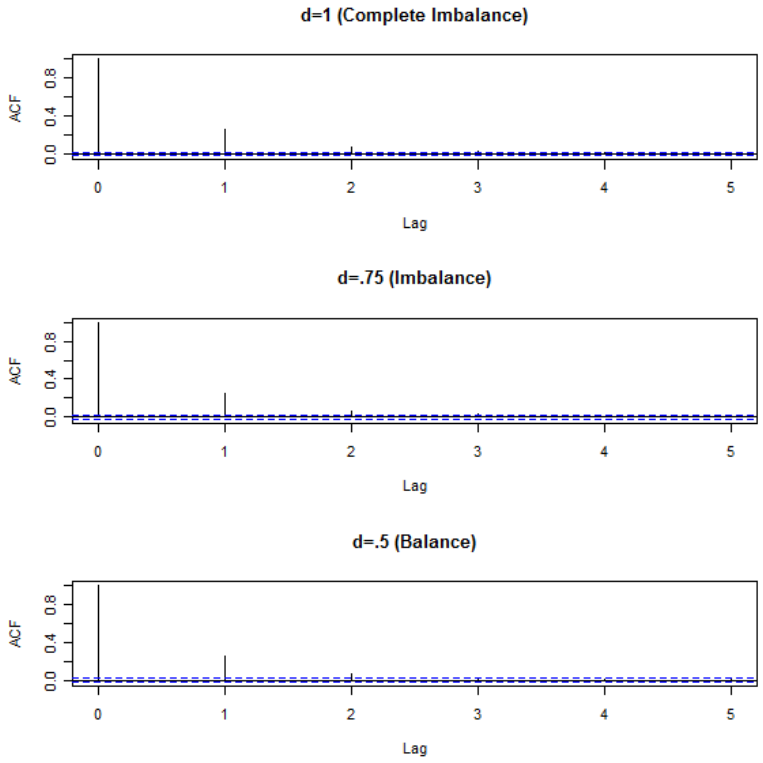
Figure 43: Autocorrelation plots for all levels of d when c is equal to 0 and range = to 5



In Figure 43, it can again be seen that when c is equal to zero, the chains are an i.i.d. sequence of elements. This pattern is similar to all cases where c is equal to 0.

Figure 44 presents the autocorrelation plots for all levels of d when c is equal to .25 and range is equal to 5. It is presented below.

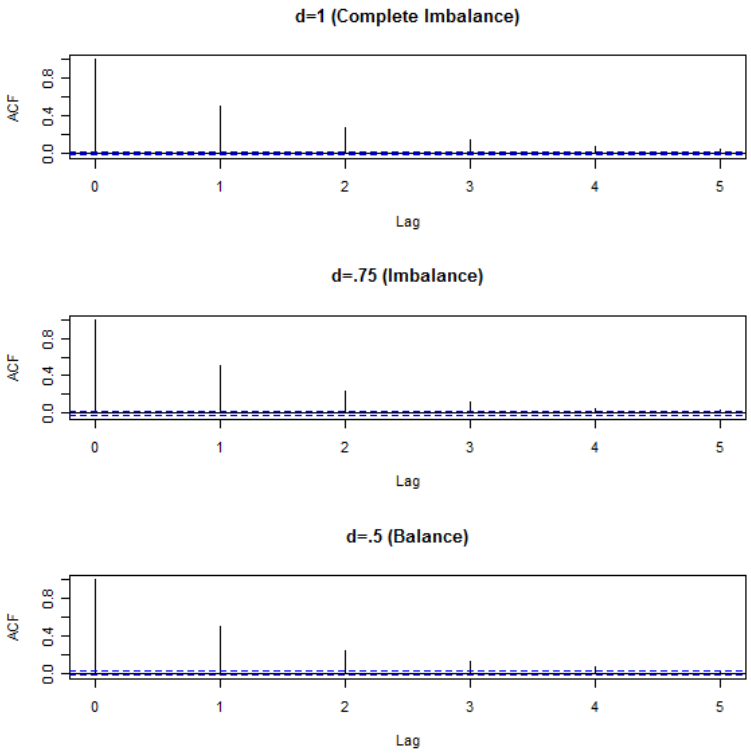
Figure 44: Autocorrelation plots for all levels of d when c is equal to .25 and range = to 5



In Figure 44, it can again be seen that when c is equal to .25, there is a small amount of autocorrelation that exists among elements in the chain. For the chains in this condition, the autocorrelation is no longer observable by a lag of three or four. This pattern is similar to all the cases where the value of c is equal to .25.

Figure 45 presents the autocorrelation plots for all levels of d when c is equal to .5 and range is equal to 5. It is presented below.

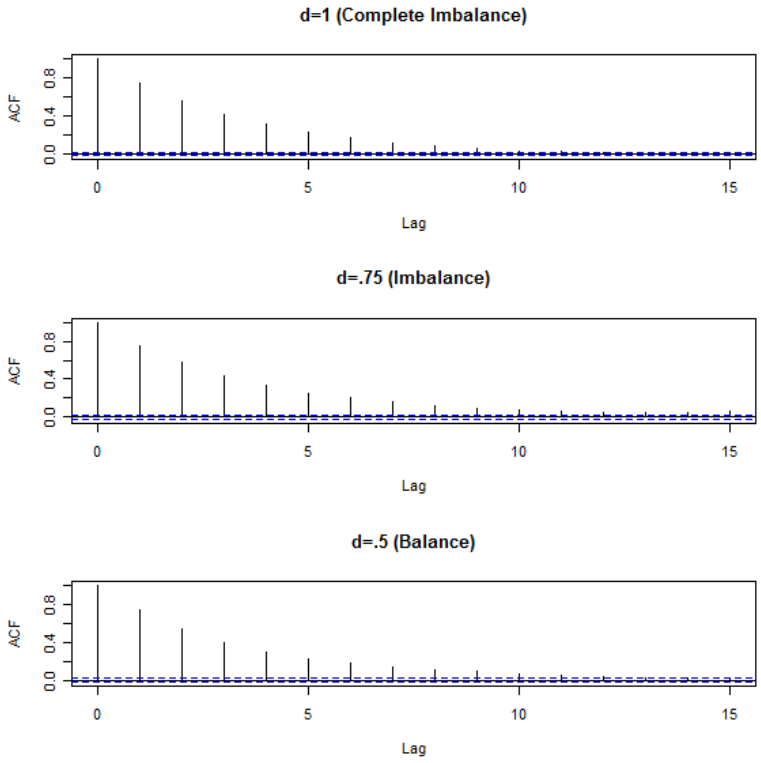
Figure 45: Autocorrelation plots for all levels of d when c is equal to .5 and range = to 5



In Figure 45, there is again an increase in the degree of autocorrelation present in the chains when c is set equal to .5. In the chains simulated for this condition it can be seen that the autocorrelation tends towards zero by a lag of roughly five. This pattern is similar for all conditions where c is equal to .5.

Figure 46 presents the autocorrelation plots for all levels of d when c is equal to .75 and range is equal to 5. It is presented below.

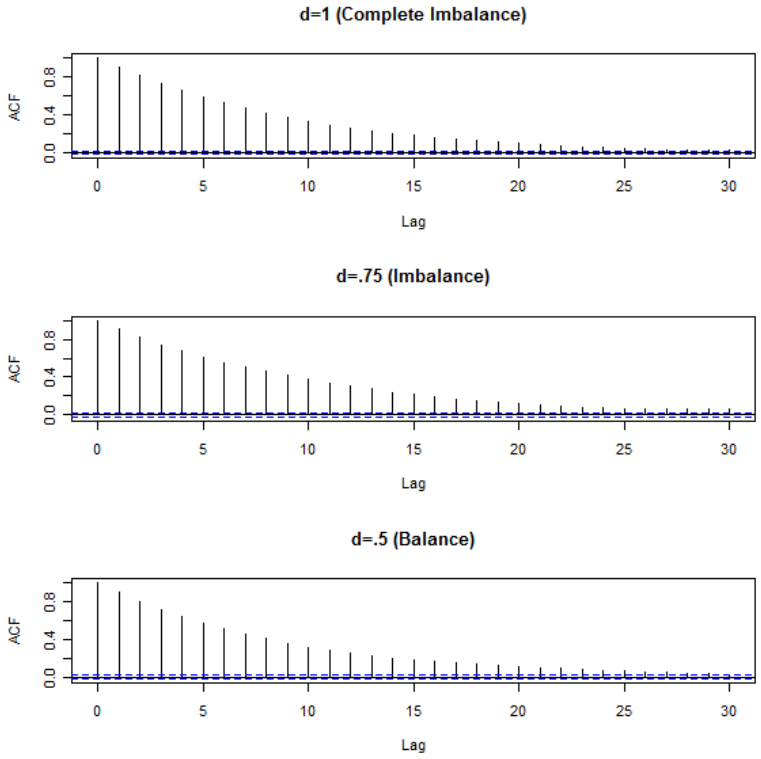
Figure 46: Autocorrelation plots for all levels of d when c is equal to .75 and range = to 5



In Figure 46, it can again be seen that when c is equal to .75 there is dependency among elements for a lag of up to 15. This pattern is similar to all other case where c is equal to .75.

Figure 47 presents the autocorrelation plots for all levels of d when c is equal to .9 and range is equal to 5. It is presented below.

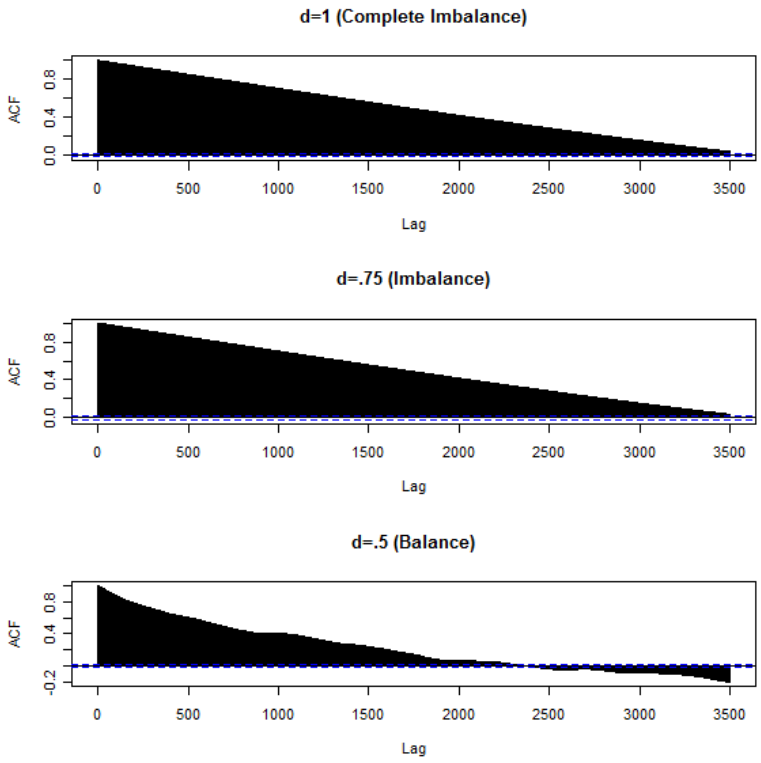
Figure 47: Autocorrelation plots for all levels of d when c is equal to .9 and range = to 5



In Figure 47, it can again be seen that when c is equal to .9 there is a relationship among elements separated by a lag of up to roughly 25 to 30 elements. This pattern is similar to all other cases where c is equal to .9.

Figure 48 presents the autocorrelation plots for all levels of d when c is equal to 1 and range is equal to 5. It is presented below.

Figure 48: Autocorrelation plots for all levels of d when c is equal to 1 and range = to 5



In Figure 48, it can again be seen that when c is equal to 1, there is a strong degree of autocorrelation present among elements. When there is any type of imbalance present, this autocorrelation persists to a lag of up to 3500 elements. When there is balance present (and the chain is equally likely to move up or down, rather than in one direction only), the autocorrelation exists among elements for a smaller lag. The lag over which the autocorrelation persists in the case of balance is roughly 2000 to 2500.

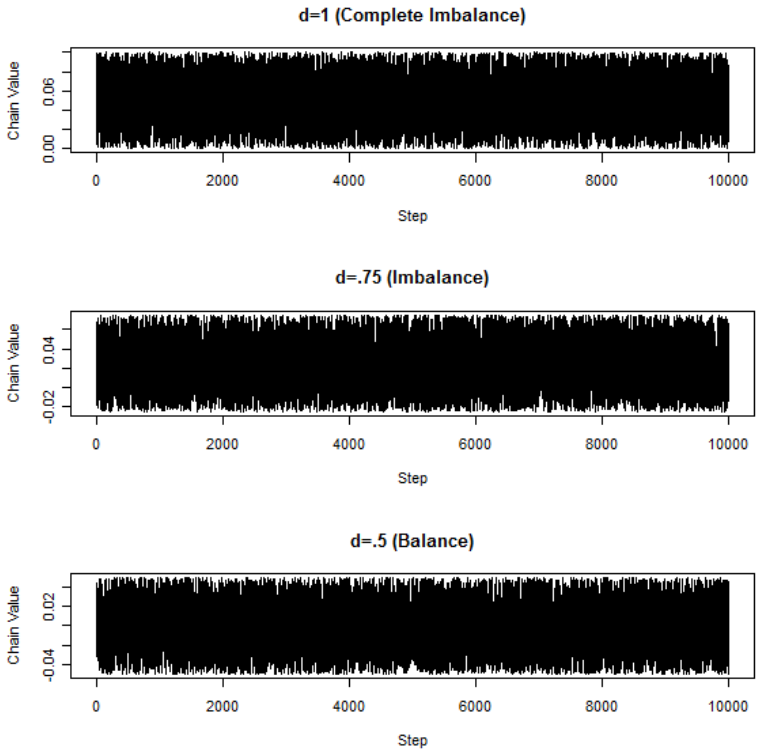
In general, the autocorrelation plots presented for the new conditions of simulation study 4 demonstrate the same patterns seen in the previous simulation studies. Taken together, the fact that the pattern of autocorrelations present across all

combinations of c and d is the same regardless of the range of the random component of the chain simulator is to be expected. Again, the range of the random component of the chain simulator affects the magnitude of movement we see between elements in a chain, but it does nothing to manipulate the relative rank orderings or associations among the elements in a chain.

The path plots for the conditions new to simulation study 4 will now be presented. Each plot will be provided and then briefly described. After all of the path plots representing the new conditions have been presented, a brief summary of the overall trends seen will be made and any similarity to previously presented findings will be addressed.

Figure 49 shows the path plots for all levels of d when c is equal to 0 and the range is equal to .1. It is presented below.

Figure 49: Path plots for all levels of d when c = 0 and range = .1



As can be seen in Figure 49, each of these chains traverses the space between the bounds of the respective distribution. As these are i.i.d. sequences, all values stay within the bounds as specified by the range of the random component of the chain simulator.

Figure 50 shows the path plots for all levels of d when c is equal to .25 and the range is equal to .1. It is presented below.

Figure 50: Path plots for all levels of d when c = .25 and range = .1

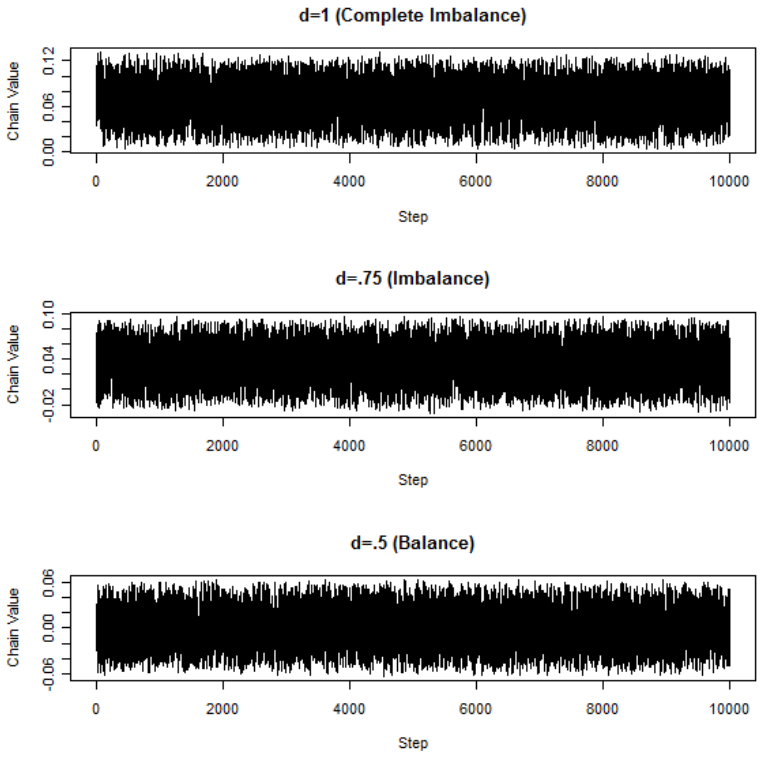


Figure 50 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of c can be seen in the expanded range that the values of the chain elements take on.

Figure 51 shows the path plots for all levels of d when c is equal to .5 and the range is equal to .1. It is presented below.

Figure 51: Path plots for all levels of d when c = .5 and range = .1

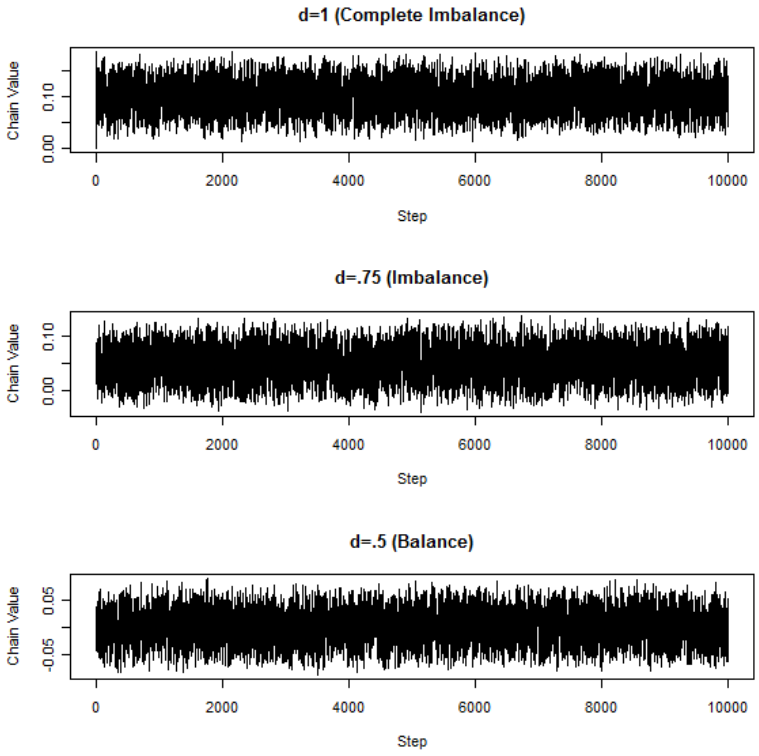


Figure 51 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of c can be seen in the expanded range that the values of the chain elements take on.

Figure 52 shows the path plots for all levels of d when c is equal to .75 and the range is equal to .1. It is presented below.

Figure 52: Path plots for all levels of d when c = .75 and range = .1

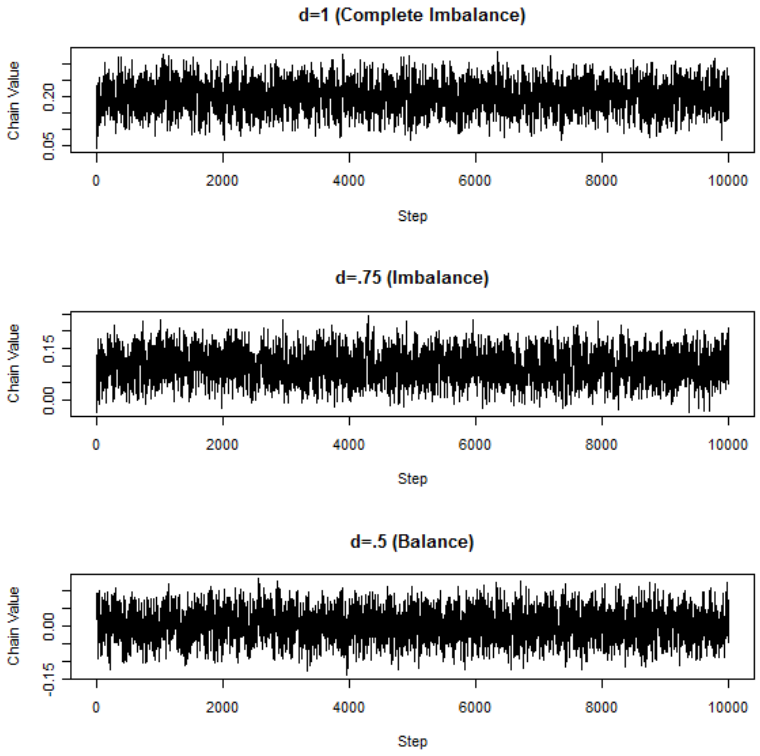


Figure 52 shows the same expansion of range that was seen in the descriptive statistics. The influence of c can be seen in the expanded range that the chain elements take on.

Figure 53 shows the path plots for all levels of d when c is equal to .9 and the range is equal to .1. It is presented below.

Figure 53: Path plots for all levels of d when c = .9 and range = .1

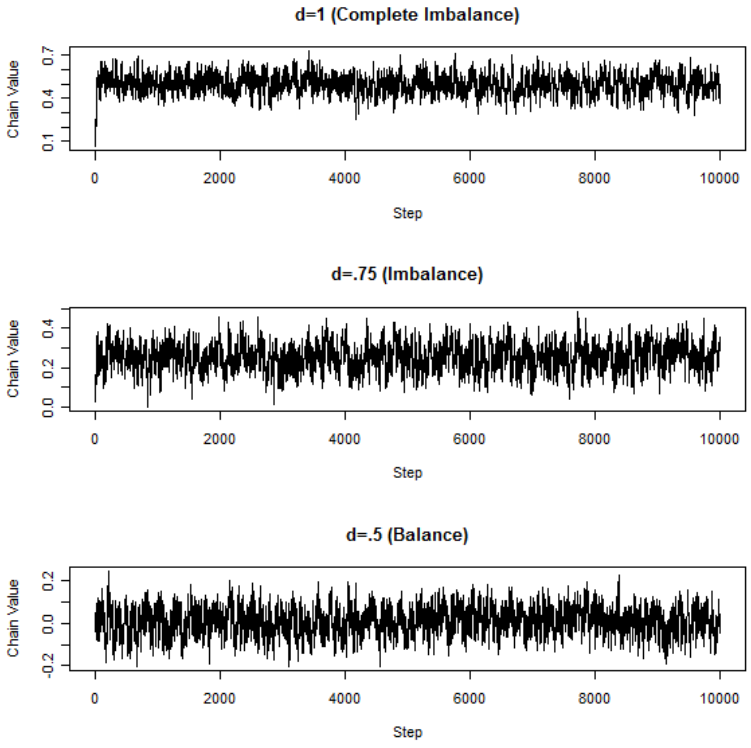


Figure 53 shows the same expansion of range that was seen in the descriptive statistics. The influence of c can be seen in the expanded range that the chain elements take on.

Figure 54 shows the path plots for all levels of d when c is equal to 1 and the range is equal to .1. It is presented below.

Figure 54: Path plots for all levels of d when c = 1 and range = .1

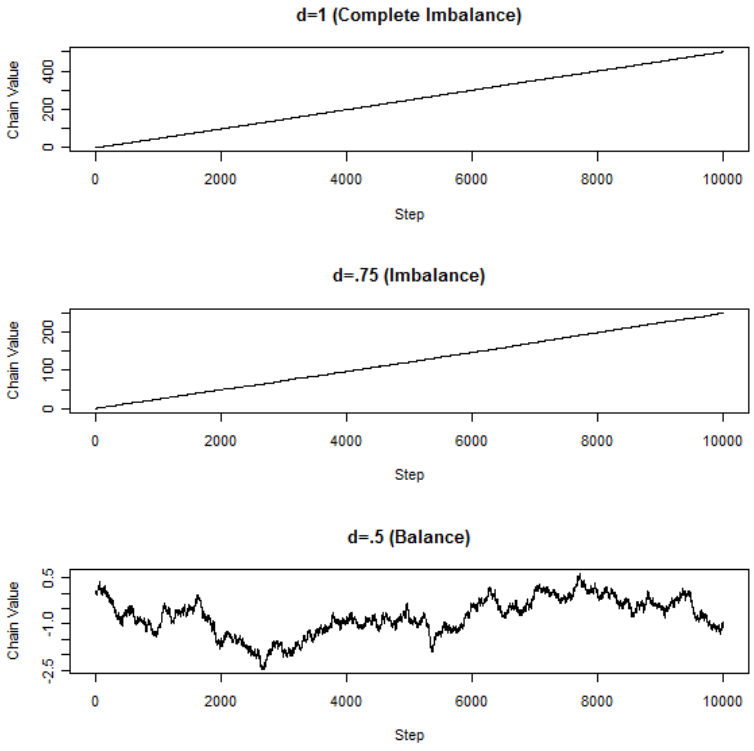
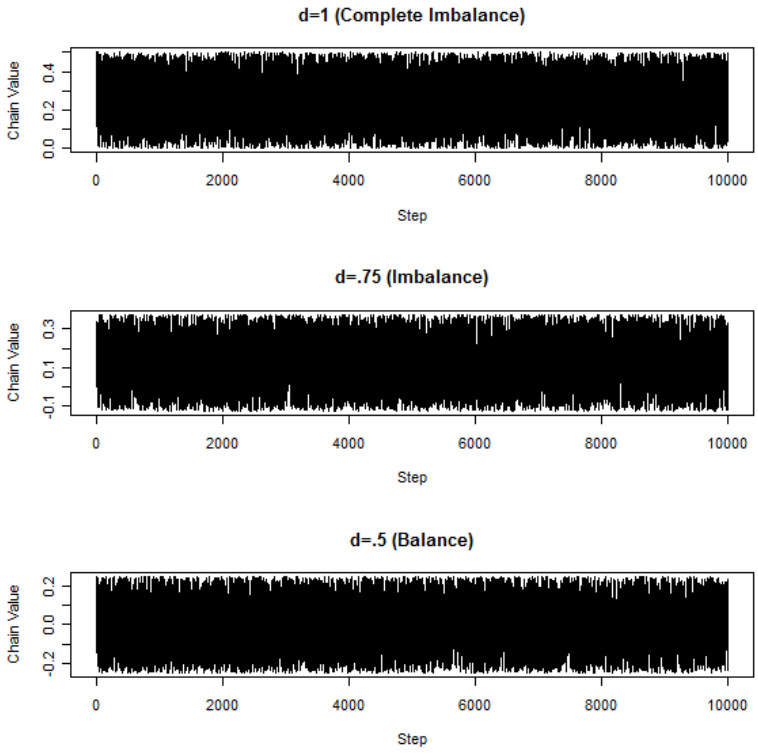


Figure 54 shows the behavior of the chain elements that was seen in the descriptive statistics for this particular set of conditions. The influence of d and c can be seen in the behavior of the chain. Specifically, because c is equal to 1, the complete imbalance condition is a strictly non-decreasing sequence of values. When there is partial imbalance present and c is equal to 1, the behavior of the chain is consistent with what would be expected. Specifically, each new element is set equal to the previous plus a random component that was twice as likely to be positive as it is to be negative. Thus the value that chain elements take on is more likely to increase rather than decrease over the length of the chain. However, chains produced in this condition are not strictly non-

decreasing. When there is balance present and c is equal to 1, it is equally likely that each new element will be greater than or less than the previous element. The result is a sequence of values that randomly increases or decreases over the length of the chain with equal frequency.

Figure 55 shows the path plots for all levels of d when c is equal to 0 and the range is equal to .5. It is presented below.

Figure 55: Path plots for all levels of d when c = 0 and range = .5



As can be seen in Figure 55, each of these chains traverses the space between the bounds of the respective distribution. The chains generated for these conditions are i.i.d.

sequences because c is equal to 0. Therefore, all values in the sequence stay within the bounds for the particular set of conditions. These chains are in agreement with the descriptive statistics presented earlier for the same case.

Figure 56 shows the path plots for all levels of d when c is equal to .25 and the range is equal to .5. It is presented below.

Figure 56: Path plots for all levels of d when c = .25 and range = .5

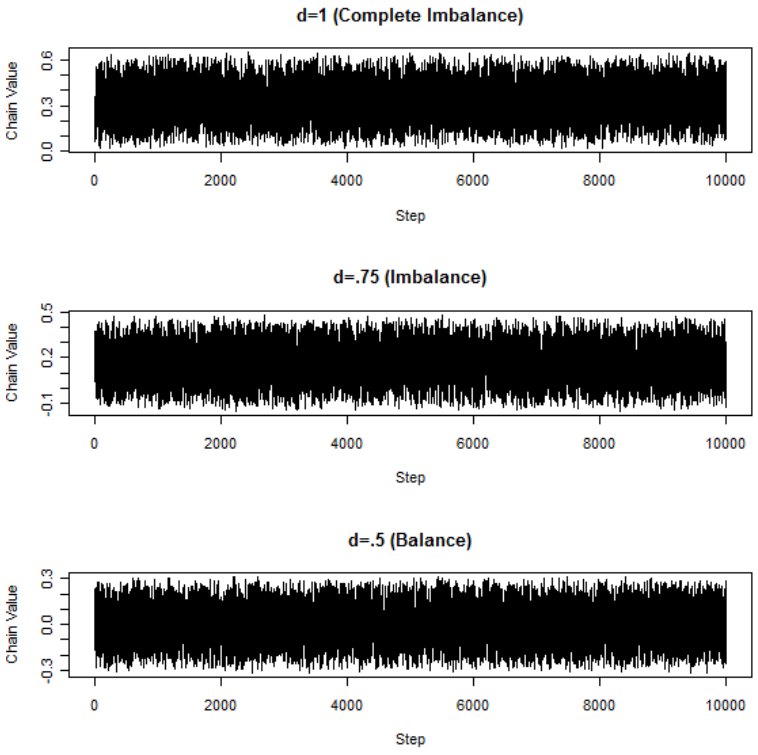


Figure 56 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range that the values of the chain elements take on, as well as the

particular values they take on, which corresponds to the direction of the imbalance.

Figure 57 shows the path plots for all levels of d when c is equal to .5 and the range is equal to .5. It is presented below.

Figure 57: Path plots for all levels of d when c = .5 and range = .5

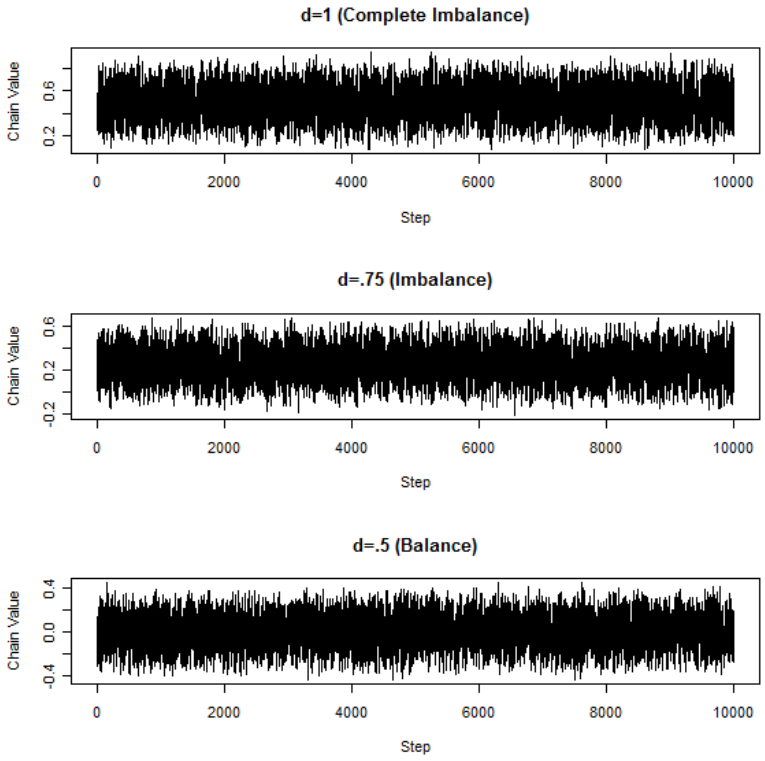


Figure 57 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range of the chain elements, and the particular values they take on, corresponding to the direction of the imbalance.

Figure 58 shows the path plots for all levels of d when c is equal to .75 and the range is equal to .5. It is presented below.

Figure 58: Path plots for all levels of d when c = .75 and range = .5

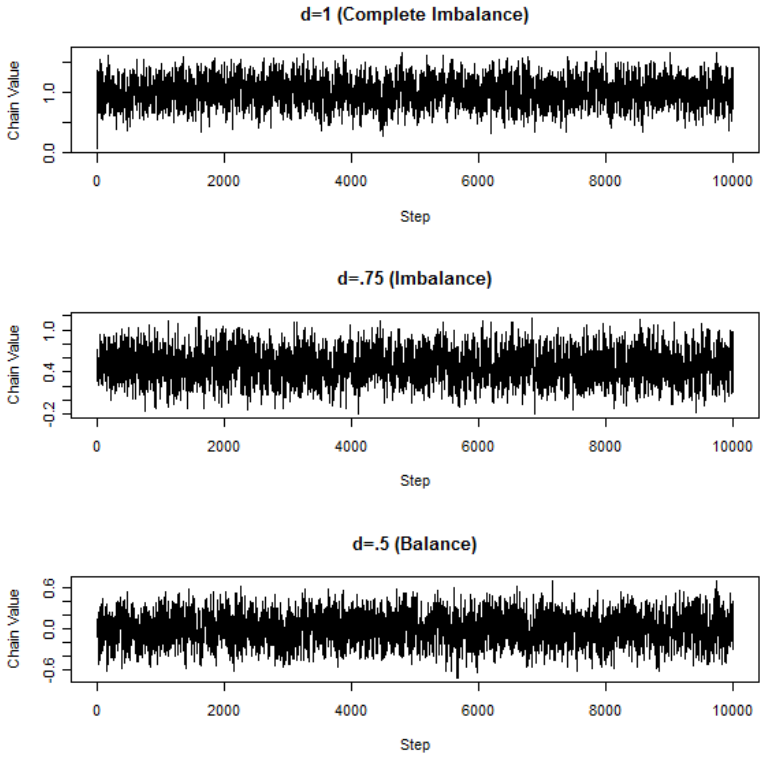


Figure 58 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range of the chain elements, and the particular values they take on, corresponding to the direction of the imbalance.

Figure 59 shows the path plots for all levels of d when c is equal to .9 and the range is equal to .5. It is presented below.

Figure 59: Path plots for all levels of d when c = .9 and range = .5

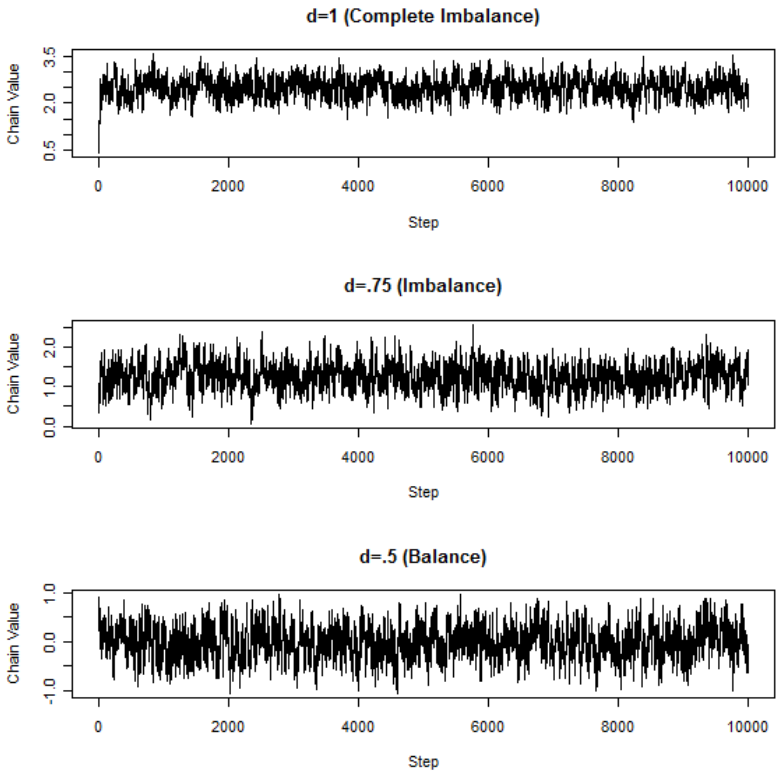


Figure 59 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range of the chain elements, and the particular values they take on, corresponding to the direction of the imbalance.

Figure 60 shows the path plots for all levels of d when c is equal to 1 and the range is equal to .5. It is presented below.

Figure 60: Path plots for all levels of d when c = 1 and range = .5

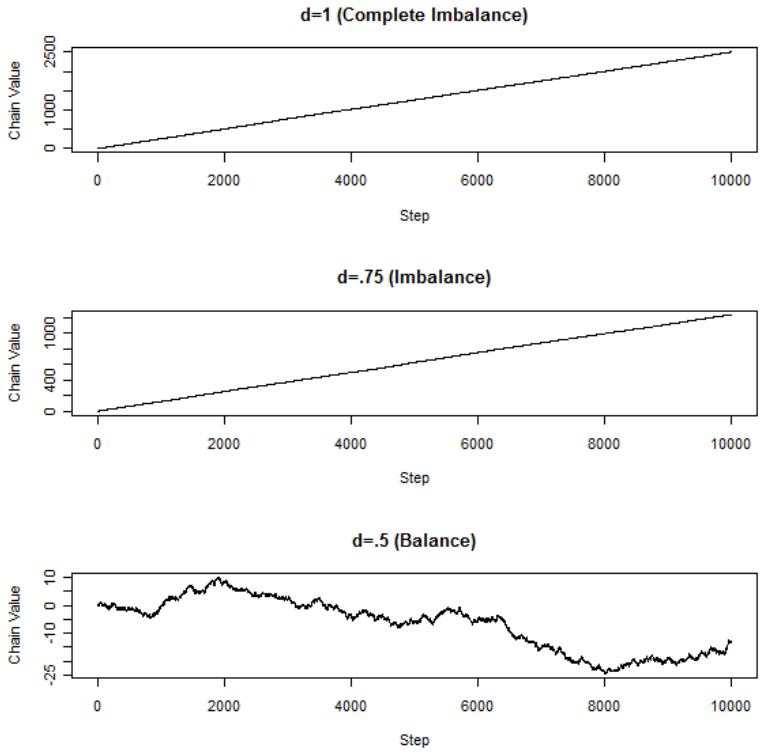
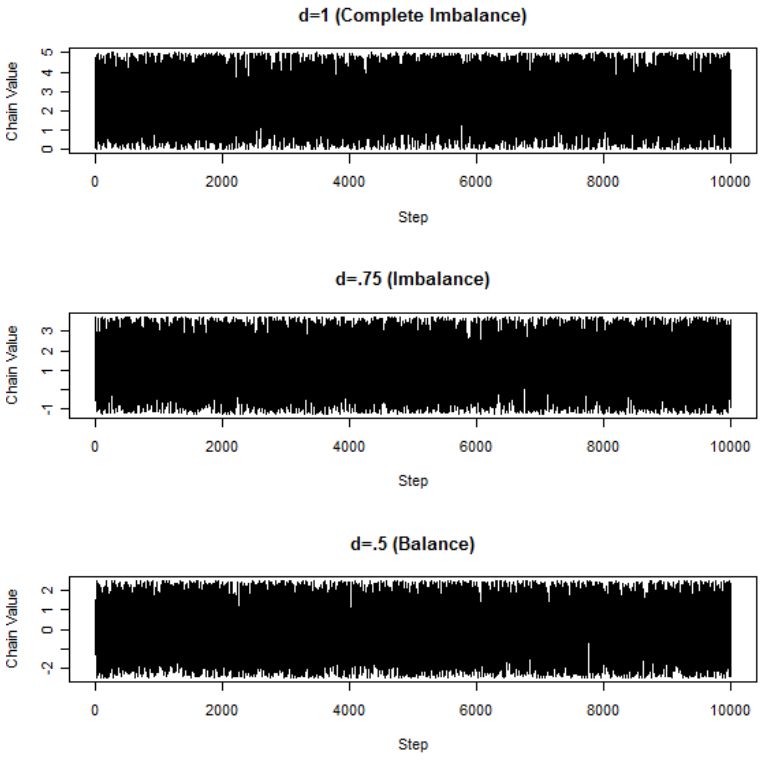


Figure 60 shows the behavior of the chain elements that was seen in the descriptive statistics for this particular set of conditions. The influence of d and c can be seen in the behavior of the chain. Specifically, because c is equal to 1, the complete imbalance condition is a strictly non-decreasing sequence of values. When there is partial imbalance present and c is equal to 1, the behavior of the chain is consistent with what would be expected. Specifically, each new element is set equal to the previous plus a random component that was twice as likely to be positive as it is to be negative. Thus the value that chain elements take on is more likely to increase rather than decrease over the length of the chain. However, chains produced in this condition are not strictly non-

decreasing. When there is balance present and c is equal to 1, it is equally likely that each new element will be greater than or less than the previous element. Thus, chains created for the case where c is equal to 1 and d is equal to .5 are sequences of elements that are equally likely to increase or decrease at each successive step.

Figure 61 shows the path plots for all levels of d when c is equal to 0 and the range is equal to 5. It is presented below.

Figure 61: Path plots for all levels of d when c = 0 and range = 5



As can be seen in Figure 61, each of these chains traverses the space between the bounds of the respective distribution. The chains generated for these conditions are i.i.d.

sequences because c is equal to 0. Therefore, all values in the sequence stay within the bounds for the particular set of conditions. These chains are in agreement with the descriptive statistics presented earlier for the same case.

Figure 62 shows the path plots for all levels of d when c is equal to .25 and the range is equal to 5. It is presented below.

Figure 62: Path plots for all levels of d when c = .25 and range = 5

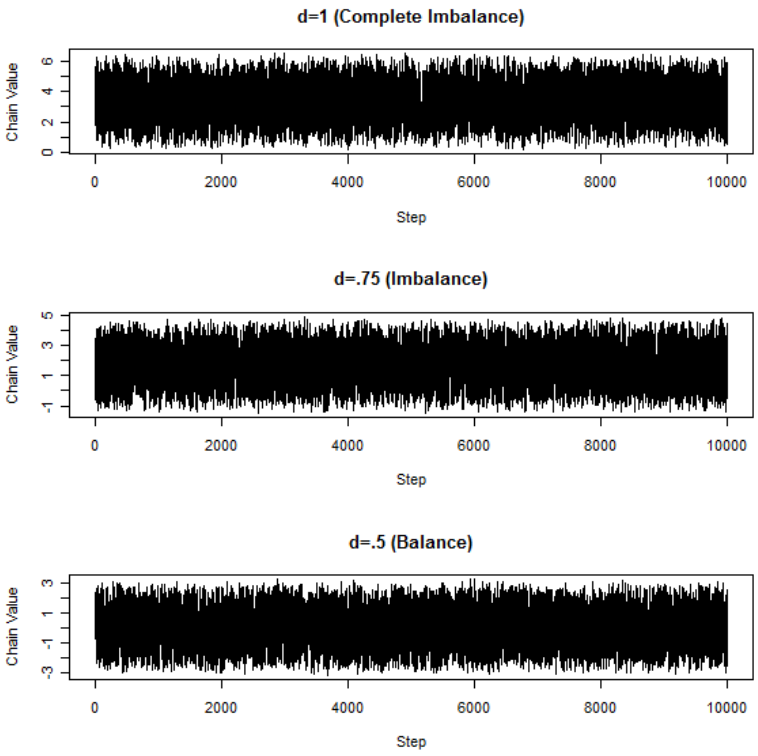


Figure 62 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range of the chain elements, and the particular values they take on, corresponding to the direction of the imbalance.

Figure 63 shows the path plots for all levels of d when c is equal to .5 and the range is equal to 5. It is presented below.

Figure 63: Path plots for all levels of d when c = .5 and range = 5

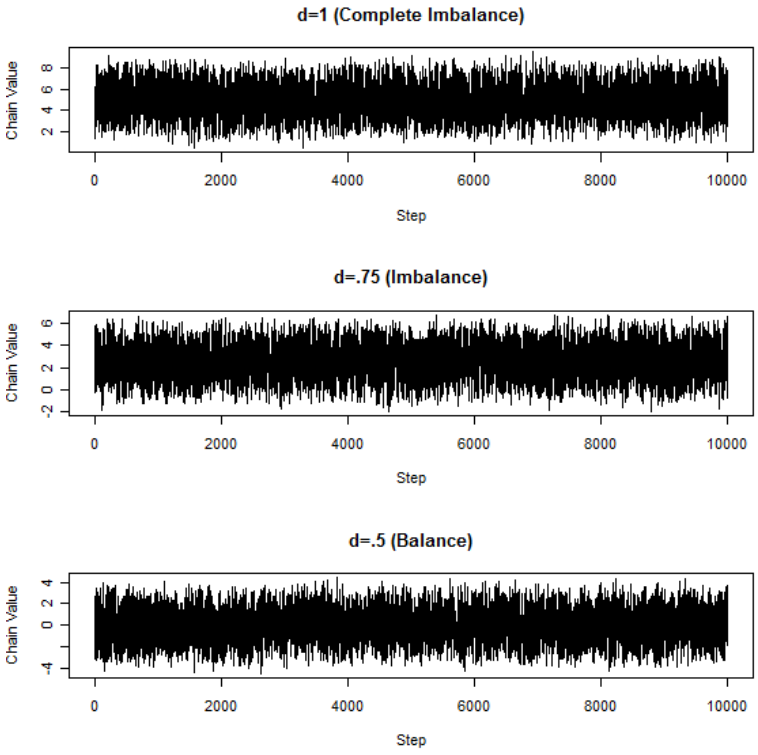


Figure 63 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range of the chain elements, and the particular values they take on, corresponding to the direction of the imbalance.

Figure 64 shows the path plots for all levels of d when c is equal to .75 and the range is equal to 5. It is presented below.

Figure 64: Path plots for all levels of d when c = .75 and range = 5

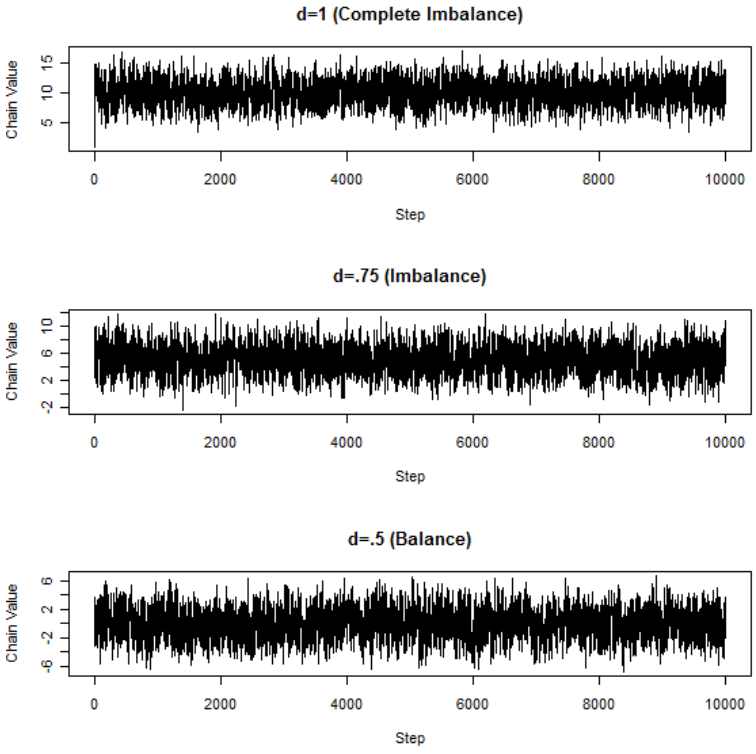


Figure 64 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range of the chain elements, and the particular values they take on, corresponding to the direction of the imbalance.

Figure 65 shows the path plots for all levels of d when c is equal to .9 and the range is equal to 5. It is presented below.

Figure 65: Path plots for all levels of d when c = .9 and range = 5

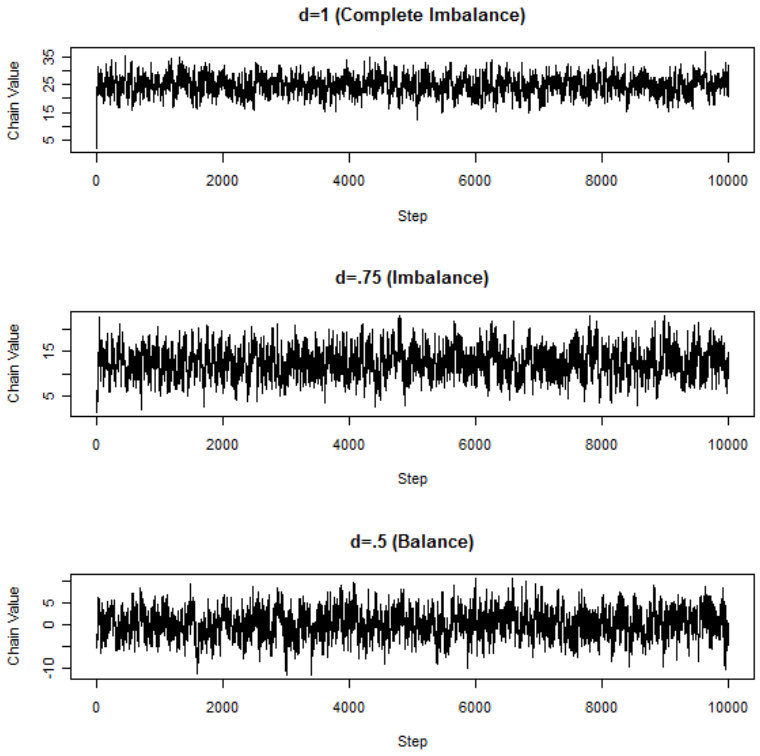


Figure 65 shows the same expansion of range that was seen in the descriptive statistics for this particular set of conditions. The influence of the factors c and d can be seen in the expanded range of the chain elements, and the particular values they take on, corresponding to the direction of the imbalance.

Figure 66 shows the path plots for all levels of d when c is equal to 1 and the range is equal to 5. It is presented below.

Figure 66: Path plots for all levels of d when c = 1 and range = 5

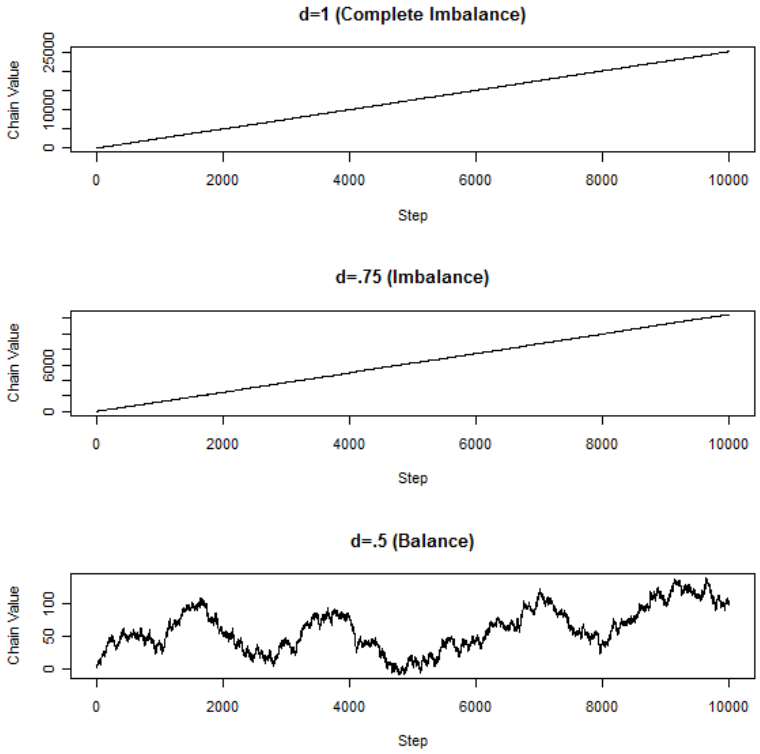


Figure 66 shows the behavior of the chain elements that was seen in the descriptive statistics for this particular set of conditions. The influence of d and c can be seen in the behavior of the chain. Specifically, because c is equal to 1, the complete imbalance condition is a strictly non-decreasing sequence of values. When there is partial imbalance present and c is equal to 1, the behavior of the chain is consistent with what would be expected. Specifically, each new element is set equal to the previous plus a random component that was twice as likely to be positive as it is to be negative. Thus the value that chain elements take on is more likely to increase rather than decrease over the length of the chain. However, chains produced in this condition are not strictly non-

decreasing. When there is balance present and c is equal to 1, it is equally likely that each new element will be greater than or less than the previous element.

Overall, there is no influence of the range of the random component of the chain simulator on the summary of the indicator statistic, D_t, or on the autocorrelation plots. However, the range of the random component does have an influence on the behavior of the chains. Specifically, as the range of the random component becomes larger, the range that the chain elements can take on becomes larger. Otherwise, the general trend that was encountered in the previous studies involving the simulated chains with these values of c and d is the same as the pattern of results observed in all tables and figures related to Research Question 3. When c is 0, the chains from all levels of d and range are i.i.d. sequences with bound equal to those implied by the level of d and range. While c is .25 through .9, the chains tend to stabilize into a specific range related to the range of the random component of the chain simulator and the level of c and d. As c increases, the chain elements tend to become more variable. As d goes from .5 to 1 the values of the chain elements tends to increase. These results will be revisited again.

Next, the results for the convergence diagnostics will be presented. The convergence diagnostics for all levels of c will presented in a single table. There will be a table for each combination of the levels of d and the levels of the range of the random component of the chain simulator. Each table contains the proportion of chains in that condition that would be deemed non-converged according to the convergence diagnostics (Pr NC). The four diagnostics are D_t, the Geweke diagnostic (G), the Heidelberger and Welch diagnostic (HW), and the Raftery and Lewis diagnostic (RL), all of which were

described earlier. The 'boa' package in R was used to compute G, HW and RL. For all diagnostics, an alpha level of .05 was selected. In addition to the proportion of non-converged chains, the table also provides the mean of the statistic/criterion used to determine convergence (where available), the maximum of the statistic, and the minimum of the statistic. The output of the software used for the stationarity test of the HW diagnostic only provided whether or not the chain passed or failed at the given significance level, therefore, descriptive statistics are not available for this diagnostic. The Geweke diagnostic can be interpreted like a z score, as it is the difference between the means for the beginning and end portion of the chains corrected for the variability present. Any value more extreme than -1.96 or 1.96 is considered to be associated with a chain that is non-converged. For the RL diagnostic, the chain length necessary to achieve convergence is reported. When this value is greater than the chain used as input (10,000), a chain is deemed non-converged. The results for each set of conditions will be briefly described, and then an overall summary of the conditions together will be provided.

Table 35 contains the convergence diagnostics for the case where d is equal to 1 and range is equal to .1. It is presented below.

Table 35: Convergence diagnostics for d = 1 and range = .1

		d=1			
c	**Statistic**	***D***	**G**	**HW**	**RL**
0.0	**Pr NC**	0.0	.08	0.0	0.0
	Mean	.667	-.05	-	384.9
	SD	.004	1.20	-	6.34
	Max	.675	2.01	-	397
	Min	.661	-2.33	-	373
.25	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.624	-.18	-	835.5
	SD	.004	1.01	-	80
	Max	.631	1.22	-	1209
	Min	.613	-2.34	-	792
.50	**Pr NC**	1.0	0	0.0	0.0
	Mean	.582	-.27	-	1475
	SD	.004	.77	-	198
	Max	.588	1.57	-	1756
	Min	.573	-1.88	-	1224
.75	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.541	-.52	-	3211
	SD	.005	.81	-	409
	Max	.550	1.05	-	3996
	Min	.532	-2.18	-	2616
.90	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.516	-.57	-	7065
	SD	.004	.84	-	903
	Max	.524	1.25	-	9320
	Min	.509	-2.23	-	5436
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	0.0	-38.7	-	34536
	SD	0.0	.32	-	0.0
	Max	0.0	-38.21	-	34536
	Min	0.0	-39.36	-	34536

In Table 35, D_t indicates that when c is equal to zero, all chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 2, 1, 0, 1, and 1, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 36 contains the convergence diagnostics for the case where d is equal to .75 and range is equal to .1. It is presented below.

Table 36: Convergence diagnostics for d = .75 and range = .1

		d=.75			
c	Statistic	D_t	G	HW	RL
0.0	Pr NC	0.0	0.0	0.0	0.0
	Mean	.668	-.35	-	385.2
	SD	.004	.86	-	7.50
	Max	.675	.89	-	401
	Min	.663	-1.70	-	369
.25	Pr NC	1.0	.04	0.0	0.0
	Mean	.625	.25	-	827.2
	SD	.004	1.20	-	19.7
	Max	.675	2.74	-	864
	Min	.663	-1.92	-	800
.50	Pr NC	1.0	.04	0.0	0.0
	Mean	.584	.24	-	1460
	SD	.004	1.10	-	193
	Max	.592	1.86	-	1764
	Min	.575	-2.48	-	1206
.75	Pr NC	1.0	0.0	.04	0.0
	Mean	.542	-.29	-	3100
	SD	.005	.73	-	388
	Max	.553	1.50	-	3915
	Min	.532	-1.80	-	2562
.90	Pr NC	1.0	0.0	0.0	0.0
	Mean	.514	-.21	-	7370
	SD	.006	.89	-	813
	Max	.524	1.62	-	8517
	Min	.503	-1.82	-	5368
1.0	Pr NC	1.0	1.0	1.0	1.0
	Mean	.375	-38.7	-	168220
	SD	.006	.72	-	369440
	Max	.387	-37.2	-	204107
	Min	.368	-40.3	-	36889

In Table 36, D_t indicates that when c is equal to zero, all chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 0, 1, 1, 0, and 0, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, only one chain is deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves as it did in the previous table. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 37 contains the convergence diagnostics for the case where d is equal to .5 and range is equal to .1. It is presented below.

Table 37: Convergence diagnostics for d = .5 and range = .1

		d=.5			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	.04	.04	0.0	0.0
	Mean	.666	-.16	-	387.1
	SD	.004	1.01	-	8.46
	Max	.676	1l72	-	403
	Min	.661	-2.21	-	371
.25	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.625	-.21	-	857
	SD	.004	1.12	-	117.4
	Max	.633	2.12	-	1254
	Min	.617	-2,11	-	774
.50	**Pr NC**	1.0	0.0	.04	0.0
	Mean	.583	.22	-	1501
	SD	.003	.91	-	206
	Max	.591	1.72	-	1784
	Min	.579	-1.38	-	1191
.75	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.542	-.27	-	3206
	SD	.005	.74	-	450
	Max	.552	1.22	-	4820
	Min	.537	-1.84	-	2694
.90	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.515	-.01	-	6794
	SD	.005	1.14	-	706
	Max	.523	2.07	-	8466
	Min	.504	-1.88	-	5291
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	.500	-38.7	-	311760
	SD	.006	.57	-	193028
	Max	.512	-38.1	-	719256
	Min	.489	-40.7	-	128084

In Table 37, D_t indicates that when c is equal to zero, all but one of the chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 2, 0, 0, and 1, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, only one chain is deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves as it did in the previous table. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 38 contains the convergence diagnostics for the case where d is equal to 1 and range is equal to .5. It is presented below.

Table 38: Convergence diagnostics for d = 1 and range = .5

		d=.75			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	.04	.04	0.0	0.0
	Mean	.666	.21	-	387
	SD	.004	.91	-	7.29
	Max	.671	1.98	-	399
	Min	.658	-1.32	-	375
.25	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.625	-.12	-	837.4
	SD	.003	.99	-	68.8
	Max	.636	1.45	-	1152
	Min	.617	-2.06	-	778
.50	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.584	-.32	-	1455
	SD	.003	.98	-	213
	Max	.595	1.86	-	1965
	Min	.578	-2.46	-	1233
.75	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.542	-.43	-	3124
	SD	.005	.66	-	434
	Max	.553	.75	-	4630
	Min	.534	-1.79	-	2610
.90	**Pr NC**	1.0	.20	0.0	0.0
	Mean	.516	-1.08	-	7440
	SD	.005	.87	-	783
	Max	.526	1.55	-	8874
	Min	.510	-2.32	-	6202
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	0	-38.7	-	34536
	SD	0	.37	-	0
	Max	0	-38.1	-	34536
	Min	0	-39.5	-	34536

In Table 38, D_t indicates that when c is equal to zero, all but one of the chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 2, 1, 0, and 1, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 39 contains the convergence diagnostics for the case where d is equal to .75 and range is equal to .5. It is presented below.

Table 39: Convergence diagnostics for d = .75 and range = .5

		d=.75			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	0.0	0.0	0.0	0.0
	Mean	.667	.114	-	387.4
	SD	.003	1.03	-	8.36
	Max	.673	1.99	-	404
	Min	.661	-1,56	-	375
.25	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.624	.02	-	859.1
	SD	.004	.92	-	112
	Max	.631	2.26	-	1245
	Min	.618	-1.38	-	780
.50	**Pr NC**	1.0	0	0.0	0.0
	Mean	.583	.19	-	1485
	SD	.004	.76	-	203
	Max	.590	1.47	-	1764
	Min	.572	-1.18	-	1248
.75	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.541	.06	-	3050
	SD	.005	.92	-	297
	Max	.550	1.90	-	3654
	Min	.535	-1.60	-	2496
.90	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.512	-.40	-	7033
	SD	.005	1.26	-	832
	Max	.528	1.83	-	8670
	Min	.510	-2.95	-	5904
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	.375	-38.9	-	285698
	SD	.005	.66	-	462363
	Max	.381	-37.29	-	1148703
	Min	.367	-40.1	-	34536

In Table 39, D_t indicates that when c is equal to zero, all chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 0, 1, 0, 0, and 2, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 40 contains the convergence diagnostics for the case where d is equal to .5 and range is equal to .5. It is presented below.

Table 40: Convergence diagnostics for d = .5 and range = .5

		d=.5			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	0.0	.04	0.0	0.0
	Mean	.667	.51	-	386.6
	SD	.004	.90	-	7.30
	Max	.674	2.04	-	402
	Min	.660	-1.48	-	372
.25	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.625	.04	-	858
	SD	.004	1.08	-	111
	Max	.634	1.73	-	1254
	Min	.619	-2.19	-	766
.50	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.584	-.14	-	1504
	SD	.003	.72	-	189
	Max	.595	1.51	-	1796
	Min	.579	-2.43	-	1260
.75	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.542	-.22	-	3130
	SD	.004	.79	-	294
	Max	.553	1.43	-	3760
	Min	.533	-1.53	-	2664
.90	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.517	.09	-	7061
	SD	.005	1.08	-	731
	Max	.524	2.07	-	8493
	Min	.510	-3.01	-	5368
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	.499	-	-	374719
	SD	.005	-	-	162218
	Max	.507	-	-	663264
	Min	.491	-	-	129980

In Table 40, D_t indicates that when c is equal to zero, all chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 2, 1, 0, and 2, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 41 contains the convergence diagnostics for the case where d is equal to 1 and range is equal to 1. It is presented below.

Table 41: Convergence diagnostics for d = 1 and range = 1

		d=1			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	0.0	.04	0.0	0.0
	Mean	.666	-.08	-	389
	SD	.004	1.08	-	7.53
	Max	.674	1.91	-	406
	Min	.661	-2.01	-	376
.25	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.624	-.03	-	814.0
	SD	.003	1.07	-	22.1
	Max	.632	2.32	-	868
	Min	.615	-2.45	-	780
.50	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.584	-.12	-	1511
	SD	.004	.96	-	197
	Max	.593	1.38	-	1768
	Min	.577	-2.04	-	1242
.75	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.543	-.02	-	3108
	SD	.005	.66	-	420
	Max	.551	1.79	-	4239
	Min	.534	-1.47	-	2556
.90	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.517	-.49	-	6921
	SD	.006	1.10	-	736
	Max	.532	2.07	-	8415
	Min	.506	-3.09	-	5772
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	0	-38.9	-	34536
	SD	0	.36	-	0
	Max	0	-38.0	-	34536
	Min	0	-39.5	-	34536

In Table 41, D_t indicates that when c is equal to zero, all chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 2, 1, 0, and 2, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 42 contains the convergence diagnostics for the case where d is equal to .75 and range is equal to 1. It is presented below.

Table 42: Convergence diagnostics for d = .75 and range = 1

		d=.75			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	0.0	.04	0.0	0.0
	Mean	.667	.26	-	387.0
	SD	.004	.97	-	7.17
	Max	.675	2.13	-	403
	Min	.659	-1.34	-	374
.25	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.626	-.01	-	824.3
	SD	.004	.79	-	166
	Max	.634	2.16	-	856
	Min	.621	-1.28	-	798
.50	**Pr NC**	1.0	0	0.0	0.0
	Mean	.584	-.14	-	1465
	SD	.005	.79	-	210
	Max	.590	1.72	-	1768
	Min	.577	-1.36	-	1248
.75	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.542	-.21	-	3195
	SD	.005	1.14	-	301
	Max	.552	1.79	-	3744
	Min	.533	-2.47	-	2670
.90	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.513	-.43	-	7318
	SD	.005	.94	-	866
	Max	.523	1.57	-	9432
	Min	.508	-1.8	-	5760
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	.376	-38.9	-	391046
	SD	.006	.71	-	530415
	Max	.386	-37.0	-	1148812
	Min	.363	-39.9	-	34532

In Table 42, D_t indicates that when c is equal to zero, all chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 1, 0, 2, and 0, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 43 contains the convergence diagnostics for the case where d is equal to .5 and range is equal to 1. It is presented below.

Table 43: Convergence diagnostics for d = .5 and range = 1

		d=.5			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	0.0	.04	0.0	0.0
	Mean	.666	.01	-	384.8
	SD	.003	.95	-	6.78
	Max	.673	2.50	-	398
	Min	.658	-1.45	-	373
.25	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.625	.09	-	839
	SD	.004	.81	-	73.1
	Max	.632	1.55	-	1176
	Min	.622	-1.25	-	778
.50	**Pr NC**	1.0	.12	0.0	0.0
	Mean	.585	-.01	-	1510
	SD	.005	1.09	-	184
	Max	.588	2.14	-	1796
	Min	.576	-2.21	-	1257
.75	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.542	-.19	-	3114
	SD	.004	.77	-	327
	Max	.551	.97	-	3792
	Min	.530	-1.43	-	2653
.90	**Pr NC**	1.0	.04	.04	0.0
	Mean	.516	-.03	-	7114
	SD	.005	1.00	-	1050
	Max	.527	1.65	-	9108
	Min	.510	-2.70	-	5568
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	.499	-	-	346903
	SD	.005	-	-	195113
	Max	.507	-	-	673937
	Min	.492	-	-	34227

In Table 43, D_t indicates that when c is equal to zero, all chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 0, 3, 0, and 1, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, all but one of the chains is deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 44 contains the convergence diagnostics for the case where d is equal to 1 and range is equal to 5. It is presented below.

Table 44: Convergence diagnostics for d = 1 and range = 5

		d=1			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	.08	.04	0.0	0.0
	Mean	.666	-.07	-	388
	SD	.004	1.01	-	9.59
	Max	.678	2.53	-	407
	Min	.658	-1.74	-	372
.25	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.626	-.06	-	857.1
	SD	.005	.95	-	107
	Max	.637	2.02	-	1227
	Min	.613	-2.06	-	760
.50	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.585	-.36	-	1568
	SD	.004	.89	-	197
	Max	.594	1.41	-	2010
	Min	.578	-2.03	-	1248
.75	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.542	-.06	-	3115
	SD	.005	.97	-	307
	Max	.552	1.49	-	3712
	Min	.533	-2.43	-	2652
.90	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.518	-.55	-	7498
	SD	.005	.75	-	729
	Max	.528	.53	-	9360
	Min	.510	-1.77	-	5832
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	0	-38.8	-	34536
	SD	0	.27	-	0
	Max	0	-38.3	-	34536
	Min	0	-39.4	-	34536

In Table 44, D_t indicates that when c is equal to zero, all but two chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 2, 1, 2, and 0, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 45 contains the convergence diagnostics for the case where d is equal to .75 and range is equal to 5. It is presented below.

Table 45: Convergence diagnostics for d = .75 and range = 5

		d=.75			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	.04	.04	0.0	0.0
	Mean	.668	.03	-	385.1
	SD	.004	1.01	-	10.15
	Max	.677	2.29	-	406
	Min	.659	-1.76	-	371
.25	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.624	.04	-	889.3
	SD	.004	.85	-	137
	Max	.631	1.65	-	1215
	Min	.616	-1.49	-	792
.50	**Pr NC**	1.0	.04	.04	0.0
	Mean	.585	-.43	-	1525
	SD	.005	.90	-	216
	Max	.592	1.08	-	2145
	Min	.573	-2.62	-	1260
.75	**Pr NC**	1.0	.08	0.0	0.0
	Mean	.541	.01	-	3189
	SD	.006	.97	-	278
	Max	.551	2.16	-	3808
	Min	.534	-1.79	-	2760
.90	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.517	-.04	-	6789
	SD	.005	.95	-	613
	Max	.529	2.04	-	8330
	Min	.506	-1.89	-	5754
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	.377	-38.8	-	257378
	SD	.007	.47	-	454875
	Max	.393	-37.8	-	1147536
	Min	.365	-39.4	-	34536

In Table 45, D_t indicates that when c is equal to zero, all but one of the chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 1, 0, 1, 2, and 1, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, all but one of the chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

Table 46 contains the convergence diagnostics for the case where d is equal to .5 and range is equal to 5. It is presented below.

Table 46: Convergence diagnostics for d = .5 and range = 5

		d=.5			
	Statistic				
c		***D_t***	**G**	**HW**	**RL**
0.0	**Pr NC**	.04	0.0	0.0	0.0
	Mean	.668	.01	-	388.3
	SD	.004	.95	-	6.79
	Max	.674	1.49	-	403
	Min	.656	-1.91	-	377
.25	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.624	.28	-	837
	SD	.005	.72	-	82.5
	Max	.632	1.91	-	1218
	Min	.618	-1.02	-	766
.50	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.582	.25	-	1541
	SD	.005	.90	-	178
	Max	.588	2.02	-	1760
	Min	.572	-1.46	-	1242
.75	**Pr NC**	1.0	0.0	0.0	0.0
	Mean	.542	.06	-	3067
	SD	.004	.82	-	321
	Max	.549	1.56	-	3933
	Min	.536	-1.27	-	2448
.90	**Pr NC**	1.0	.04	0.0	0.0
	Mean	.517	-.03	-	7065
	SD	.004	1.23	-	793
	Max	.525	1.91	-	8415
	Min	.510	-2.63	-	5784
1.0	**Pr NC**	1.0	1.0	1.0	1.0
	Mean	.501	-	-	336239
	SD	.005	-	-	205545
	Max	.509	-	-	817920
	Min	.489	-	-	33450

In Table 46, D_t indicates that when c is equal to zero, all but one of the chains are deemed converged. For all cases where c is greater than zero, each and every chain is deemed non-converged. Whenever there is some degree of autocorrelation present in the sequence, the value of D_t over both patterns satisfying the indicator statistic decreases to the point that it falls outside the bounds as specified by Brooks (1998c), thus these chains are deemed non-converged. For the Geweke diagnostic, very few chains are deemed non-converged while c is less than 1. For example, when c is equal to 0, .25, .5, .75 and 9, the number chains deemed non-converged are 0, 0, 1, 0, and 1, respectively. This is a roughly chance level of detection of non-convergence for each level of c. When c is equal to one, all chains are deemed non-converged by the G diagnostic. For the Heidelberger and Welch diagnostic, when c is less than 1, no chains are deemed non-converged. When c is equal to 1, then the HW diagnostic deems every chain non-converged. The Raftery and Lewis diagnostic behaves similarly to the HW diagnostic. When c is less than 1, the RL diagnostic indicates that no chains are non-converged. When c is equal to 1, the RL diagnostic deems all chains non-converged.

The amount of agreement between D_t and each of the other diagnostics can be quantified with kappa. Kappa was calculated for all levels of c, d and range. The observed kappa values for all levels of c and d are summarized in Table 47 below for the case where the range is equal to .1.

Table 47: Agreement (kappa) between D_t and G, HW, and RL for all levels of c and d when range is equal to .1

c	d = 1.0			d = .75			d = .5		
	G	**HW**	**RL**	**G**	**HW**	**RL**	**G**	**HW**	**RL**
0.0	.84	1	1	1	1	1	.84	.92	.92
.25	-.92	-1	-1	-.92	-1	-1	-.84	-1	-1
.50	-1	-1	-1	-.92	-1	-1	-1	-.92	-1
.75	-.92	-1	-1	-1	-.92	-1	-1	-1	-1
.90	-.92	-1	-1	-1	-1	-1	-.92	-1	-1
1.0	1	1	1	1	1	1	1	1	1

Kappa quantifies the agreement between the new diagnostic and the existing diagnostics. Table 47 demonstrates that there are cases where the new diagnostic agrees with the existing diagnostics, and there are cases where there is disagreement. There is some evidence that the level of c influences the value of Kappa. Specifically, when c is equal to 0 or 1, the new diagnostic agrees with the existing ones. However, for all other levels of c, the new diagnostic largely disagrees with the existing diagnostics. There is no evidence that d influences Kappa.

The observed kappa values for all levels of c and d are summarized in Table 48 below for the case where the range is equal to .5.

Table 48: Agreement (kappa) between D_t and G, HW, and RL for all levels of c and d when range is equal to .5

c	**d**								
	1.0			**.75**			**.5**		
	G	**HW**	**RL**	**G**	**HW**	**RL**	**G**	**HW**	**RL**
0.0	.92	.92	.92	.92	1	1	.92	1	1
.25	-.92	-1	-1	-.84	-1	-1	-.84	-1	-1
.50	-1	-1	-1	-.92	-1	-1	-.92	-1	-1
.75	-1	-1	-1	-1	-1	-1	-1	-1	-1
.90	-.84	-1	-1	-.6	-1	-1	-.84	-1	-1
1.0	1	1	1	1	1	1	1	1	1

Table 48 is very similar to Table 47. It demonstrates that there are cases where the new diagnostic agrees with the existing diagnostics, and there are cases where there is disagreement. There is some evidence that the level of c influences the value of Kappa. Specifically, when c is equal to 0 or 1, the new diagnostic agrees with the existing ones. However, for all other levels of c, the new diagnostic largely disagrees with the existing diagnostics. There is no evidence that d influences Kappa.

The observed kappa values for all levels of c and d are summarized in Table 49 below for the case where the range is equal to 1.

Table 49: Agreement (kappa) between D_t and G, HW, and RL for all levels of c and d when range is equal to 1

c	**d**								
	1.0			**.75**			**.5**		
	G	**HW**	**RL**	**G**	**HW**	**RL**	**G**	**HW**	**RL**
0.0	.84	1	1	1	1	1	.84	.92	.92
.25	-.92	-1	-1	-.92	-1	-1	-.84	-1	-1
.50	-1	-1	-1	-.92	-1	-1	-1	-.92	-1
.75	-.92	-1	-1	-1	-.92	-1	-1	-1	-1
.90	-.92	-1	-1	-1	-1	-1	-.92	-1	-1
1.0	1	1	1	1	1	1	1	1	1

Table 49 is similar to Table 47 and Table 48. It also demonstrates that there are cases where the new diagnostic agrees with the existing diagnostics, and there are cases where there is disagreement. There is some evidence that the level of c influences the value of Kappa. Specifically, when c is equal to 0 or 1, the new diagnostic agrees with the existing ones. However, for all other levels of c, the new diagnostic largely disagrees with the existing diagnostics. There is no evidence that d influences Kappa.

The observed kappa values for all levels of c and d are summarized in Table 50 below for the case where the range is equal to 5.

Table 50: Agreement (kappa) between D_t and G, HW, and RL for all levels of c and d when range is equal to 5

c	**d**								
	1.0			**.75**			**.5**		
	G	**HW**	**RL**	**G**	**HW**	**RL**	**G**	**HW**	**RL**
0.0	.84	1	1	1	1	1	.84	.92	.92
.25	-.92	-1	-1	-.92	-1	-1	-.84	-1	-1
.50	-1	-1	-1	-.92	-1	-1	-1	-.92	-1
.75	-.92	-1	-1	-1	-.92	-1	-1	-1	-1
.90	-.92	-1	-1	-1	-1	-1	-.92	-1	-1
1.0	1	1	1	1	1	1	1	1	1

Table 50 agrees with the results presented in Tables 47, 48, and 49. It also demonstrates that there are cases where the new diagnostic agrees with the existing diagnostics, and there are cases where there is disagreement. There is some evidence that the level of c influences the value of Kappa. Specifically, when c is equal to 0 or 1, the new diagnostic agrees with the existing ones. However, for all other levels of c, the new diagnostic largely disagrees with the existing diagnostics. There is no evidence that d influences Kappa.

In summary, when c is equal to 0 or 1, there is a great deal of agreement among all four of the convergence diagnostics. Very few of the chains produced for the case where c is equal to 0 were deemed non-converged. The rate of detection is what would be expected according to chance. Because an alpha level of .05 was used for all diagnostics and 25 chains were produced for each condition in the study, it is not surprising that one or maybe even two of the chains produced diagnostics that would be deemed non-converged by chance alone. When c is equal to 1, all chains from all conditions were deemed non-converged by all diagnostics. This finding was anticipated because chains produced from all cases where c is equal to 1 showed a high degree of autocorrelation and tended to be unstable. Depending on the level of d, these chains moved up, down, or both up and down over the length of the chain. The finding of most importance deals with the cases where there is a large discrepancy among the diagnostics. When c was equal to .25, .5, .75, or .9, the value of D_t was always outside the boundaries specified by Brooks (1998c). Each of the chains from these conditions was deemed non-converged by the new method. However, the other diagnostics rarely, if ever, deemed chains produced from conditions where c was .25, .5, .75, and .9 to be non-converged. The Geweke diagnostic would sometimes identify 1, 2, or even 3 chains from these conditions that were deemed non-converged. The Heidelberger and Welch diagnostic would occasionally identify a single chain from these conditions that was deemed non-converged. The Raftery and Lewis diagnostic never identified a chain from these conditions as being non-converged. These findings must be interpreted keeping in mind that the simulated chains are essentially producing converged chains. This issue will be redressed in the Discussion.

CHAPTER V

DISCUSSION

The discussion will be divided into several sections. First, the results of the simulation studies will be briefly revisited in order to summarize the information provided by each study as it pertains to the associated research question. As each study and its findings are discussed, an attempt will be made to link data to aspects of the specific research question under consideration. Second, the strengths and limitations of the current method will be summarized. Both the convergence diagnostic being developed, D_t, and the method of simulating chains will receive consideration. Finally, future research that needs to be done to address limitations and additional questions raised by this research will be discussed.

Summarizing the results

The current research has been undertaken to determine whether or not the current method of assessing convergence is capable of being used for that purpose. Each of the simulation studies was designed to provide information concerning a particular issue that needs to be addressed before the new method can be put into use. Some of these issues were phrased in the form of the research question addressed in this study. Together, the simulation studies provide information about how the indicator statistic might be used in practice. The studies provide information about how the indicator statistic will be influenced by some of the known characteristics of Markov chains, such as linear

dependence among elements and stability of location of the chains, how the method should behave with and without thinning, and how the method compares to some existing methods. Each simulation study will be presented in order, and the contribution of the evidence from each study will be considered in light of the question it was intended to answer.

Research question 1

The purpose of the first research question is to learn the effect that autocorrelation and balance (in the form of c and d) have on the value of D_t. This question was primarily asked because of the known problems with thinning the chain to assess convergence (MacEachern and Berliner, 1995). If there is to be any chance of using this method without thinning the chains prior to applying the indicator statistic, it must be known exactly what influence autocorrelation has on the summary of the indicator statistic.

Autocorrelation

Overall, as anticipated, as the amount of autocorrelation present in the chains increased, the value of D_t decreased. Thus, the answer to the first research question is in part revealed. Increased amounts of autocorrelation are associated with reduced values of D_t. Without thinning of the chains to reduce the amount of dependence present among elements, the expected value of the summary of the indicator statistic would need to be adjusted downwards. What remains to be determined is exactly what amount of autocorrelation at each lag is associated with the exact amount of reduction in $\underline{D_t}$. This would also in turn affect the value of the boundaries as proposed by Brooks (1998c) which are centered at the expected value of D_t. However, it must also be considered that

there are other things that can reduce the value of D_t other than autocorrelation. For example, other evidence shows that the number of ties in the chain, especially consecutive ties, can influence the value of the summary of the indicator statistic.

Balance

When it comes to the balance of the chains, the answer of whether or not this factor influences the summary of the indicator statistic is not so clear. An unanticipated outcome occurred by which the simulated solutions were in disagreement with the analytical solution. This unanticipated outcome requires further explanation. The findings indicate that one of the two solutions is incorrect. There is reason to believe that the simulated solutions are somewhat incomplete in their representation of the analytical solutions. First the analytical solutions will be briefly reviewed and then the simulated solutions will be discussed and critiqued.

The analytical solutions showed the expected relationship among c, d, and D_t. As the 'autocorrelation' factor c increased, the value of D_t decreased, and did so at a greater rate when imbalance was present to greater degrees. These findings seem to be clearly in line with initial expectations. There is an implicit assumption in the analytical solution that the range of the random component shares a steady relationship with the value of the previous element (which is multiplied by c). This assumption was not incorporated into the method for simulating chains, and is presumed to be the primary source of the difference between the two solutions.

Now attention is turned to the simulated solutions for the value of D_t based on the simulated chains. Investigation of the simulated chains showed that there was some relationship between the factor c and the range of the random component of the chain simulator (1 in this study) that prevented the simulated chains from behaving as initially predicted. Essentially, the chains will move in one direction until the proportional reduction, c, of element x_{i-1} is greater in degree than the range of the continuous uniform random component. This relationship prevents the imbalance from behaving as initially predicted across the length of the chain. The chain has some of the desired characteristics (e.g., a mean that is unstable in the short term), but didn't behave exactly as predicted in regard to the influence of balance. It should be noted here that when there was no autocorrelation (c = 0), or the strongest degree of autocorrelation (c = 1), the simulated solutions did match the analytical solutions. Initially, it was thought that when an imbalanced chain is being simulated, it is more likely to move in one direction than another and that this would take place gradually over the length of the chain. Instead, the chains tended to move in one direction very quickly, and then change direction abruptly. This pattern of changing directions then alternates again and again throughout the chains.

Let us briefly consider the chain as it is simulated from the first element. As the chain moves in one direction more so than another, the value that any individual element takes on tends to become quite large (as compared to the range of the continuous random uniform component). Once this happens, then taking some proportion of that element, say c = .5, moves the new element far enough towards zero that the following element is constrained in such a way as to produce a pattern among elements that satisfies the

indicator statistic being equal to one. The value of d did have an influence on the particular location where the chain elements were simulated. The indicator statistic is only sensitive to the patterns of rank orderings, so all of these chains produce a similar value of D_t. Thus, regardless of the degree of imbalance, the simulated chains all behaved in similar fashion as far as the current method is concerned.

Summarizing the first research question

In summary for the first research question, it is not entirely clear what the full relationship is among c, d and D_t. The analytical solutions fit with expectations regarding c and d influencing D_t. However, the simulated solutions aren't in line with expectations, so it raises questions concerning the strength of assertions made about the influence of c and d. It does seem to be the case for the simulated solutions, though, that as autocorrelation increases, D_t decreases. Because this occurs in both sets of solutions it provides some convergent evidence for the influence of autocorrelation on the value of the summary of the indicator statistic. However, the fact that the simulated chains do not behave as expected should give reason to interpret these results with the explanation of why the two sets of solutions differ in mind.

Research question 2

The second research question addressed the practice of thinning the chain prior to characterizing it by way of the indicator statistic. For the first simulation study regarding this question, chains were simulated in the same fashion as in the first simulation study, and the degree of autocorrelation present in the chains was used to determine the amount of thinning necessary for the chains. Thinning was done by inspecting the autocorrelation

plots to determine the lag at which the value of the autocorrelation was not significant at alpha = .05. Visual inspection of the autocorrelation plots is the most efficient way of proceeding, but for the sake of this research, the study is informative about the influence of thinning.

Due to the uniformity of results across replications, the same lag was used to thin all chains for a given level of c. Except for the case with the strongest possible degree of autocorrelation (c = 1), the thinning left a subsample of elements that was linearly dependent. When characterized by way of D_t, all of the chains that showed linear independence were also deemed converged according to the criterion employed as described in the methods. This finding provides evidence that the practice of thinning removes the linear dependence among elements, thus increasing the proportion of elements that satisfy the indicator statistic being equal to one for the simulated chains. When considering the fact that the method used for simulating chains in this work essentially produced converged chains, the results of simulation study 2 become clearer. In short, these chains behave as if converged, but have some amount of autocorrelation which influences the value of the summary of the indicator statistic. The thinning removes the autocorrelation, thus, the value of D_t moves towards .67.

A fair criticism of the current method is that the act of thinning may be making the remaining elements look like a converged sequence, regardless of the true state of the Markov chain under consideration. There is some evidence that this is not the case, though. When the chains from the real MCMC samplers were investigated for simulation study 3, the thinned chains showed a great deal of variability. In fact, many of the chains

would have been deemed non-converged according to the criteria (D_t above or below the upper and lower bounds, respectively) in place for this study. So, there are chains that when thinned and characterized by the indicator statistic are sometimes inside the bounds, and sometimes above or below the bounds. This finding at least provides some evidence that the criticism of the method simply giving the appearance of convergence may be invalid.

It is not clear from simulation study 1 or 2, however, whether or not the thinning is affecting the quality of the estimate. In simulation study 3, the true values of the parameters are known. The third simulation study involved generating chains from a real sampler rather than simulating chains. Also, because the true parameters are known it is possible to determine the accuracy of the chains in their estimates of the parameters. Tables 16 and 17 and Table 21 and 22 showed that the MADs for the chains produced by the conditions manipulated in this study produced estimates that were accurate. While, MADs have no clear objective criterion to indicate 'good fit', it is clear from this study that in general it is possible to recapture the true values used to generate the data. While the conditions of this study are not overly rigorous, it at least provides some evidence that the thinning may not be detrimental enough to prohibit the practice of thinning.

Simulation study 3 manipulated the ratio of the standard deviations of the proposal and target distributions to affect the degree of autocorrelation present in the chains. There was some evidence that the level of RATIO influenced the value of D_t, and did so by affecting the autocorrelation present in the chains. However, there was a strong influence on the value of D_t that was due to the number of ties present in the chains

produced from the real MCMC sampler. The large number of ties was due to the acceptance rates of the samplers. As the value of RATIO increased, the number of ties increased in the Markov chains. As the indicator statistic is currently phrased in terms of strict inequalities, the large amount of tied, consecutive elements depresses the value of D_t. In the extreme case of RATIO being equal to 4, there were often so many ties and one directional changes that no elements in the chain were coded as 1, and the value of D_t was equal to 0. Brooks (1998c) suggests that all ties be set equal to the expected value of the summary of the indicator statistic. However, some of the chains presented show clearly that setting ties equal to the expected value of 2/3 would bias them towards looking like 'converged chains' when they are in fact not good estimates.

Further research will need to address the questions remaining regarding the appropriateness of thinning the chains. However, it has been clearly demonstrated that thinning will remove the dependence among elements, and when conditions are such that the thinning by autocorrelation leaves a good deal of the original elements, then this may be a viable technique to employ. At least for the chains produced in this study, it has been shown that thinning did not necessarily degrade the quality of the estimates obtained.

Research question 3

The third research question compares the current diagnostic to three existing methods. In simulation study 4, chains were again simulated and then characterized by each of the convergence diagnostics. Attempts were made to standardize the criteria used by each technique to make a determination of convergence. For example, the alpha level used for all diagnostics was .05. No thinning was performed on these chains. The

proportion of chains deemed non-converged was presented for each method, as well as a measure of agreement (kappa).

The current (and previous) results demonstrate that the new method being developed is especially sensitive to the patterns of rank orderings of the elements in these chains. When there is no autocorrelation present in the chains, all chains were deemed to fall within the boundaries. When there is any degree of autocorrelation present in the chains, the value of D_t decreases to the point it falls outside the boundaries. The boundaries are influenced by the number of elements in the chain. As the number of elements increases, the boundaries get narrower. The chains in this study all contain 10,000 elements, so the boundaries for differentiating convergence from non-convergence are quite narrow. So, all cases where c was greater than 0, the current method deemed the chains non-converged. Perhaps the criterion used to distinguish between convergence and non-convergence is too strict.

For the other three methods, all chains are deemed non-converged when c is equal to 1. The chains in these conditions have a strong degree of autocorrelation, and they also tend to have unstable locations. Specifically, the mean at any point in the chains is relatively likely to be different than the mean at any other point in the chain. When c is less than 1, the Geweke and Heidelberger and Welch diagnostics only deem a chain non-converged at chance levels. The Raftery and Lewis diagnostic never deems a chain non-converged when c is less than 1. All of these methods are sensitive to both bias and variance (Cowles & Carlin, 1996). However, these methods appear to be insensitive to the patterns of rank orderings among elements.

The evidence provided by simulation study 4 is tempered by the fact that it is not entirely clear which of these chains would be deemed converged or not. That is, there is no clear distinction between what is and is not converged except for the extreme conditions of c equal to 0 and c equal to 1 as converged and non-converged, respectively. When c was not equal to 0 or 1, the simulated chains tended to settle into locations as influenced by the experimental conditions. This 'settling' into fairly well-defined boundaries seems to satisfy the three existing methods, but the new method ignores the settling and focuses on the patterns of rank orderings.

The results of simulation study 4 need to be interpreted in light of the fact that the simulated chains were essentially converged. The new method being developed is sensitive to patterns of rank orderings and autocorrelations. The three existing methods included in this study are sensitive to the stability of the mean for the chains. The discrepancies among the existing techniques and the new technique may be largely due to the interplay of the method used to simulate chains and the way that the diagnostics characterize convergence.

Strengths and weaknesses

Strengths

The strengths and weaknesses of the current method will now be discussed. It is not always clear what characteristics of the new method are strengths or weaknesses, so every attempt will be made to characterize the new method fairly. One of the reasons why the interpretation of the findings is not always clear has to do with uncertainty regarding the simulated chains. As it is not the case that the simulated chains were completely

understood at the outset of these studies, any attempt to draw inference based on the simulated chains must be tempered with caution. Another reason that interpretation of the findings is not straightforward is that the methods employed in this study were focused specifically on the research questions, and the research questions were not aimed at answering all possible questions of interest. Many unforeseen nuances of interest were uncovered during the course of this research, but unfortunately not all can be addressed.

To begin, a potential strength of the new method is that it is fundamentally unlike any of the methods it was compared to in this study. The current method is sensitive to the pattern of rank orderings of elements, and the other methods are not. In addition, the new method tends to disagree with the other methods in characterizing the chains simulated in this study. The new method deems all chains with any degree of dependence to be non-converged (using .67 as the expected value of D_t). It is because the new method is sensitive to particular patterns of rank orderings of the elements in the chain.

The current method provides information about the particular pattern of rank orderings present in the chains. This characterization of chains is unique among convergence diagnostics. As such, it is providing information that other diagnostics simply do not provide. It raises an interesting question about the fundamental nature of what we mean by convergence in MCMC estimation.

For the sake of example, if we have a chain that stays within reasonable bounds for a given parameter but is either: 1) constantly wandering up for say 20 iterations, and then down and up again, or 2) shows a pattern of rank orderings similar to what would be expected if they were just random draws from a distribution, do we call both of these

chains converged? Probably so for the second case, and probably not for the first case. Diagnostics like the one Geweke proposed offer no clear direction that helps users to make a distinction between these two general types of chains. However, D_t is able to distinguish between chains like the ones just described.

When a chain is being produced from a stationary distribution with no degree of linear dependence among elements, it is clear what value of $\underline{D_t}$ to expect. However, when dealing with chains from real samplers, it is not clear exactly how to proceed with this technique. When it is the case that thinning of chains isn't so restrictive as to remove the vast majority of sampled values, then this technique provides an alternative to the existing techniques.

The way the indicator statistic is defined allows the method to be sensitive to a case where the location of the chain is fluctuating up (and/or down) more often than would be expected. The other methods treat convergence from different standpoints. For example, the Geweke (1992) diagnostic compares the mean from the first tenth of the chain to the last half of the chain. When applied to the simulated chains in this study, it deemed most chains converged.

The stationarity test of Heidelberger and Welch (1983) also is relatively unlikely to classify the simulated chains as non-converged. The HW diagnostic is also essentially focused on the mean of the chain and attempts to determine when the transient phase at the beginning of a Markov Chain has passed.

The Raftery and Lewis (1992) diagnostic focuses on the accuracy of estimating user-specified quantiles of the target distribution of interest. In this study, the quantile

that was specified was .5. This quantile was chosen because it represents the center of the distribution. The RL method essentially boils down to estimating the mean and variability of the distribution.

So, each of the existing methods to which the new method is being compared is focused on a fundamentally different aspect of the behavior of the chains. Thus, it is not surprising that the three existing diagnostics behave similarly to one another, and that each tends to disagree with the new method. The new diagnostic takes a new perspective on convergence.

At this point, it has not been determined whether the new method is an improvement or not over the existing methods. It could be an improvement by virtue of its sensitivity to the patterns of rank orderings within the chains. No other diagnostics considered have this kind of sensitivity, so the new method may be providing a useful new characterization of convergence. However, perhaps the new technique is simply sensitive to autocorrelation in the chains. Autocorrelations are known to go hand in hand with MCMC samplers. Just because there is dependence among elements, it does not mean that there is lack of convergence. That is to say, the new method may be sensitive to something that is not informative about the convergence of Markov chains.

Weaknesses

One of the primary weaknesses of this set of studies is its limited scope. There are many questions of interest that are not addressed by these research questions. Future research will address the limited scope of the current set of research questions.

The primary limitation of this set of studies is that it provides only limited information on how to proceed in using the new method for the assessment of convergence. One of the reasons why the current study provides only limited information is that it contains only a few simple simulation studies. If the method is to be used in practical settings, it needs to be further refined so there are clear guidelines for how to apply it. There is also a need to develop more complex simulation studies to address the usefulness of the technique. Also, it would be very informative to apply the method to chains from real samplers. Also, the method needs to be applied to real datasets.

Another reason why the current set of studies only provides limited information is that there was a discrepancy between what was expected and what was observed with the solutions for the value of the summary of the indicator statistic. The analytical and simulated solutions do not match. The reason for the disagreement likely lies in the fact that the boundaries specified for the simulated solutions don't accurately reflect the assumptions of the analytical solutions.

Also, although the simulated chains behaved somewhat as expected (e.g., as c increased so did the amount of autocorrelation), it may be the case that elements of chains from real MCMC samplers such as those created using the MH algorithm have a fundamentally different relationship to one another which is defined by the algorithm. Thus, the reason why elements are associated with one another in real chains may be due to a fundamentally different mechanism than those provided by the simulated chains. If this is the case, then the simulated solutions provided only limited insight into the behavior of the method being developed here. The only remedy for this situation is to

apply the method to real chains from real samplers. The one study contained herein using real samplers provided a glimpse of all the issues remaining to be resolved. For example, the specification of the variability of the proposal distribution can greatly affect the chains produced by the sampler. Based just on the very limited scope of the study with real samplers in this paper, if the proposal distribution has too great a variance, then the chains will have a great deal of autocorrelation over very long lags. Chains of this type would immediately pose problems for applying the current method. Also, it is well known that any dependence among parameters in the model can result in interdependencies among chains. This issue was not addressed at all by the current research, but is certainly germane to the use of the method.

Another limitation has to do with how the method characterizes the chains. The indicator statistic provides information about the pattern of rank orderings of the chains. However, as currently defined the indicator statistic is not capable of making any claims about convergence to a location. The indicator statistic is a simple way to describe the pattern of rank orderings of the chain elements. In this way, the new method is an interesting and possible useful alternative to the existing methods. It is probably best to pair this technique with other methods that do provide information about the particular location of chain elements.

Finally, the most serious limitation of this technique is that is deals with convergence in a post hoc fashion. This limitation is not unique to the current method, but it is a limitation nonetheless. Ideally, it would be best to be able to specify ahead of time how long a chain should be allowed to run to achieve convergence. As there is no

currently accepted method to determine this, convergence diagnostics will all remain 'fundamentally flawed'.

Future directions

Much work still remains to be done in order for this technique to become a viable way of assessing convergence of MCMC samplers. This work falls into three general categories; operational, graphical, multivariate.

In order for the technique to become operational, the issue of thinning needs to be settled. It has been shown by other authors to be undesirable to thin, as it goes against the practice of obtaining the best estimate possible. If thinning is to continue, guidelines must be provided for exactly when and how to thin. For example, how much dependence among elements is too much before thinning would require the removal of too many elements? In the studies presented here, even a mild amount of autocorrelation led to the reduction of the initial chain by 1/2 or 1/3. When there was a substantial amount of autocorrelation over longs lags, the thinning could lead to reductions of the chain by 1/30 or more. This large degree of thinning is clearly counterproductive to applying the method in practice. MCMC samplers tend to run for a long time to produce chains, so throwing away the vast majority of the observations seems especially wasteful.

In order to avoid thinning, it seems necessary to know how the autocorrelation present can influence the value of the summary of the indicator statistic. However, other factors can influence the value of $\underline{D_t}$ besides the amount of autocorrelation present in the chains. The analytical and simulated solutions for the value of $\underline{D_t}$ presented herein do not reveal a simple relationship between autocorrelation, balance/stability of the chain, and

the value of $\underline{D_t}$. Therefore, statements about the utility of the method are limited. To use the method without thinning, we would expect the value of $\underline{D_t}$ to be lower than .67, but by how much? Also, if we see a decrease in the value of $\underline{D_t}$, is it due to autocorrelation or instability? Some work remains to be done to answer these questions.

Also, one of the aspects of this method was largely overlooked in order to provide information to the basic research questions presented. This method can be represented graphically. It is hoped that the combined use of graphical and quantitative representations of the chains can provide more information than either method alone. What this method lacks in sensitivity to location can be easily informed by path plots and CUSUM plots. Just because the method under development is no longer tied to Cumulative Sums doesn't mean that this aspect should be overlooked or disregarded. The purpose of the current research was to learn about the potential of this method to characterize Markov chains. To this end, the current research has met with some success, however, to fully benefit from all this technique has to offer, further work needs to be done to integrate the strengths of this technique with other existing methods. It is unwise to develop a method that does not complement and add to existing methods. The CUSUM path plots are indicative about the behavior of the chain in reference to the mean. This type of characterization contains information that the current method does not.

The gold standard of convergence diagnostics is the Multivariate Potential Scale Reduction factor (Gelman and Rubin, 1992). This method takes into account the relationships among the parameters that the chains represent. In this way, any dependence

among parameters that exists is incorporated into the assessment of convergence. Following in this fashion, it would be ideal to expand the method to be used with multiple chains. Assuming that the chains are independent is probably incorrect in many practical applications. Also, investigation of the cross-correlations may be informative for the purpose of thinning.

Also, Brooks (1998c) suggested that this type of method could be incorporated into the MCMC sampling procedure. Essentially, as the chain is being produced by the sampler, it would also be analyzed by way of the indicator statistic. In this way, the sampler could be programmed to run long enough to satisfy some pre-specified criterion chosen by the user. Once the chain satisfies the criterion, the sampler could be terminated. A note of caution here is appropriate, though. It is generally unwise to try to automate the assessment of convergence for MCMC samplers (Cowles & Carlin, 1996).

Also, something must be done to deal with the issue of ties in the sampler. As the diagnostic is currently defined in terms of strict inequalities, any ties will be coded as zeroes. Brooks (1998c) suggests setting ties equal to the expected value of the indicator statistic. Research needs to be done to verify if this is wise. MCMC samplers built using the MHA are known to be effective, even though the fact that they are rejection samplers means ties are common. To make this method applicable to MHA samplers, something needs to be done to address how the method should deal with ties. An alternative to Brooks' (1998c) modification of dealing with ties is to redefine the indicator statistic by relaxing the idea of strict inequalities. Although, relaxing the strict inequalities of the indicator statistic would mean that samplers like those produced in simulation study 3 for

RATIO equal to 4 would set nearly every element to 1. There is much research needed to address this issue.

Thinning is a big issue to resolve. The issue was only indirectly broached by the current set of studies. The studies contained in this paper provided evidence that agreed with predictions about thinning's effects on the values of $\underline{D_t}$ achieved. When a chain with any degree of autocorrelation was thinned to achieve a linearly independent sequence, the result was that the value of $\underline{D_t}$ increased when applied to the thinned chain. A much more thorough investigation of the effect of thinning on the quality of the estimates obtained from chains would be needed to make further statements about the appropriateness of thinning as a practice in the assessment of convergence.

Relevance

The convenience of MMLE typically makes it preferable to MCMC when both techniques are easily implemented. The advantage of MCMC becomes apparent when models become overly complex and/or highly dimensional. In cases where the psychometric model in question is complex, it is not always possible to implement MMLE procedures. MCMC techniques are much more easily implemented when the complexity of the model in question increases. Where the current technique may find some usefulness is with models that are formulated to represent highly multidimensional constructs. Examples would include Diagnostic Classification Models (DCMs) such as the Log-linear Cognitive Diagnostic Model (LCDM; Henson, Templin and Willse, 2007).

When MCMC techniques become the only or best way to obtain parameter estimates, it is necessary to check the quality of the chains produced. While there are

currently many diagnostics that exist for use with MCMC procedures, the current technique clearly provides information different than the other three techniques considered in this paper. While this paper does not speak to the overall efficacy of the technique under development, it does raise the question of the usefulness of the technique for assessing convergence. It is necessary to investigate the technique further to determine if it is useful for assessing the quality of estimates obtained in situations where the more accepted MMLE approach isn't feasible. This paper is a first step in a sequence of directed research to address the utility of the new diagnostic.

REFERENCES

Albert, J. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. Journal of Educational Statistics, 17, 251-269.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. Psychometrika, 66(4), 541–561.

Brooks, S. P. (1998). Quantitative convergence diagnosis for MCMC via CUSUMS. Statistics and Computing, 8, 267-274.

Brooks, S.P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4), 434-455.

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis Hastings Algorithm. The American Statistician, 49(4), 327.335.

Cowles, M. K., and Carlin, B. P. (1996), 'Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,' Journal of the American Statistical Association, 91, 883-904.

De la Torre, J., Stark, S., and Chernyshenko, O. (2006). Markov Chain Monte Carlo Estimation of Item Parameters for the Generalized Graded Unfolding Model. Applied Psychological Measurement, 30(3), 216-232.

Fox, J. P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. Psychometrika, 68(2), 169–191.

Glas, Cees A. W., Meijer, Rob R. (2003). A Bayesian Approach to Person Fit Analysis in Item Response Theory Models. Applied Psychological Measurement, 27 (3) 217-233.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika, 72, 711-732.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. Journal of Educational Measurement, 26, 333-352.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Academic Publishers.

Hanson, and Cunningham (1998). Posterior sampling with improved efficiency. In Medical Imaging: Image Processing, K. M. Hanson (ed.), pp. 371-382.

Hastings, W. K. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, 57, 97-109.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. Psychometrika, 55(4), 577–601.

Kim, J. S., Bolt, D.(2007). Estimating Item Response Theory Models Using Markov Chain Monte Carlo Methods. Educational Measurement v. 26 no. 4 (Winter 2007) p. 38-51.

Lin, Z. Y. (1992) On the increments of partial sums of a phi-mixing sequence. Theoretical Probability and its Applications, 36, 316-326.

McLeod, L., Lewis, C., and Thissen, D.(2003). A Bayesian Method for the detection of Item Pre-knowledge in Computerized Adaptive Testing. Applied Psychological Measurement, 27 (2), 121-137.

Meijer, R., and van Krimpen-Stoop, E. (2003). The Use fo Statistical Process Control Charts for Person Fit Analysis on Computerized Adaptive Testing. LSAC Research Report Series.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), 'Equations of State Calculations by Fast Computing Machines,' Journal of Chemical Physics, 21, 1087-1092.

Page, E.S. (1954). Continuous inspection schemes. Biometrika, 41(1-2), 100-115.

Patz, R., & Junker, B. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. Journal of Educational and Behavioral Statistics, 24, 146–178.

Patz, R., & Junker, B. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. Journal of Educational and Behavioral Statistics, 24, 342–366.

Sinharay, S. (2004). Experiences with Markov Chain Monte Carlo convergence assessment in two psychometric examples. Journal of Educational and Behavioral Statistics, 29(4), 461-488.

von Neumann, J. (1951). Various techniques used in connection with random digits. Monte Carlo Methods. National Bureau of Standards, 12, pp. 36-38.

Yu, B. (1995) Discussion to Besag et al. Statistical Science, 10, 3-66.

Yu, B. and Mykland, P. (1994) Looking at Markov sampler through cusum path plots: a simple diagnostic idea. Technical Report 413, University of California at Berkeley, Dept. of Statistics.

Yu, B. and Mykland, P. (1998) Looking at Markov sampler through cusum path plots: a simple diagnostic idea. Statistics and Computing, 8, 275-286.