# Methods in Hypothesis Testing, Markov Chain Monte Carlo and Neuroimaging Data Analysis

# Methods in Hypothesis Testing, Markov Chain Monte Carlo and Neuroimaging Data Analysis

A dissertation presented

by

Xiaojin Xu

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

May 2013

**Professors Joseph K. Blitzstein and Xiao-Li Meng** **Xiaojin Xu**

**Methods in Hypothesis Testing, Markov Chain Monte Carlo and Neuroimaging Data Analysis**

# Abstract

This thesis presents three distinct topics: a modified K-S test for autocorrelated data, improving MCMC convergence rate with residual augmentations, and resting state fMRI data analysis.

In Chapter 1, we present a modified K-S test to adjust for sample autocorrelation. We first demonstrate that the original K-S test does not have the nominal type one error rate when applied to autocorrelated samples. Then the notion of mixing conditions and Billingsley's theorem are reviewed. Based on these results, we suggest an effective sample size formula to adjust sample autocorrelation. Extensive simulation studies are presented to demonstrate that this modified K-S test has the nominal type one error as well as reasonable power for various autocorrelated samples. An application to an fMRI data set is presented in the end.

In Chapter 2 of this thesis, we present the work on MCMC sampling. Inspired by a toy example of random effect model, we find there are two ways to boost the efficiency of MCMC algorithms: direct and indirect residual augmentations. We first report theoretical investigations under a class of normal/independece models, where we find an intriguing phase transition type of phenomenon. Then we present an application of the direct residual augmentations to the probit regression, where we also include a numerical comparison with other existing algorithms.

In Chapter 3, we present a statistical analysis of resting state fMRI data. The functional connectivity, which can be measured as spatial correlation (or partial correlation), is of great interest to the researchers. In the literature, the default estimator is the standard sample correlation matrix obtained by ignoring the temporal correlation (autocorrelation). We propose fitting the covariance separable model and using the MLE. The relative efficiency of the MLE is demonstrated both in theory and simulations. We also propose an empirical Bayes model and fit it to the resting state fMRI data.

# Contents

# Acknowledgments

First and foremost, I would like to thank Professor Joseph K. Blitzstein and Professor Xiao-Li Meng, without whom this dissertation would have only been a dream. They have been always advising and supporting me along the way. The invaluable guidance from them has not only helped me learn Statistics, but also helped me become an independent researcher. Outside the field of Statistics, they have also been the mentors and role models of my life. I feel greatly honored to be their students. I also owe tremendously to my other committee member, Professor Samuel Kou. Professor Kou guided me in the resting state fMRI project. He taught me the importance of doing statistical research in principles and always focusing on the scientific goal in the domain. I am truly grateful to him.

I also want to take this opportunity to thank Professor Yaming Yu, Professor Nanyin Zhang and Zhifeng Liang, for their collaboration. Their experience and help is invaluable to the completion of my thesis. I also want to show my sincere gratitude to other faculty members within the Department of Statistics: Professor Jun Liu, Donald Rubin, Carl Morris, David Harrington and Tirthankar Dasgupta. It is my great honor to learn from and work with them. I would also like to extend my thanks to the department staff: Betsey Cogswell, James Matejek, Dale Rinkel, Maureen Stanton and Ellen Weene, who helped me in a lot of aspects during the five years.

Meanwhile, I can not imagine going through the Ph.D. study without the help and accompany of my friends and fellow students: Kevin Rader, Kari Lock, Zhan Li, Xianchao Xie, Martin Lysy, Sergiy Nesterko, Li Zhu, Bo Jiang, Simeng Han, Ke Deng, Thomas Tong, Samuel Wong, Nathan Stein, Valeria Espinosa, Jonathan Hennessy, Joseph Kelly, Daniel Fernandez to name a few.

Last but most importantly, I would like to thank my parents and my wife, for their unreserved love and support. This journey has eventually become real because of you.

*To my wife Yuan.*

# Chapter 1

# A modified K-S test for autocorrelated data – ESS adjustment

## 1.1 Introduction

The Kolmogorov-Smirnov (K-S) test has been used to compare a sample with a reference probability distribution (one-sample K-S test), or to compare two samples (two-sample K-S test). This is a very widely used nonparametric test. However, critical values of the K-S statistics are sensitive to sample autocorrelation. The actual Type I error rate tends to be much bigger when there exists a positive sample autocorrelation. As a result, direct use of the K-S test might give misleading result in an autocorrelated sample.

For example, in studying fMRI data, researchers have proposed the use of the

K-S statistic to test normality. However, Aguirre et al. (2005) pointed out that the K-S test is an invalid statistical test for use with BOLD fMRI: the K-S test does not control the false-positive rate in BOLD fMRI data mainly due to the violation of the independence assumption. So indeed the use of the K-S test with non-independent data is problematic. In this paper we suggest a modified K-S test to correct it.

We first consider the simplest autocorrelated process – AR(1) process with Normal stationary distributions. Based on generalized Billingsley's Theorem, the asymptotic distribution of the K-S statistic could be derived in theory. Based on it, we use Monte Carlo method to estimate the critical values from the distribution. And then a modified K-S test with effective sample size (ESS) adjustment is constructed, with ESS formula $n_{ESS} = n(1 - \rho)$, where $\rho$ is the coefficient (or estimated coefficient) of the underlying AR(1) process. It is illustrated in this paper that this modified K-S test not only applies to data generated by AR(1) processes, but also applies to data that can be reasonably fitted by such AR(1) processes. Several simulation studies indicate that it has its nominal significance level (being slightly conservative) as well as reasonable power. Therefore it could be used for a general autocorrelated sample, though it's inspired by AR(1) processes.

Further more, this test can be easily generalized to two sample test. In two sample test, the effective sample sizes are calculated for two samples respectively. Then one simply plugs them instead of the actual sample sizes into the test. It is very useful since it provides a way of comparing two samples even if there is autocorrelation within each sample. Hence it's a big advantage of our method over some other existing method.

The structure of this chapter is as follows: we introduce the K-S test in Section 1.1.1. A simple example is included in Section 1.2 to illustrate the influence of sample autocorrelation. For an AR(1) process with positive coefficient $\rho$, the actual Type I error rate tends to increase as $\rho$ increases. As a result, we need to adjust critical values of test statistics. However, the asymptotic distribution of the K-S statistic of autocorrelated data is much more involved than that of independent data. With results in empirical process literature, especially the generalized Billingsley's Theorem, we characterized the asymptotic distribution of the K-S statistic for AR(1) processes with normal stationary distribution. Then a modified K-S test is constructed on it, using effective sample size adjustment. It's very easy to implement since the only difference from the original K-S test is to replace the actual sample size by the effective sample size $n_{eff}$. In Section 1.5, we implement this modified K-S test to several simulated data: AR(1) process, MA(2) process, Metropolis algorithm and two sample test. In order to make comparison with, we also implement a naive "thinning" K-S test, which is explained in Section 1.5.1. Simulations on AR(1) process indicate that our modified K-S test has almost the nominal significance level (being slightly conservative with actual Type I error rate around 0.04) as well as reasonable power (bigger than the "thinning" K-S test). It is not surprising since it is derived from AR(1) processes. More importantly, the results on MA(2) process, Metropolis algorithm are also promising. They demonstrate that this method can be used to a general autocorrelated sample as well. Further more, we also include a simulation study about two sample test in the end. A summary of Type I error rate and power is illustrated for all simulations. Section 1.6 provides an application of this method to fMRI data.

### 1.1.1 Introduction to the K-S test

Comparing a sample to some specified distribution is frequently met in statistics. The K-S test has been proposed by Kolmogoroff (1933) and Smirnov (1939) to solve this problem. This nonparametric test is constructed upon the asymptotic distribution of the Kolmogorov-Smirnov (K-S) statistic. More specifically, suppose our null hypothesis is that $n$ i.i.d. observations $X_i, i = 1, 2, \cdots, n$ follow a distribution $F(x)$. Define the empirical distribution function $F_n$ as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leqslant x}.$$

The K-S statistic for this given cumulative distribution function $F(x)$ is defined as:

$$D_n = \sup_x |F_n(x) - F(x)|.$$

When $F$ is continuous, Kolmogorov has shown that under the null hypothesis

$$\sqrt{n} D_n \xrightarrow{n \to \infty} \sup_{t \in [0,1]} |B(t)|$$

in distribution, where $B(t)$ is the Brownian bridge. Noticing that $\sup_{t \in [0,1]} |B(t)|$ (named Kolmogorov distribution) is independent of the distribution $F$, the K-S test compares the statistic $\sqrt{n} D_n$ to some quantile of the Kolmogorov distribution.

For two sample test, where the null hypothesis is that two i.i.d. samples (with

sample sizes $n_1$ and $n_2$ respectively) follow the same distribution, the K-S statistic is:

$$D_{n_1,n_2} = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)|,$$

where $F_{1,n_1}$ and $F_{2,n_2}$ are the empirical distribution functions respectively. Smirnov has shown that under the null hypothesis,

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}} D_{n_1,n_2} \xrightarrow{n_1,n_2 \to \infty} \sup_{t \in [0,1]} |B(t)|$$

in distribution. So similarly, the test is constructed upon the asymptotic distribution of $D_{n_1,n_2}$.

## 1.2 Influence of autocorrelation on the K-S test

A fundamental assumption of the K-S test, independent sampling, is sometimes violated in application. If we still want to apply the K-S test, the first question is whether it is still a valid test (has its nominal significance level). A simple example of autocorrelated data is a first-order autoregressive (AR(1)) process. Suppose $\{X_n\}$ are generated from the following process:

$$X_1 \sim N(0,1) \tag{1.1}$$

$$X_n = \rho X_{n-1} + \sqrt{1-\rho^2} Z_n, n = 2, 3, \cdots \tag{1.2}$$

where $Z_n$ follows i.i.d. standard normal distribution and is independent of $X_j, j = 1, 2, \cdots, n-1$. The underlying stationary distribution of this process is $N(0,1)$. If

$\rho = 0$, they are essentially i.i.d. observations; if $\rho \neq 0$, they are autocorrelated. Suppose our null hypothesis is that the stationary distribution is $N(0, 1)$, is the K-S test still valid in this case? The following simulation study answers this question: for a chosen $\rho$, 5000 such processes were independently simulated and the K-S test (significance level $\alpha = 0.05$) was conducted on each sample. The rejection rate (type I error rate) was recorded for each $\rho$. We chose 15 different $\rho$'s from $-1$ to 1 and different sample sizes ($n = 100, 200, 500$) to see the effect of sample size. Below is the summary plot:

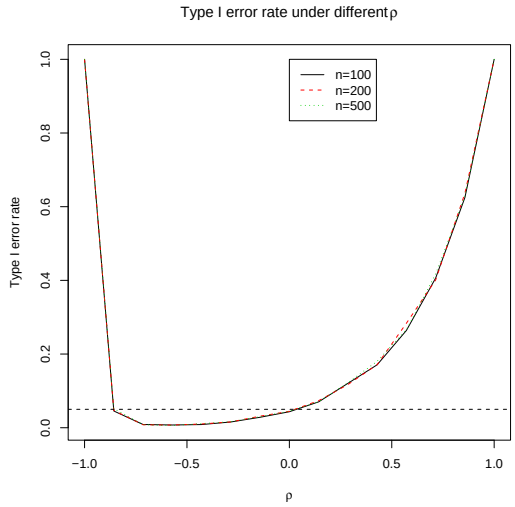


Figure 1.1: Type I error rate for autocorrelated data

There are several things that can be observed from this graph:

- This curve is not symmetric: it is not monotone when $\rho < 0$. However, since $\rho$ is positive in most of the problems we want to solve, we ignore the part $\rho < 0$ in the discussion.

- When $\rho > 0$, the rejection rate is monotonely increasing as $\rho$ increases. If

samples are positively autocorrelated, we expect a larger K-S statistic than i.i.d. sample, which leads to higher rejection rates. Therefore if we still want to apply the K-S test, the critical value of the K-S statistic should be adjusted with $\rho$.

- There is little influence of sample size. This ensures us it is sample autocorrelation that breaks the K-S test.

As a summary, the K-S test is not necessarily valid if there is an autocorrelation within the sample. It has a larger rejection rate than its nominal type I error rate when $\rho > 0$. In order to correct it, we need to find out the asymptotic distribution of the K-S statistic when the i.i.d. assumption is violated. As a first step, we go back to see what breaks down in the K-S test if there exists an autocorrelation.

There are two key facts in deriving the asymptotic distribution of the K-S statistic (Doob (1949)):

- Distribution free property:

$$\begin{aligned} F_n(x) - F(x) &= \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x} - F(x) && (1.3) \\ &= \frac{1}{n}\sum_{i=1}^{n} I_{F(X_i) \le F(x)} - F(x) && (1.4) \\ (t \triangleq F(x)) &= \frac{1}{n}\sum_{i=1}^{n} I_{U_i \le t} - t && (1.5) \end{aligned}$$

  where $U_i$ are independent uniform random variables and hence the quantity is independent of the distribution $F(\cdot)$.

- The above function converges weakly to a Gaussian process. This is proved by

functional central limit theorem for i.i.d. random variables.

However, if there exists an autocorrelation within the sample, we can still rewrite $F_n(x) - F(x)$ in terms of uniform random variables, but those random variables are not independent any more. One natural question is if there is some corresponding functional central limit theorem for sum of correlated random variables. The literature of empirical process gives some sufficient conditions and we cite some useful results in the next section.

## 1.3 Functional central limit theorem of empirical process

In this part, we cite some useful theorems in studying functional central limit theorem of empirical process. Based on these theorems, the asymptotic distribution of the K-S statistic for an AR(1) process with Normal stationary distribution is derived theoretically. Section 1.3.1 introduces Billingsley's theorem and its generalization in proving weak convergence of empirical CDF; Section 1.3.2 verifies that AR(1) process with Normal stationary distribution satisfies conditions in Billingsley's theorem and hence we can theoretically derive the asymptotic distribution of the K-S statistic.

### 1.3.1 Billingsley's theorem and its generalization

Let's first introduce some widely used concepts in studying weak convergence. We will use notations by Bradley (Bradley (2005)): let the probability space be $(\Omega, \mathcal{F}, P)$. For any $\sigma$-field $\mathcal{A} \subset \mathcal{F}$, let $\mathcal{L}^2(\mathcal{A})$ be the space of square-integrable, $\mathcal{A}$-measurable

random variables. For any two $\sigma$-fields $\mathcal{A}$ and $\mathcal{B} \subset \mathcal{F}$, define the following measures of dependence:

$$\begin{aligned} \alpha(\mathcal{A}, \mathcal{B}) &:= \sup |P(A \cap B) - P(A)P(B)|, A \in \mathcal{A}, B \in \mathcal{B}; && (1.6) \\ \rho(\mathcal{A}, \mathcal{B}) &:= \sup |\mathrm{cor}(f, g)|, f \in \mathcal{L}^2(\mathcal{A}), g \in \mathcal{L}^2(\mathcal{B}); && (1.7) \\ \phi(\mathcal{A}, \mathcal{B}) &:= \sup |P(B|A) - P(B)|, A \in \mathcal{A}, B \in \mathcal{B}, P(A) > 0. && (1.8) \end{aligned}$$

These coefficients measure the dependence between two $\sigma$-fields in different aspects, but there are also some connections between them. For instance, one useful relationship between $\alpha$ and $\rho$ is the following inequality (Bradley (2005)):

$$4\alpha(\mathcal{A}, \mathcal{B}) \leq \rho(\mathcal{A}, \mathcal{B}).$$

Moreover, mixing conditions for a sequence of random variables can be defined as follows: suppose $\{X_k : k \in \mathbb{Z}\}$ is a (strictly) stationary sequence of random variables. For $-\infty \leq J \leq L \leq \infty$, define the $\sigma$-field

$$\mathcal{F}_J^L := \sigma(X_k, J \leq k \leq L).$$

For each $n \geq 1$, define the following dependence coefficients:

$$\begin{aligned} \alpha(n) &:= \alpha(\mathcal{F}_{-\infty}^0, \mathcal{F}_n^\infty) && (1.9) \\ \rho(n) &:= \rho(\mathcal{F}_{-\infty}^0, \mathcal{F}_n^\infty) && (1.10) \end{aligned}$$

The random sequence $\{X_k\}$ is said to be

"$\alpha$ mixing" (or "strong mixing") if $\alpha(n) \to 0$ as $n \to \infty$,

"$\rho$ mixing" if $\rho(n) \to 0$ as $n \to \infty$,

"$\phi$ mixing" if $\phi(n) \to 0$ as $n \to \infty$.

If $\{X_k\}$ is furthermore a Markov Chain, the mixing coefficients become:

$$\alpha(n) := \alpha(\sigma(X_0), \sigma(X_n)), \tag{1.11}$$

$$\rho(n) := \rho(\sigma(X_0), \sigma(X_n)), \tag{1.12}$$

$$\phi(n) := \phi(\sigma(X_0), \sigma(X_n)). \tag{1.13}$$

These different types of mixing conditions characterize the dependence between observations with lag $n$. Recall from last section that the functional central limit theorem is used in the second step of deriving the asymptotic distribution of the K-S statistic for "independent" case. There exists parallel functional central limit theorem for "dependent" random variables under some above mixing conditions. With these mixing conditions satisfied, the asymptotic distribution of the K-S statistic can therefore be derived and used to construct a hypothesis test. Billingsley (2009) established one of the most fundamental results:

**THEOREM(Billingsley).** Suppose now that $0 \leq U_0 \leq 1$, and $U_0$ has a continuous distribution function $F$ on $[0, 1]$. Let $\{F_n(t) : 0 \leq t \leq 1\}$ be the empirical process for

$U_1, U_2, \cdots, U_n$, i.e., $F_n(t) = n^{-1} \sum_{i=1}^{n} I_{[0,t]}(U_i)$. Normalize $F_n(t)$ as

$$Y_n(t) = \sqrt{n}(F_n(t) - F(t)),$$

where $0 \leq t \leq 1$. Further more, define

$$g_t(x) = I_{[0,t]}(x) - F(t).$$

Let $\{U_n : n = 0, 1, 2, \cdots\}$ be defined as above and suppose $\{U_n\}$ satisfies the mixing condition:

$$\sum_{n \geq 1} n^2 \phi(n)^{\frac{1}{2}} < \infty. \tag{1.14}$$

Then the sequence $\{Y_n(t) : 0 \leq t \leq 1\}$ of normalized empirical processes converges weakly in $D[0, 1]$ to a Gaussian random function $\{Y(t) : 0 \leq t \leq 1\}$ specified by

$$E[Y(t)] = 0$$

and

$$E[Y(s)Y(t)] = E[g_s(U_0)g_t(U_0)] + \sum_{k=1}^{\infty} E[g_s(U_0)g_t(U_k)] + \sum_{k=1}^{\infty} E[g_s(U_k)g_t(U_0)]. \tag{1.15}$$

Furthermore, the series above converges absolutely and the sample paths of $Y$ are continuous with probability one.

The key message of this theorem is that as long as the mixing condition (1.14) is satisfied, the corresponding empirical process converges weakly to a Gaussian process

with covariance function specified by (1.15). This is a fundamental theorem, providing a condition when the functional central limit theorem holds for dependent variables. Followed by it, Deo (1973) generalized the result to $\alpha$-mixing condition as follows: Billingsley's theorem remains true if the condition (1.14) is replaced by

$$\sum_{n\geq 1} n^2 \alpha(n)^{\frac{1}{2}-\tau} < \infty \tag{1.16}$$

for some $0 < \tau < \frac{1}{2}$.

In the next section, we verify that the AR(1) process with Normal stationary distribution satisfies condition (1.16).

### 1.3.2 Weak convergence in AR(1) process

We want to show that for the process defined in (1.1), the K-S statistic asymptotically converges to the supremum of some Gaussian process and then we can use that Gaussian process to calculate critical values of the K-S test. Recall that if $X_n = \rho X_{n-1} + \sqrt{1-\rho^2} Z_n$ has $N(0,1)$ as its stationary distribution, $U_n := \Phi(X_n)$[1] has a marginal uniform distribution. We can verify that $\{U_n\}$ satisfies the mixing condition (1.16) given by the generalization of Billingsley's theorem (see Appendix). Thus it converges weakly to a Gaussian process specified in the theorem. The covariance function (1.15), although not analytically expressible, can be calculated numerically. With the corresponding Gaussian process simulated, the critical value of the K-S statistic can be estimated from the corresponding quantile.

[1] $\Phi(\cdot)$ is the CDF of $N(0,1)$.

For a chosen $\rho$, a multivariate normal distribution with dimension 200 was simulated to approximate the Gaussian process. Its covariance matrix was numerically calculated by the form (1.15). This was repeated 2000 times to estimate the 95% quantile of the supreme absolute value of the corresponding Gaussian process. The following plot shows the results:

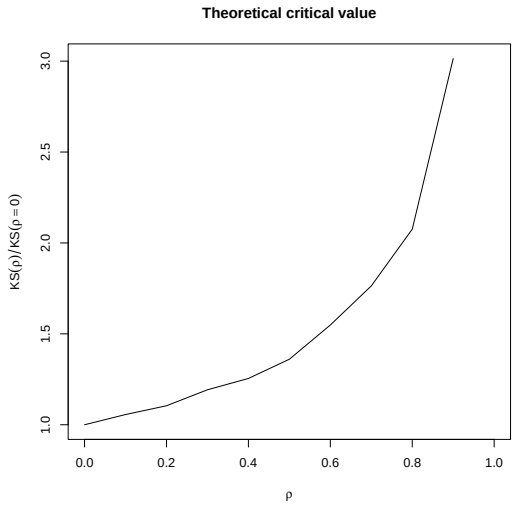


Figure 1.2: Theoretical critical value

The black line represents the estimated critical values for different coefficient $\rho's$.

## 1.4 A modified K-S test

Given these critical values in Figure 1.2, one can construct a test, at least for AR(1) process with normal stationary distribution. However, these critical values can only be computed through simulations and are not easy to use or interpret. Moreover, this method only applies to data generated from an AR(1) process with normal stationary distribution. We hope to find some more general method.

Surprisingly, it could be found that a simple function $\frac{1}{\sqrt{1-\rho}}$ fits the curve of the critical values simulated above. Here is a plot of the theoretical values and fitted values:

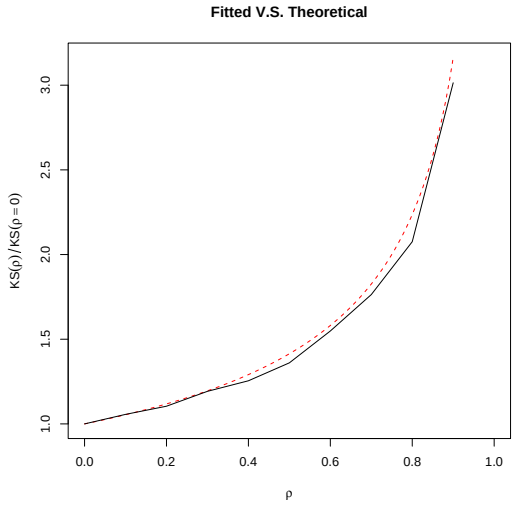


Figure 1.3: Fitted and theoretical values

The black line represents the theoretical critical values (though derived through simulation) for different $\rho's$; the red dashed line represents the function $\frac{1}{\sqrt{1-\rho}}$ used to calculate effective sample size. The difference between these two lines, measured in terms of the proportion $\frac{|\frac{KS(\rho)}{KS(\rho=0)} - \frac{1}{\sqrt{1-\rho}}|}{\frac{KS(\rho)}{KS(\rho=0)}}$ has a maximum 15.6% with all the others below 10%[2]. This difference won't allow our modified K-S test has exactly the nominal significance level, but it's acceptable as we will see later. Based on this fitted function, the modified K-S test can be constructed as follows: for an AR(1) process $X_n = \rho X_{n-1} + \epsilon_n$, where $\{\epsilon_n\}$ are i.i.d. error terms which are not necessarily normal, the critical value $KS(\rho)$ corresponding to a 0.05 level test is estimated by the product of

[2]The median difference is 6.4%.

$\frac{1}{\sqrt{1-\rho}}$ and $KS(\rho = 0)$. Then the rejection area is determined as $(KS(\rho), \infty)$. Since the estimated critical value is close to the empirical 95% quantile of the K-S statistic, the type I error rate of the modified K-S test should be close to 0.05 (simulation study in Section 1.5.1 confirms that it is actually smaller than 0.05 most of the time).

### 1.4.1 ESS interpretation

This modified K-S test not only has its nominal significance level (being slightly conservative), but also has a nice interpretation of effective sample size (ESS) adjustment. Since the sample has an autocorrelation, the information contained in the data is (usually) less than an i.i.d. sample with the same size. In other words, the number of equivalent independent observations is fewer than the sample size $n$. If we can somehow adjust for it, we may be able to modify the test.

To formalize this idea, let's take a simple example: suppose the sample comes from an AR(1) process defined in (1.1), our goal is to estimate the mean of the stationary distribution. The estimator would still be the sample mean, but the variance of the estimator is not $\frac{\sigma^2}{n}$ as in i.i.d. case. According to Brooks et al. (2010), the variance of the estimator is:

$$Var(\bar{x}) = \frac{\sigma^2}{n}[\frac{1+\rho}{1-\rho} - \frac{2}{n}\frac{\rho(1-\rho^n)}{(1-\rho)^2}],$$

where $\sigma^2$ is the variance of the stationary distribution, which is 1 here and $\rho$ is the coefficient of the AR(1) process. If an estimator from i.i.d. sample has the same

variance as it, the sample size is:

$$n_{ESS} = \frac{n}{\frac{1+\rho}{1-\rho} - \frac{2}{n}\frac{\rho(1-\rho^n)}{(1-\rho)^2}}.$$

As $n \to \infty$, the effective sample size $n_{ESS} = n\dfrac{1-\rho}{1+\rho}$ asymptotically. So if we replace $n$ by $n_{ESS}$ in the regular variance calculation $\dfrac{\sigma^2}{n}$, we end up with a correct variance calculation.

Bayley and Hammersley (1946) discussed the effective number of independent observations in an autocorrelated time series, specifically about estimating mean and variance of a distribution. Further more, besides estimation problem, it has been demonstrated that some hypothesis test can be also modified for autocorrelated data using ESS adjustment. For example, Yue and Wang (2004) proposed a modified Mann-Kendall test by ESS adjustment to detect trend in a time series. Similar to the K-S test, the original Mann-Kendall test has an assumption of independent sampling, which is not satisfied in most time series data. Yue and Wang proposed use ESS adjustment to calculate the "correct" variance that leads to the correct rejection rate. In these problems, usually the only required modification is to replace the actual sample size $n$ by the effective sample size $n_{eff}$. And the modified estimator/test works reasonably well.

In our problem, It turns out that $f(\rho) = \frac{1}{\sqrt{1-\rho}}$ corresponds to the following ESS formula:

$$n_{ESS} = n(1-\rho).$$

In fact, after adjusting the effective sample size, $\sqrt{n(1-\rho)}D_n$ is compared with

the critical value of the K-S test $KS(\rho = 0)$, which is equivalent to comparing $\sqrt{n}D_n$ with $\frac{KS(\rho=0)}{\sqrt{1-\rho}}$. Theory about why this is the proper formula of ESS adjustment for comparing two distributions hasn't been found yet, but heuristically it makes sense. Recall that the ESS formula for estimating the mean is $n_{ESS} = n\frac{1-\rho}{1+\rho}$. In general, distinguishing the difference between two distributions is easier than distinguishing the means of them. So information about the difference of two distributions is more than information about the difference of two means in the same sample, wherein indeed $\frac{1-\rho}{1+\rho} < 1 - \rho$ if $\rho > 0$.

### 1.4.2 Comparison with Weiss's method

Weiss (1978) suggested a way of modifying the K-S test for autocorrelated data. They considered data generated by an AR(2) process with coefficients $\rho_1, \rho_2$: he first fitted the empirical relationship (a linear function) between $\frac{KS(\rho_1,\rho_2)}{KS(\rho_1=\rho_2=0)}$ and $\tau, n$, where n is the sample size and $\tau$ is derived from the second and fourth moments of the normalized spectrum of the autoregressive process; then he used this empirical relationship to predict the critical values of the K-S statistics for a given AR(2) process with known $\rho_1, \rho_2, n$. In comparison to it, the method we suggest is essentially fitting the empirical relationship between $\frac{KS(\rho_1)}{KS(\rho_1=0)}$ and $\rho_1$, and then using this to predict the critical values. As a result, the type I error rate of Weiss's method is closer to 0.05 since it incorporates both coefficients $\rho_1, \rho_2$ and sample size $n$ (simulations illustrate that the type I error rate of our method is less than 0.05 most of the time). However, our method is easier to implement and has a nice interpretation of ESS adjustment. More importantly, our method can be easily generalized to two sample test, which

will be discussed in Section 1.5.2.

## 1.5 Implementation of the modified K-S test

### 1.5.1 Implementation on AR(1) processes

We have discussed the construction of the modified K-S test and its interpretation. In this section, a Monte Carlo simulation study was conducted to investigate power of the modified K-S test under different alternatives.

We still considered the following AR(1) processes:

$$X_1 \sim \epsilon_n \tag{1.17}$$

$$X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\epsilon_n, n = 2, 3, \cdots \tag{1.18}$$

where $\epsilon_n$ are iid error terms following a distribution which is not necessarily normal. Again $\epsilon_n$ is independent of $X_j, j = 1, 2, \cdots, n-1$. When $\epsilon_n \sim N(0,1)$, the underlying stationary distribution is $N(0,1)$; when $\epsilon_n$ follows some other distribution, such as exponential distribution, the underlying stationary distribution is not $N(0,1)$ anymore (though we don't know it exactly). According to Mallows (1967), we know the following:

Suppose $\{Y_u, u \in U\}$ (U is the set of all integers) is a process of standardized, independent and identically distributed random variables with finite third moment and with a common absolutely continuous distribution function $G$, and $\{a_u, u \in U\}$ is a sequence of real numbers such that $\sum_u a_u^2 = 1$, then $X_u = \sum_w a_w Y_{u-w}$ defines a

stationary linear process with $E(X_u) = 0, E(X_u^2) = 1$ for $u \in U$. Let $f$ be the density function of $X_0$. If $\max_u |a_u|$ is small, then for each $w \in U$, $X_w$ is close to Gaussian in the sense that $\int_{-\infty}^{\infty}(f(y)-\phi(y))^2 dy \leq g \max_u |a_u|$, where $\phi(\cdot)$ is the standard Gaussian density function and $g$ only depends on $G$.

If we carefully rewrite our process in the above formula, we can find that $X_n = \sum_{k=1}^{n} a_{n-k}\epsilon_k$, where $a_k = \rho^k\sqrt{1-\rho^2}$. As $n \to \infty$, this limiting distribution is the one described by Mallows above. $\max_u |a_u| = \sqrt{1-\rho^2}$, which means as $\rho$ increases, the limiting distribution will be closer to $N(0,1)$ in the sense described before. This will be observed and discussed later in this section.

We want to investigate power of the modified K-S test under these different alternatives. Moreover, it is generally easier to distinguish the difference between two distributions that have different means or standard deviations. So in order to make fair comparisons, we chose alternative distributions such that the stationary distributions still have mean 0 and standard deviation 1. For instance, if the alternative is the exponential distribution, we chose $\epsilon_n \sim$ Exp-1. It can be easily verified that the underlying stationary distribution has mean 0 and standard deviation 1. As a summary, the following six distributions were chosen as the error term distributions:

Except for "Normal2"[3], all the other five distributions were modified such that the corresponding stationary distributions have mean 0 and standard deviation 1. The null hypothesis being tested is that the stationary distribution is $N(0,1)$. It's worth noting that in the "Normal" case the rejection rate is actually the type I error

[3]In "Normal2" case, the starting point $X_1$ also has a different distribution: we draw $X_1 \sim N(0.1, 1)$ instead of $N(\frac{0.1(1-\rho)}{\sqrt{1-\rho^2}}, 1)$ so that the stationary distribution is exactly $N(0.1, 1)$.

Table 1.1: Error term distribution

| Name | Error term distribution ($\epsilon_n$) | Underlying stationary distribution |
|---|---|---|
| Normal | $N(0,1)$ | $N(0,1)$ |
| Uniform | $Unif(-\sqrt{3},\sqrt{3})$ | unknown [mean=0, sd=1] |
| $t_8$ | $\frac{\sqrt{3}}{2}t_8$ | unknown [mean=0, sd=1] |
| Exponential | $Exp-1$ | unknown [mean=0, sd=1] |
| Lognormal | $e^{N(0,\sigma^2)}-e^{\frac{\sigma^2}{2}}$, $(\sigma^2=\log\frac{1+\sqrt{5}}{2})$ | unknown [mean=0, sd=1] |
| Normal2 | $N(\frac{0.1(1-\rho)}{\sqrt{1-\rho^2}},1)$ | $N(0.1,1)$ |

rate; while in other cases the rejection rates are power of the test under different alternatives.

In order to make comparison with, another alternative to the modified K-S test was also performed here–we call it "thinning" K-S test. The idea is quite simple: we know that autocorrelation breaks the validation of the K-S test. If we can somehow remove the autocorrelation, problem will be solved. So a simple but inefficient way of doing is "thinning" the data, meaning that we discard some data until the remaining is almost independent of each other. For example, we can choose every other $m$ observations. Then the K-S test was performed on the reduced sample. This method is of course not ideal: despite the reduction of power by discarding data, how to choose an appropriate $m$ is unclear. If we choose a relatively small $m$, the autocorrelation may not be removed, which makes the validation of the test doubtful. On the other hand, if we choose too big an $m$, the sample size of the reduced data is very small, leading to unsatisfying power. Just heuristically as a rule of thumb, we chose $m$ in such a way that it is the smallest integer satisfying $\rho^m < 0.01$.[4] We wanted to make a comparison of power between our method and this "thinning" method.

[4]It means the correlation of two observations with lag $m$ is smaller than 0.01.

For each AR(1) process, the first 200 iterations were discarded as burning. 1000 samples were simulated from each of the six distributions and the rejection rates of both tests were recorded. This was replicated 20 times to estimate the standard error of rejection rates. This study was conducted on sample sizes (length of the process) 100, 300, 1000, 3000, 10000 and $\rho = 0.6, 0.75, 0.9$. Below are selected plots:

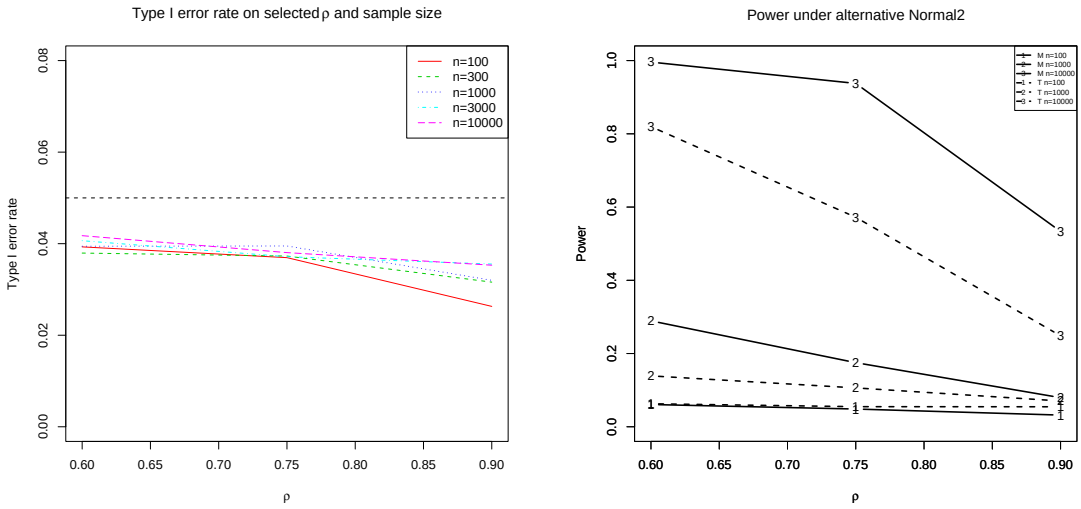


Figure 1.4: Selected summary plots

The left graph is the rejection rate under the distribution "Normal", e.g. type I error rate of the modified K-S test. It's immediately seen that the type I error rate is smaller than its nominal rate 0.05. Besides that, there are two important features of this graph: as $\rho$ increases, the type I error rate decreases; as the sample size $n$ increases, the type I error rate gets closer to 0.05. The right graph is the rejection rate under the distribution "Normal2", e.g. power of those two tests[5] under "Normal2" alternative. Several important features of this graph: power of both tests

[5]In the graph, "M" stands for the modified K-S test and "T" stands for the "thinning" K-S test.

increase as $\rho$ decreases and sample size $n$ increases; power of the modified K-S is significantly larger than that of the "thinning" K-S test. More detailed results are provided in Table 1.2.

There are several things that can be observed from Table 1.2:

- When the underlying stationary distribution is $N(0,1)$ (column Normal), the rejection rates of the modified K-S test are less than its nominal level 0.05 while the rejection rates of the "thinning" K-S test are roughly around 0.05. This tells us that the modified K-S test is slightly conservative.

- When the distribution of the error term is symmetric, such as Uniform, and $t_8$ in the table, power is relatively small for both tests, especially for big $\rho$. For example when $\rho = 0.9$, power of two tests is almost the same as the type I error rates. This has been mentioned before. When $\rho$ is close to 1, the limiting distribution is close to $N(0,1)$[6]. As a result, power is relatively low. But we still can see that for $\rho = 0.6, 0.75$ power of the modified K-S test is greater than that of the "thinning" K-S test.

- When the distribution of the error term is not symmetric about 0, such as Exponential, Lognormal and Normal2 in the table, power is relatively large. Moreover, power of the modified K-S test is almost as twice big as the "thinning" K-S test.

- Power increases as sample size increases, decreases as $\rho$ increases with all other parameters fixed. And asymptotically it's increasing to 1.

[6]Especially if the distribution of $\epsilon_n$ is symmetric.

Table 1.2: Rejection rates of the modified and "thinning" K-S tests

| | | Rejection rate(sd) under the distribution | | | |
|---|---|---|---|---|---|
| | | Normal | | Uniform | |
| | | Modified | Thinning | Modified | Thinning |
| 0.6 | 100 | 0.039 ( 0.007 ) | 0.054 ( 0.006 ) | 0.043 ( 0.007 ) | 0.062 ( 0.009 ) |
| | 1000 | 0.039 ( 0.006 ) | 0.047 ( 0.007 ) | 0.09 ( 0.008 ) | 0.071 ( 0.009 ) |
| | 10000 | 0.042 ( 0.006 ) | 0.051 ( 0.006 ) | 0.829 ( 0.01 ) | 0.282 ( 0.012 ) |
| 0.75 | 100 | 0.037 ( 0.006 ) | 0.057 ( 0.008 ) | 0.033 ( 0.008 ) | 0.055 ( 0.006 ) |
| | 1000 | 0.04 ( 0.007 ) | 0.052 ( 0.008 ) | 0.04 ( 0.007 ) | 0.056 ( 0.007 ) |
| | 10000 | 0.038 ( 0.008 ) | 0.05 ( 0.006 ) | 0.106 ( 0.007 ) | 0.076 ( 0.006 ) |
| 0.9 | 100 | 0.026 ( 0.004 ) | 0.053 ( 0.006 ) | 0.026 ( 0.006 ) | 0.052 ( 0.006 ) |
| | 1000 | 0.032 ( 0.006 ) | 0.054 ( 0.008 ) | 0.031 ( 0.005 ) | 0.057 ( 0.007 ) |
| | 10000 | 0.035 ( 0.006 ) | 0.05 ( 0.008 ) | 0.035 ( 0.007 ) | 0.051 ( 0.006 ) |
| | | Rejection rate(sd) under the distribution | | | |
| | | Exponential | | $t_8$ | |
| | | Modified | Thinning | Modified | Thinning |
| 0.6 | 100 | 0.127 ( 0.01 ) | 0.08 ( 0.009 ) | 0.041 ( 0.008 ) | 0.048 ( 0.009 ) |
| | 1000 | 0.951 ( 0.008 ) | 0.382 ( 0.013 ) | 0.054 ( 0.007 ) | 0.048 ( 0.007 ) |
| | 10000 | 1 ( 0 ) | 1 ( 0 ) | 0.265 ( 0.011 ) | 0.109 ( 0.011 ) |
| 0.75 | 100 | 0.068 ( 0.009 ) | 0.062 ( 0.005 ) | 0.036 ( 0.008 ) | 0.049 ( 0.005 ) |
| | 1000 | 0.37 ( 0.018 ) | 0.164 ( 0.01 ) | 0.042 ( 0.005 ) | 0.05 ( 0.007 ) |
| | 10000 | 1 ( 0 ) | 0.921 ( 0.006 ) | 0.077 ( 0.007 ) | 0.063 ( 0.008 ) |
| 0.9 | 100 | 0.03 ( 0.006 ) | 0.053 ( 0.009 ) | 0.028 ( 0.007 ) | 0.052 ( 0.007 ) |
| | 1000 | 0.073 ( 0.006 ) | 0.065 ( 0.007 ) | 0.035 ( 0.007 ) | 0.051 ( 0.008 ) |
| | 10000 | 0.504 ( 0.018 ) | 0.206 ( 0.008 ) | 0.038 ( 0.006 ) | 0.048 ( 0.007 ) |
| | | Rejection rate(sd) under the distribution | | | |
| | | Lognormal | | Normal2 | |
| | | Modified | Thinning | Modified | Thinning |
| 0.6 | 100 | 0.122 ( 0.012 ) | 0.077 ( 0.01 ) | 0.061 ( 0.007 ) | 0.063 ( 0.007 ) |
| | 1000 | 0.9 ( 0.012 ) | 0.364 ( 0.014 ) | 0.289 ( 0.016 ) | 0.139 ( 0.008 ) |
| | 10000 | 1 ( 0 ) | 1 ( 0 ) | 0.996 ( 0.002 ) | 0.82 ( 0.012 ) |
| 0.75 | 100 | 0.061 ( 0.01 ) | 0.055 ( 0.007 ) | 0.048 ( 0.004 ) | 0.055 ( 0.008 ) |
| | 1000 | 0.292 ( 0.011 ) | 0.133 ( 0.01 ) | 0.175 ( 0.011 ) | 0.106 ( 0.009 ) |
| | 10000 | 1 ( 0 ) | 0.802 ( 0.013 ) | 0.938 ( 0.006 ) | 0.572 ( 0.02 ) |
| 0.9 | 100 | 0.028 ( 0.006 ) | 0.05 ( 0.006 ) | 0.032 ( 0.006 ) | 0.055 ( 0.006 ) |
| | 1000 | 0.051 ( 0.007 ) | 0.054 ( 0.006 ) | 0.08 ( 0.01 ) | 0.07 ( 0.008 ) |
| | 10000 | 0.221 ( 0.016 ) | 0.112 ( 0.013 ) | 0.533 ( 0.016 ) | 0.248 ( 0.012 ) |

### 1.5.2 Two sample test

As is mentioned in Section 1.4.2, one advantage of this modified K-S test over Weiss's method is that it can be easily generalized to two sample test problem. If the null hypothesis is that two samples (with sample sizes $n_1$ and $n_2$ respectively) follow the same distribution, the K-S test can be modified in this way:

$$n_{ESS,1} = n_1(1-\rho_1), \tag{1.19}$$

$$n_{ESS,2} = n_2(1-\rho_2). \tag{1.20}$$

As a result, $\sqrt{\frac{n_{ESS,1}n_{ESS,2}}{n_{ESS,1}+n_{ESS,2}}}D_{n_1,n_2}$ instead of $\sqrt{\frac{n_1n_2}{n_1+n_2}}D_{n_1,n_2}$ is compared to the critical value of the Kolmogorov distribution. Similar to the one sample test, this modification just replaces the sample sizes by their effective sample sizes respectively, where $\rho_1$ and $\rho_2$ are coefficients of the two AR(1) processes.

The following simulation study is similar to the previous section: the null hypothesis is that two samples follow the same stationary distribution. Each time two samples were simulated from AR(1) processes defined by (1.1): in one sample, $\epsilon_n \sim N(0, 1-\rho^2)$ while in the other sample $\epsilon_n \sim$ $N(0, 1-\rho^2)$, $\sqrt{1-\rho^2}$(Exp$-1$) or $N(0.1(1-\rho), 1-\rho^2)$. Both modified K-S and "thinning" K-S tests were performed. Different sample sizes 100, 300, 1000 and 3000 were chosen with combinations of $\rho = 0.6, 0.9$. Below are selected plots:

Similar patterns can be observed as in one sample test: the type I error rate is smaller than its nominal value; power increases as the sample size increases and sample correlation decreases; power of the modified K-S test is significantly larger

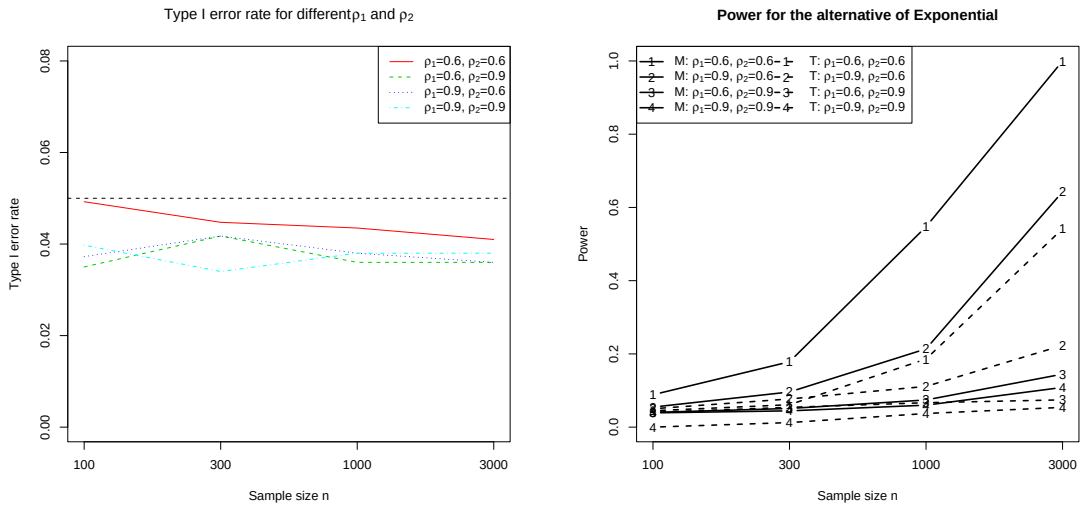


Figure 1.5: Selected summary plots

than the "thinning" K-S test. More detailed results are provided in Table 1.3:

Again, the first part tells us that even when two samples come from AR(1) processes with different coefficients, type I error rate of the modified K-S test is still close to 0.05, being slightly conservative. The second and third parts tell us power of the modified K-S test asymptotically increases to 1, being greater than that of the "thinning" K-S test.

### 1.5.3 Implementation on autocorrelated processes beyond AR(1)

We derive the modified K-S test from AR(1) processes, so it's not surprising that it performs well on AR(1) processes. What remains interesting to us is how general our method is. We wonder how our method behaves on data generated from some autocorrelated processes other than AR(1). Moving average (MA) process is one commonly used time series model. The other example is Metropolis algorithm generated

Table 1.3: Rejection rates for two sample test

| | | Rejection rate under the distribution | | | |
|---|---|---|---|---|---|
| | | Normal | | | |
| | | $\rho = 0.6$ | | $\rho = 0.9$ | |
| | | Modified | Thinning | Modified | Thinning |
| $\rho = 0.6$ | n= 100 | 0.049 | 0.03 | 0.035 | 0.043 |
| | n= 300 | 0.045 | 0.028 | 0.042 | 0.056 |
| | n= 1000 | 0.044 | 0.048 | 0.036 | 0.052 |
| | n= 3000 | 0.041 | 0.041 | 0.036 | 0.043 |
| $\rho = 0.9$ | n= 100 | 0.037 | 0.046 | 0.04 | 0 |
| | n= 300 | 0.042 | 0.05 | 0.034 | 0.008 |
| | n= 1000 | 0.038 | 0.056 | 0.038 | 0.034 |
| | n= 3000 | 0.036 | 0.045 | 0.038 | 0.036 |
| | | Exponential | | | |
| | | $\rho = 0.6$ | | $\rho = 0.9$ | |
| | | Modified | Thinning | Modified | Thinning |
| $\rho = 0.6$ | n= 100 | 0.088 | 0.045 | 0.04 | 0.04 |
| | n= 300 | 0.178 | 0.061 | 0.051 | 0.054 |
| | n= 1000 | 0.547 | 0.186 | 0.074 | 0.067 |
| | n= 3000 | 1 | 0.542 | 0.145 | 0.074 |
| $\rho = 0.9$ | n= 100 | 0.055 | 0.05 | 0.039 | 0 |
| | n= 300 | 0.096 | 0.077 | 0.044 | 0.012 |
| | n= 1000 | 0.214 | 0.111 | 0.06 | 0.037 |
| | n= 3000 | 0.643 | 0.223 | 0.108 | 0.054 |
| | | Normal2 | | | |
| | | $\rho = 0.6$ | | $\rho = 0.9$ | |
| | | Modified | Thinning | Modified | Thinning |
| $\rho = 0.6$ | n= 100 | 0.052 | 0.032 | 0.044 | 0.051 |
| | n= 300 | 0.081 | 0.04 | 0.054 | 0.051 |
| | n= 1000 | 0.156 | 0.079 | 0.076 | 0.07 |
| | n= 3000 | 0.412 | 0.187 | 0.16 | 0.089 |
| $\rho = 0.9$ | n= 100 | 0.041 | 0.046 | 0.045 | 0 |
| | n= 300 | 0.058 | 0.06 | 0.046 | 0.011 |
| | n= 1000 | 0.077 | 0.06 | 0.054 | 0.038 |
| | n= 3000 | 0.16 | 0.089 | 0.11 | 0.059 |

data. By design, this type of data is also highly autocorrelated. By implementing the test on these two models, we explore the generality of our method.

The simulation procedure is similar to the AR(1) case. The only difference is $\rho$–the AR(1) coefficient– is unknown. As a result, for each simulated process, we first fit an AR(1) process and then use the fitted coefficient $\hat{\rho}$ to calculate effective sample size.

**MA(2) process**

In this simulation study, data are generated from the following MA(2) processes:

$$Y_n = \rho_1 X_n + \rho_2 X_{n-1} \tag{1.21}$$

$$X_n \overset{\text{iid}}{\sim} F, \tag{1.22}$$

where $F$ is chosen to be the standard Normal distribution or Cauchy distribution due to their stability. For different combinations of $\rho_1$ and $\rho_2$, a number of MA(2) processes are simulated. Then the modified K-S test is implemented on each process against the null hypothesis ($H_0 : N(0,1)$ if $F \sim N(0,1)$; $H_0 : Cauchy$ if $F \sim Cauchy$). Rejection rates are calculated by repeating this procedure.

Similar to the AR(1) case, we implement our test on two scenarios: under the null hypothesis and under the alternative hypothesis. For $F$ being $N(0,1)$, the combination $\{(\rho_1, \rho_2) : \rho_1^2 + \rho_2^2 = 1\}$ correspond to the null hypothesis; for $F$ being $Cauchy$, the combination $\{(\rho_1, \rho_2) : |\rho_1| + |\rho_2| = 1\}$ correspond to the null hypothesis. Table 1.4 summarizes the actual type I error rate for different $(\rho_1, \rho_2)$ and sample sizes:

A similar pattern to the AR(1) case is observed here. The modified K-S test is slightly conservative. Power is also investigated. For $F$ being $N(0,1)$, we choose the combination $\{(\rho_1, \rho_2) : |\rho_1| + |\rho_2| = 1\}$ as the alternative; for $F$ being $Cauchy$, we

Table 1.4: Type I error rates of the modified and "thinning" K-S tests

| | | Type I error rate(sd) under the distribution | | | |
|---|---|---|---|---|---|
| | | Normal | | Cauchy | |
| | | Modified | Thinning | Modified | Thinning |
| $r_1 = 0.2$ | n= 100 | 0.053 ( 0.006 ) | 0.061 ( 0.006 ) | 0.063 ( 0.007 ) | 0.054 ( 0.006 ) |
| | n= 300 | 0.044 ( 0.008 ) | 0.046 ( 0.006 ) | 0.057 ( 0.004 ) | 0.049 ( 0.006 ) |
| | n= 1000 | 0.043 ( 0.007 ) | 0.048 ( 0.007 ) | 0.055 ( 0.009 ) | 0.045 ( 0.007 ) |
| $r_1 = 0.5$ | n= 100 | 0.027 ( 0.003 ) | 0.05 ( 0.004 ) | 0.023 ( 0.002 ) | 0.05 ( 0.008 ) |
| | n= 300 | 0.024 ( 0.003 ) | 0.052 ( 0.008 ) | 0.019 ( 0.005 ) | 0.046 ( 0.007 ) |
| | n= 1000 | 0.019 ( 0.007 ) | 0.044 ( 0.003 ) | 0.017 ( 0.005 ) | 0.043 ( 0.006 ) |
| $r_1 = 0.8$ | n= 100 | 0.024 ( 0.004 ) | 0.052 ( 0.005 ) | 0.061 ( 0.007 ) | 0.054 ( 0.008 ) |
| | n= 300 | 0.019 ( 0.003 ) | 0.05 ( 0.008 ) | 0.06 ( 0.006 ) | 0.049 ( 0.005 ) |
| | n= 1000 | 0.017 ( 0.003 ) | 0.045 ( 0.01 ) | 0.052 ( 0.009 ) | 0.048 ( 0.008 ) |

choose the combination $\{(\rho_1, \rho_2) : \rho_1^2 + \rho_2^2 = 1\}$ as the alternative. The following table summarizes power for different $(\rho_1, \rho_2)$ and sample sizes:

Table 1.5: Power of the modified and "thinning" K-S tests

| | | Power(sd) under the distribution | | | |
|---|---|---|---|---|---|
| | | Normal | | Cauchy | |
| | | Modified | Thinning | Modified | Thinning |
| $r_1 = 0.2$ | n= 100 | 0.094 ( 0.007 ) | 0.086 ( 0.009 ) | 0.113 ( 0.01 ) | 0.09 ( 0.008 ) |
| | n= 300 | 0.251 ( 0.012 ) | 0.133 ( 0.012 ) | 0.162 ( 0.012 ) | 0.106 ( 0.012 ) |
| | n= 1000 | 0.936 ( 0.007 ) | 0.54 ( 0.012 ) | 0.42 ( 0.011 ) | 0.262 ( 0.014 ) |
| $r_1 = 0.5$ | n= 100 | 0.097 ( 0.007 ) | 0.056 ( 0.01 ) | 0.106 ( 0.014 ) | 0.097 ( 0.008 ) |
| | n= 300 | 0.54 ( 0.018 ) | 0.162 ( 0.012 ) | 0.262 ( 0.014 ) | 0.152 ( 0.012 ) |
| | n= 1000 | 1 ( 0 ) | 0.733 ( 0.015 ) | 0.847 ( 0.004 ) | 0.374 ( 0.013 ) |
| $r_1 = 0.8$ | n= 100 | 0.099 ( 0.008 ) | 0.087 ( 0.008 ) | 0.097 ( 0.01 ) | 0.099 ( 0.008 ) |
| | n= 300 | 0.26 ( 0.01 ) | 0.136 ( 0.008 ) | 0.252 ( 0.009 ) | 0.147 ( 0.015 ) |
| | n= 1000 | 0.94 ( 0.007 ) | 0.535 ( 0.016 ) | 0.86 ( 0.011 ) | 0.37 ( 0.017 ) |

Again, our method strictly dominates the "thinning" method, implying a reasonable power.

**Metropolis algorithm generated data**

In this simulation study, data were sampled from a Metropolis algorithm. The null hypothesis stays the same, testing whether the stationary distribution is $N(0,1)$. Three markov chains were designed based on Metropolis algorithm such that their stationary distributions are $N(0,1)$, $Unif(-\sqrt{3},\sqrt{3})$ and $Exp(1)-1$ respectively.[7] Similar to the simulation in 5.1, the first 200 iterations were discarded in each sample. 1000 samples were simulated and the rejection rates were recorded. This was repeated 20 times to estimate the standard error of the rejection rate. Different sample sizes (length of the process) 300, 1000, 3000, 1000 were chosen. Below is a summary table:

Table 1.6: Rejection rates for Metropolis generated data

| | Rejection rate under the distribution | | | | | |
|---|---|---|---|---|---|---|
| | Normal | | Exponential | | Uniform | |
| | Modified | Thinning | Modified | Thinning | Modified | Thinning |
| n= 300 | 0.151 | 0.086 | 0.343 | 0.21 | 0.258 | 0.178 |
| n= 1000 | 0.068 | 0.063 | 0.318 | 0.194 | 0.143 | 0.106 |
| n= 3000 | 0.041 | 0.055 | 0.463 | 0.226 | 0.145 | 0.102 |
| n= 10000 | 0.031 | 0.049 | 0.968 | 0.393 | 0.318 | 0.146 |

There are several things that can be observed from this table:

- When the underlying stationary distribution is $N(0,1)$ (column Normal), rejection rates of both tests are larger than its significance level 0.05 when sample size is not large enough. However, both rejection rates decrease as sample size increases. Similar to the previous results, the modified K-S test is slightly conservative while the "thinning" K-S test roughly has its nominal significance

[7]We intent to design in this way such that the stationary distributions all have mean 0 and standard deviation 1.

level.

- When the underlying stationary distribution is not $N(0,1)$, power asymptotically increases to 1 as sample size increases and power of the modified K-S test is larger than that of the "thinning" K-S test with the same sample size. The uniform distribution (symmetric) is harder to distinguish than the exponential distribution (asymmetric).

As a brief summary, the Metropolis based markov chain constructed here is not exactly but can be reasonably fitted as an AR(1) process. The modified K-S test on this is conservative when sample size is large enough. Its power asymptotically increases to 1 as sample size increases, faster than the power of the "thinning" K-S test.

## 1.6 Application to fMRI data analysis

As is mentioned earlier, the use of the K-S test in fMRI analysis was criticized in Aguirre et al. (2005) due to the presence of autocorrelation. Now we show that the modified K-S test provides a remedy to this issue.

We followed the procedure in Aguirre et al. (2005) except that the data set used in Aguirre et al. (2005) no longer exits, so we used a similar data set instead. A brief description of the problem is as follows: functional MRI (fMRI) is a technique to determine how the human brain works, specifically speaking, precisely which part of the brain is handling critical functions such as thought, speech, movement, and sensation. The general principle is that if some stimulus is given to a subject, the

area of the brain in charge of that stimulus will be activated. As a result, the fMRI signal values of that area will increase, which can be detected through experiment. Researchers can find plausible areas that are handling that stimulus through such an experiment. A typical experiment goes as follows: researchers collect fMRI signal values on a subject over a period. During that period, a stimulus is given to the subject periodically (say every 20 seconds) and the stimulus can last some time (say 10 seconds). Despite issues as the delay of response, noise, etc., we can compare the signal values during the control period to the experiment period and find the area whose signal values significantly differ. One way is to use the K-S test. However, as Geoffrey K. Aguirre et al. pointed out, the temporal correlation (autocorrelation) among the time series data leads to more false positive results under the null hypothesis.

To repeat the experiment by Geoffrey K. Aguirre et al., we downloaded data set from NITRC. The data set consists of BOLD fMRI signal values collected on 20 subjects under resting state (no stimulus was given at all). As suggested by Geoffrey K. Aguirre et al., the temporal structure of the assumed paradigm was designated to be a boxcar with an 80-s period (i.e., two 40-s epochs; 0.0125 Hz). In other words, the signal values within the time slots 0-40s, 80-120s, etc. are treated as control period while the signal values within the time slots 40-80s, 120-160s, etc. are treated as experiment period. For each voxel, observations from the two periods are compared via the K-S test. For each subject, the overall rejection rate on all voxels represents the actually Type I error rate (false positive rate), since observations are sampled under the null hypothesis (control period and experiment period have the same distribution of signal values). Linear signal drift was removed, but motion correction was not

conducted due to lack of corresponding data. The K-S tests were conducted on the 20 subjects and we had a similar result as in Aguirre et al. (2005): The significance level of the K-S test was chosen as 0.05. However, the actual false positive rates were much higher than 0.05. The median false positive rate among the 20 subjects is 0.12. If we use the modified K-S test instead, the median false positive rate falls to 0.05. Below is a graph demonstrating the false positive rates of two methods for each subject:

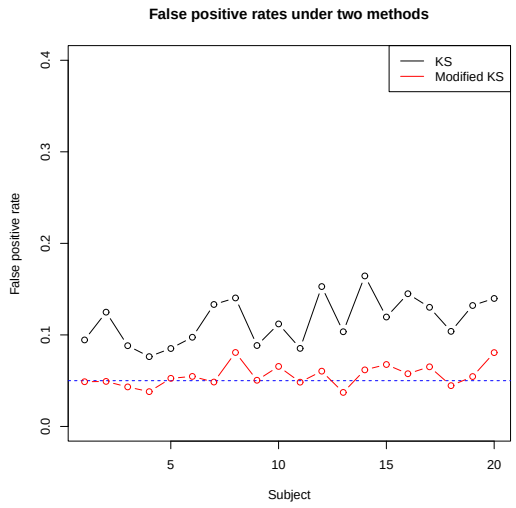


Figure 1.6: Actual Type I error rate

The false positive rate has been pulled down significantly, very close to the nominal level of 0.05. The main reason of the high false positive rate of the K-S test is due to the presence of autocorrelation. On the other hand, the fMRI signal value processing can be thought approximately an AR process. As a result, our modified K-S test based on AR(1) fitting does a good job no surprisingly. Similar results are derived if we model the hypothetical experiment in different frequencies (corresponding to

higher or lower autocorrelation).

One main critique of the use of the K-S test in this field is that it can't control false positive rates as well as other parametric methods when autocorrelation exists. Our method provides a remedy of this issue. It demonstrates that the K-S test can still be used, only with a slight modification of ESS adjustment. On the other hand, it also demonstrates that our modified K-S test is applicable to real problems.

## 1.7 Concluding remarks

The main contribution of this article is to provide a modified K-S test when the *independent* assumption is violated. This method, though derived from a simple autocorrelated process–AR(1) process with Normal stationary distribution, has been demonstrated to be reasonably used for other autocorrelated data as well. The idea of effective sample size adjustment is not new, but has been shown powerful in dealing with autocorrelated/time series data. However, along with direction of this article, several questions remain interesting to us:

1. In our article, we justify the mixing condition (1.16) and hence can estimate the critical value of the K-S statistic from those Gaussian Processes. However in general, for the underlying AR(1) process

$$X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\epsilon_n,$$

where $\epsilon_n$ is not Normal, does the K-S statistic converge to some quantity? If so, what does it converge to? To answer this, two questions need to deal with:

- Does the sequence $\{\sqrt{n}(\sum_{i=1}^{n} I_{\Phi(X_i)\leq t} - t) : 0 \leq t \leq 1\}$ converge? If we still want to apply Billingsley's theorem, we need to verify the mixing condition. But it's not as easy as in the Normal case to evaluate those mixing coefficients. Moreover, Billingsley's theorem gives a sufficient but not necessary condition for weak convergence.
- Even if we assume that the sequence $\{\sqrt{n}(\sum_{i=1}^{n} I_{\Phi(X_i)\leq t} - t) : 0 \leq t \leq 1\}$ converges, what does it converge to? How to evaluate the formula (1.15) is unclear for a general AR(1) process, since the stationary distribution is unknown to us. But if we can find a way to estimate it, we can again simulate those Gaussian processes to estimate the critical values for the K-S statistic.

There questions still remain open to us. It would be interesting and useful if we can find a general answer to the asymptotic distribution of the K-S statistic for an AR(1) process (or even more general process).

2. We derived the ESS formula by fitting a curve to the simulated critical values. However, is there a theoretical justification for the simple ESS formula $n_{ESS} = n(1-\rho)$? Can we prove that it is (approximately) the correct formula?

3. We have seen that the ESS adjustment method is very powerful in dealing with problems for autocorrelated data. But under what circumstances is this method applicable? Is there a systematic way of finding the ESS formula? If we can answer these type of questions, a large amount of time series problems can be solved.

# Chapter 2

# Improving MCMC convergence rate with Residual Augmentations

## 2.1 Residual Augmentations: A Unified Strategy

### 2.1.1 Creative Re-Parameterization and Over-Parameterization

Designing algorithms that are *simple, stable, and speedy* is a dream shared by virtually anyone working on Markov chain Monte Carlo (MCMC) or more generally on statistical computing. For data augmentation (Tanner and Wong (1987)) and Gibbs sampling (Geman and Geman (1984); Gelfand and Smith (1990)) type of algorithms, it is well known that parameterizations can affect substantially both convergence and ease of implementation (e.g., Gelfand et al. (1995, 1996); Van Dyk and Meng (2010)). By using parameterizations creatively, a variety of strategies have been proposed to accelerate convergence while maintaining implementation simplicity. In particular,

Papaspiliopoulos et al. (2003, 2007) study the *centered*, *noncentered*, and *partially noncentered* parameterizations. The idea of partial noncentering is to introduce a family of parameterizations (or data augmentation schemes), and then to seek the optimal parameterization for fastest convergence. This is mathematically equivalent to the conditional augmentation approach (Meng and Van Dyk (1999); Van Dyk and Meng (2001)), where the family of data augmentation schemes are indexed by a *working parameter*.

Formally, consider the model $p(\theta|Y_{obs}) \propto p(Y_{obs}|\theta)p(\theta)$ where $\theta$ is the parameter of interest and $Y_{obs}$ denotes observed data. A data augmentation (DA) model $p(Y_{obs}, Y_{mis}|\theta)$ is any joint distribution of $Y_{mis}$ (the missing or augmented data) and $Y_{obs}$ given $\theta$ such that the marginal $p(Y_{obs}|\theta)$ is preserved. In other words, we can write $p(Y_{obs}, Y_{mis}|\theta) = p(Y_{obs}|\theta)p(Y_{mis}|Y_{obs}, \theta)$. In conditional augmentation, a working parameter $c$ is introduced such that

$$p(Y_{obs}, Y_{mis}|\theta, c) = p(Y_{obs}|\theta)p(Y_{mis}|Y_{obs}, \theta, c).$$

Whereas $p(Y_{obs}, Y_{mis}|\theta, c)$ clearly is a legitimate DA because it preserves the desired margin $p(Y_{obs}|\theta)$, it is a form of over-parameterization because the working parameter $c$ is not identifiable by the observed data $Y_{obs}$. For conditional augmentation, the value of $c$ is obtained by optimizing a certain criterion, e.g., based on the convergence rate of the closely related EM algorithm (Meng and Van Dyk (1998, 1999, 2002) and Van Dyk and Meng (2001)). The resulting algorithm alternates between drawing $\theta$ given $(Y_{obs}, Y_{mis})$ and drawing $Y_{mis}$ given $(\theta, Y_{obs})$, conditioning on the chosen value of $c$. Finding a good conditional augmentation scheme requires a careful balance

between the theoretical speed and ease of implementation, as illustrated in detail by Van Dyk and Meng (2001, 2010).

This *conditional augmentation* approach contrasts with the *marginal augmentation* approach (Meng and Van Dyk (1999)), which is closely related to parameter-expanded DA (PX-DA; Liu and Wu (1999)). In marginal augmentation, the working parameter $c$ is marginalized out after being assigned a *working prior* $p(c)$. The resulting algorithm is a standard DA—labeled Scheme 2 in Van Dyk and Meng (2001) – alternating between drawing $Y_{mis}$ given $(\theta, c, Y_{obs})$ and drawing $(\theta, c)$ given $(Y_{mis}, Y_{obs})$ based on the joint posterior

$$p(Y_{mis}, \theta, c|Y_{obs}) \propto p(Y_{obs}, Y_{mis}|\theta, c)p(\theta)p(c). \tag{2.1}$$

We can also sample from (2.1) by alternating between drawing $(Y_{mis}, c)$ given $(\theta, Y_{obs})$ and drawing $(\theta, c)$ given $(Y_{mis}, Y_{obs})$, as in PX-DA (Liu and Wu (1999)). Obviously this is algorithmically equivalent to the DA sampler that alternates between drawing $Y_{mis}$ given $(\theta, Y_{obs})$ and drawing $\theta$ given $(Y_{mis}, Y_{obs})$, which was labeled Scheme 1 in Van Dyk and Meng (2001).

The strategies discussed above all amount to using a single data augmentation scheme in the actual implementation. For conditional augmentation, this is rather obvious by construction. For marginal augmentation, if the working prior $p(c)$ is proper, then Scheme 1 is the standard DA using

$$\tilde{p}(Y_{mis}|Y_{obs}, \theta) = \int p(Y_{mis}|Y_{obs}, \theta, c)p(c)\mu(dc) \tag{2.2}$$

as the data augmentation, where $\mu$ is the dominating measure for the working prior, typically the Lebesgue measure. However, when $p(c)$ is improper, Scheme 1 is not feasible. In contrast, Scheme 2 still is implementable, just as an improper prior can still lead to a proper posterior. But this does not automatically imply that the algorithm will converge properly. Minimally it should be clear that the resulting joint chain for $(\theta, c, Y_{mis})$ cannot be positive recurrent because its target distribution (2.1) is improper when $p(c)$ is improper. By a result of Hobert (2001b,a), this also automatically implies that the corresponding (major) sub-chain for $(\theta, c)$ cannot be positive recurrent either. However, when the improper working prior is the limit of a sequence of proper priors, then under regularity conditions, the *sub-sub-chain* produced by Scheme 2 for $\theta$ will converge to the desired target distribution $p(\theta|Y_{obs})$. Intriguingly, when $p(c)$ corresponds to the right Haar measure, this sub-sub-chain actually represents the fastest algorithm among a class of DA algorithms as formulated in Liu and Wu (1999) with their elegant group-theoretic argument.

Even more intriguingly, there is often a simpler way to reach this optimality by using two standard data augmentation schemes (i.e., no improper prior is involved), and the new strategy is demonstrably more powerful and versatile than all known strategies based on a single (limiting) data augmentation, for reasons presented in the following section.

### 2.1.2 Alternating versus Interweaving

Suppose $p(Y_{mis}, \theta|Y_{obs})$ and $p(\tilde{Y}_{mis}, \theta|Y_{obs})$ are two augmentation schemes (i.e., both preserving the target posterior $p(\theta|Y_{obs})$). An obvious strategy is to concate-

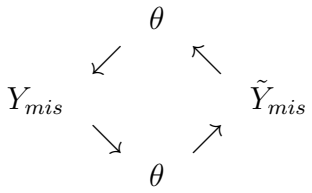


Figure 2.1: Alternating Scheme

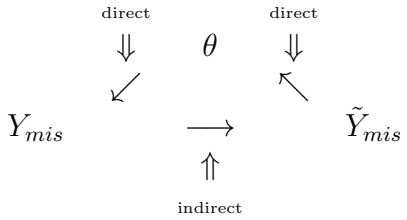


Figure 2.2: Interweaving Scheme

nate two iterations, one based on each of the two schemes, that is, by *alternating* between the two algorithms. This may be represented schematically as in Figure 2.1 where each arrow indicates a sampling step. For example $Y_{mis} \to \theta$ means drawing $\theta$ given the current $Y_{mis}$ (and $Y_{obs}$). Somewhat surprisingly, Yu and Meng (2011) demonstrate that an alternative *interweaving strategy* holds much more promise than the simple alternating scheme. Specifically, the interweaving strategy simply cuts out the $\theta$ between $Y_{mis}$ and $\tilde{Y}_{mis}$, and hence it leads to the triangular diagram given in Figure 2.2. That is, each iteration cycles through the parameter $\theta$ and the two sets of augmented data by first drawing $Y_{mis}$ given $\theta$, then $\tilde{Y}_{mis}$ given $Y_{mis}$, and then $\theta$ given $\tilde{Y}_{mis}$. (Henceforth we suppress the conditioning on $Y_{obs}$ when there is no confusion.)

The triangular diagram also reveals a fundamental insight about the power of the interweaving strategy. Similar to the usual DA algorithm, whose convergence rate is the square of the maximal correlation between $Y_{mis}$ and $\theta$ in their joint posterior, the interweaving strategy has a convergence rate that is bounded above by the product of three maximal correlations as indicated by the three links in the above diagram. That is, let the geometric convergence rate of DA under $Y_{mis}$ and $\tilde{Y}_{mis}$ be $r_1$ and $r_2$, respectively, and the rate for the interweaving scheme be $r_{1\&2}$. Then Yu and Meng (2011) proved that

$$r_{1\&2} \leq \mathcal{R}(Y_{mis}, \tilde{Y}_{mis})\mathcal{R}(\theta, Y_{mis})\mathcal{R}(\theta, \tilde{Y}_{mis}) = \mathcal{R}_{1,2}\sqrt{r_1 r_2} \tag{2.3}$$

where $\mathcal{R}_{1,2} \equiv \mathcal{R}(Y_{mis}, \tilde{Y}_{mis})$, and

$$\mathcal{R}(X_1, X_2) = \sup_{g,h \in L^2} \text{Corr}\{g(X_1), h(X_2)\}$$

is the maximal correlation between (generic) $X_1$ and $X_2$. (Note in our application, the joint distribution is the joint posterior predictive distribution $p(Y_{mis}, \tilde{Y}_{mis}|Y_{obs})$.)

As discussed in Yu and Meng (2011), the key insight here is that we can make $r_{1\&2}$ small (which means a faster algorithm) by making *any one of* $\{\mathcal{R}_{1,2}, r_1, r_2\}$ small. Indeed, it is even possible that $r_1 = r_2 = 1$, that is, neither of the two DAs being interwoven is geometrically convergent, and yet $r_{1\&2} = 0$, that is, the interwoven algorithm will deliver i.i.d draws! See Yu and Meng (2011) for such an example.

In general, achieving i.i.d. draws is obviously too much of a dream, but the interweaving strategy provides us with a new way to combat the common problem of high dependence among consecutive MCMC draws. Specifically, with either alternating or interweaving, we can reduce the dependence between $\theta^{(t)}$ and $\theta^{(t+1)}$—where $t$ indexes the iteration—by reducing either $r_1$ or $r_2$ or both. Schematically, this corresponds to "breaking" either or both of the two *direct* links marked in Figure 2.2; here a direct link is an arrow with $\theta$ as one of its two end points. However, the interweaving strategy allows us to achieve the same goal by breaking an *indirect* link, which does not involve $\theta$, and clearly it exists only in Figure 2.2.

Therefore, given the original augmentation as represented by the arrow from $\theta$ to

$Y_{mis}$, we now have two ways to break the cycle. The first is to make $\tilde{Y}_{mis}$ independent of $\theta$ and hence to break the $\tilde{Y}_{mis} \to \theta$ link, which is what partially non-centering or conditional augmentation aims to achieve. The second is to make $\tilde{Y}_{mis}$ independent of $Y_{mis}$, thereby breaking the $Y_{mis} \to \tilde{Y}_{mis}$ link, which is what marginal augmentation and the interweaving strategy try to accomplish. In particular, Yu and Meng (2011) advocate an ancillarity-sufficiency interweaving strategy (ASIS) that takes advantage of the existing competing nature between sufficient augmentation and ancillary augmentation to reduce their *a posteriori* dependence.

### 2.1.3 Direct and Indirect Residual Augmentations

Such considerations lead to the idea of residual augmentations (Yu and Meng (2011), Rejoinder), as a way to "break links" by judiciously choosing $\tilde{Y}_{mis}$ for a given (original) DA scheme $(Y_{mis}, \theta)$. For the *direct residual augmentation* (DRA), we attempt to break the direct link $\tilde{Y}_{mis} \to \theta$ by choosing $\tilde{Y}_{mis}$ to be a residual from regressing $Y_{mis}$ on $\theta$. The central idea here is that a residual is constructed to be uncorrelated (though rarely independent) with the regression function, which is $\theta$ here. The obvious choice is the usual additive residual from regressing $Y_{mis}$ on $\theta$ :

$$\tilde{Y}_{mis} = Y_{mis} - \mathrm{E}[Y_{mis}|\theta, Y_{obs}]. \tag{2.4}$$

A less obvious one is its multiplicative variant:

$$\tilde{Y}_{mis} = \frac{Y_{mis}}{\mathrm{E}[Y_{mis}|\theta, Y_{obs}]} \tag{2.5}$$

in the scalar case. It is straightforward to show that both $\tilde{Y}_{mis}$'s are uncorrelated with $\theta$ with respect to the joint posterior distribution $p(Y_{mis}, \theta|Y_{obs})$, as long as the correlation exists. (But note the condition of having correlation does not hold for (2.5) as often as it does for (2.4)).

For the *indirect residual augmentation* (IRA), the aim is to break the indirect link $Y_{mis} \to \tilde{Y}_{mis}$, and hence we need to regress $\theta$ on $Y_{mis}$. This naturally leads to the counterparts of (2.4) and (2.5) by swapping $\theta$ and $Y_{mis}$, that is,

$$\tilde{Y}_{mis} = \theta - \mathrm{E}[\theta|Y_{mis}, Y_{obs}] \tag{2.6}$$

and

$$\tilde{Y}_{mis} = \frac{\theta}{\mathrm{E}[\theta|Y_{mis}, Y_{obs}]}. \tag{2.7}$$

For all these constructions, the implementation $Y_{mis} \to \tilde{Y}_{mis}$ is typically straightforward. We accomplish this by first drawing $\theta$ from $p(\theta|Y_{mis}, Y_{obs})$, which is a step required by the original DA algorithm based on $Y_{mis}$ alone. We can then compute $\tilde{Y}_{mis}$ as a deterministic function of $\theta$, $Y_{mis}$ and $Y_{obs}$. This computation typically is straightforward for DRA, because $\mathrm{E}[Y_{mis}|\theta, Y_{obs}]$ is simply the mean function of the full conditional $p(Y_{mis}|\theta, Y_{obs})$ already needed by the original DA algorithm; it can also be carried out by Monte Carlo if necessary. For IRA, this task typically is even simpler, because it calls only for $\mathrm{E}[\theta|Y_{mis}, Y_{obs}]$, the complete-data posterior mean.

Therefore, the simplicity of a residual augmentation algorithm depends critically on how easy it is to implement the $\tilde{Y}_{mis} \to \theta$ step. To implement it exactly requires us to derive the conditional distribution of $\theta$ given $\tilde{Y}_{mis}$ as implied by one of

(2.4)-(2.7). This may not be an easy task when the regression function involved (i.e., $\mathrm{E}[Y_{mis}|\theta, Y_{obs}]$ or $\mathrm{E}[\theta|Y_{mis}, Y_{obs}]$) is non-linear. This issue, however, can be dealt with pragmatically by adopting a convenient global or local approximation, with the trade-off of achieving less reduction in auto-correlations for implementation simplicity. Such a pragmatic approach also helps us to compromise appropriately between implementation simplicity and the desire to find suitable transformations of $g(\theta)$ and $h(Y_{mis})$ such that the low correlation between them is a reasonable indicator of their lack of dependence. Note ideally we would want a joint one-to-one transformation $T(\theta, Y_{mis})$ for better joint normality because under joint normality low linear correlation is the same as low maximal correlation. Unfortunately, this joint transformation typically will destroy the simplicity of the original Gibbs setup that alternates between $\theta$ and $Y_{mis}$.

For the rest of the chapter, in Section 2.2 we first illustrate some theoretical properties of residual augmentations using the simplest normal hierarchical model and its extensions, which include $t$ distributions. In particular, we note an interesting "safe zone" for the choice of augmentation schemes and show how ASIS can be viewed as a "minimax" strategy, always staying within the safe zone regardless of the prior specification and the configuration of observed data. Our pragmatic strategy is illustrated in Section 2.3 with a probit regression example. We conclude in Section 2.4 with a host of open problems.

## 2.2 Theoretical Illustrations and a Phase Transition Phenomenon

### 2.2.1 Illustrating DRA and IRA

A common illustrative example in the DA literature is the one-way random effect model (Liu and Wu (1999); Yu and Meng (2011); Hobert and Roman (2011)). Instead of repeating the standard setup, here we adopt a simpler representation capturing its essence that is relevant for our algorithmic investigation. Specifically, suppose $\theta$ is the parameter of interest and $Y_{mis}$ is the missing datum or latent variable, and their joint posterior distribution (given $Y_{obs}$) can be standardized into

$$\left( \begin{array}{c} \theta \\ Y_{mis} \end{array} \right) \Bigg| Y_{obs} \sim N \left[ \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} 1 & r \\ r & 1 \end{array} \right) \right]. \tag{2.8}$$

Here $r$ is a known function of $Y_{obs}$ and, without loss of generality, we can assume $0 \leq r < 1$. The standard DA based on $Y_{mis}$ then iterates between sampling $\theta$ given $Y_{mis}$ and sampling $Y_{mis}$ given $\theta$ (all conditioning on $Y_{obs}$ of course). Clearly this DA has the convergence rate $r_1 = r^2$.

Now consider a conditional augmentation or partially non-centering scheme $\tilde{Y}_{mis} = Y_{mis} - c\theta$, with $c$ being a working parameter to be determined. Clearly

$$\left( \begin{array}{c} \theta \\ \tilde{Y}_{mis} \end{array} \right) \Bigg| Y_{obs} \sim N \left[ \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \left( \begin{array}{cc} 1 & r-c \\ r-c & 1+c^2-2rc \end{array} \right) \right]. \tag{2.9}$$

This implies that the DA algorithm using $\tilde{Y}_{mis}$ as the augmentation will have con-

vergence rate $r_2 = (r-c)^2/(1+c^2-2rc)$. Now because of their joint normality, the maximal correlation between $Y_{mis}$ and $\tilde{Y}_{mis}$ is the same as the absolute value of their linear correlation. Therefore $\mathcal{R}_{1,2} = |\text{Corr}(\tilde{Y}_{mis}, Y_{mis})| = |1-cr|/\sqrt{1+c^2-2rc}$. Because the bound in (2.3) is sharp for this normal setting (Yu and Meng (2011)), we see that the rate of convergence from interweaving the DA based on $Y_{mis}$ and the DA based on $\tilde{Y}_{mis}$ is

$$r_{1\&2} = \mathcal{R}_{1,2}\sqrt{r_1 r_2} = \frac{|r(r-c)(1-cr)|}{1+c^2-2rc}. \tag{2.10}$$

We see immediately that when $c = r$ or $c = r^{-1}$, $r_{1\&2} = 0$, and hence the interweaving strategy will produce i.i.d. draws. The $c = r$ case corresponds to DRA because $\text{E}[Y_{mis}|\theta, Y_{obs}] = r\theta$, and hence taking $c = r$ in $\tilde{Y}_{mis} = Y_{mis} - c\theta$ is the same as making $\tilde{Y}_{mis}$ the additive residual, which is independent of $\theta$ because of normality. Consequently, the link $\tilde{Y}_{mis} \to \theta$ is completely broken, yielding i.i.d. draws. On the other hand, because $\text{E}[\theta|Y_{mis}, Y_{obs}] = rY_{mis}$, taking $c = r^{-1}$ in $\tilde{Y}_{mis} = Y_{mis} - c\theta = -c(\theta - c^{-1}Y_{mis})$ is equivalent to setting $\tilde{Y}_{mis} = \theta - rY_{mis}$, which is the IRA. The joint normality ensures that $\tilde{Y}_{mis}$ is independent of $Y_{mis}$, and hence IRA completely breaks the indirect link $Y_{mis} \to \tilde{Y}_{mis}$, again resulting in i.i.d. draws.

### 2.2.2 A Phase Transition Phenomenon

In real applications, rarely can the direct or indirect link be broken completely. Even under the normality assumption, we may not be able to compute the regression slopes with infinite precision. A natural question then arises: What happens if we use a $c$ that approximates a regression slope (i.e., from regressing $Y_{mis}$ on $\theta$ or $\theta$ on $Y_{mis}$)?

Does it still retain approximately the benefit of residual augmentation? Common wisdom would suggest so, based on the usual continuity argument.

Unfortunately, the continuity argument would fail here. A clue is offered by considering what happens when $c = 1$, which corresponds to using ASIS for this model (see Yu and Meng (2011)), and when $r$ approaches 1. On the one hand, when $c = 1$, it is easy to see from (2.10) that

$$r_{1\&2} = \frac{r(1-r)}{2} \leq \frac{1}{8} \tag{2.11}$$

for all $0 \leq r < 1$. On the other hand, for any $c \neq 1$, if we let $r \to 1$, $r_{1\&2}$ will approach 1. Clearly therefore there is a discontinuity at $c = r = 1$. More interestingly or even magically, as proved in the Appendix, the 1/8 bound in (2.11) holds whenever $c$ falls between the two regression slopes, that is, whenever $r \leq c \leq r^{-1}$, with the bound 1/8 achieved if and only if $r = 1/2$ and $c = 1$.

However, as seen in the perspective plot Figure 2.3 and the contour plot Figure 2.4, as soon as $c$ leaves this "safe" zone $[r, r^{-1}]$, the convergence rate $r_{1\&2}$—as a function of $(c, r)$ denoted by $g(c, r)$—increases dramatically, exhibiting essentially a phase transition type of phenomenon at the two boundaries $c = r$ and $c = r^{-1}$. As hinted previously, this phenomenon is most extreme at the point $(c, r) = (1, 1)$: If we fix $c = 1$, then $g(c, r) = r(1-r)/2 \to 0$ as $r \to 1$; if we fix $r = 1$, then $g(c, r) = 1$ for any $c$ (including $c = 1$ by a limiting argument).

A geometric interpretation of this phenomenon can help us to understand it better. The joint (degenerate) normality of $(\theta, Y_{mis}, \tilde{Y}_{mis})$ allows us to visualize the three pairwise (maximal) correlations in a *single* triangle, as in Figure 2.5, where each vec-

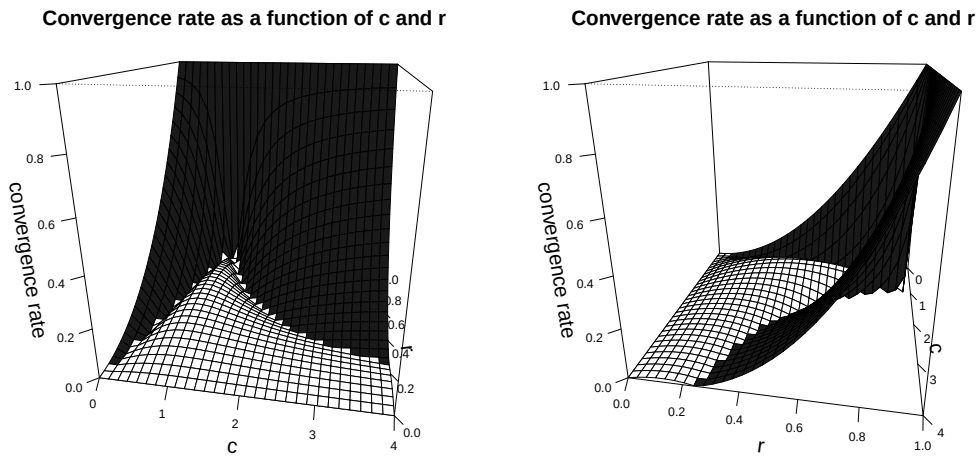


Figure 2.3: Convergence rate as a function of $(c, r)$ viewed in two perspectives: the light area is the "safe" zone, where the convergence rate is bounded by $\frac{1}{8}$; the dark area is outside the two regression lines, where the convergence rate increases dramatically.

tor represents a random variable, and the cosine of the (directional) angle between two vectors is their correlation. Denote the pairwise correlations between $(\theta, Y_{mis})$, $(\theta, \tilde{Y}_{mis})$ and $(Y_{mis}, \tilde{Y}_{mis})$ as $\cos\alpha_1 (> 0)$, $\cos\alpha_2$ and $\cos\alpha_3$ respectively. From geometry, we know that $\alpha_2 = \alpha_1 + \alpha_3$. The convergence rate of the interweaving strategy is

$$r_{1\&2} = |\cos\alpha_1 \cos\alpha_2 \cos\alpha_3| = |\cos\alpha_1 \cos(\pi - (\alpha_1 + \alpha_3)) \cos\alpha_3|. \tag{2.12}$$

For a nonobtuse triangle, the product of cosines of its three angles cannot exceed $8^{-1}$, hence the same bound is achieved when $\tilde{Y}_{mis}$ falls in the shaded area. Moreover, within the "safe" zone,

$$\mathrm{Corr}(\tilde{Y}_{mis}, Y_{mis})\mathrm{Corr}(\tilde{Y}_{mis}, \theta) \leq 0.$$

This says that the pairwise correlations of $(\theta, \tilde{Y}_{mis})$ and $(Y_{mis}, \tilde{Y}_{mis})$ should have oppo-

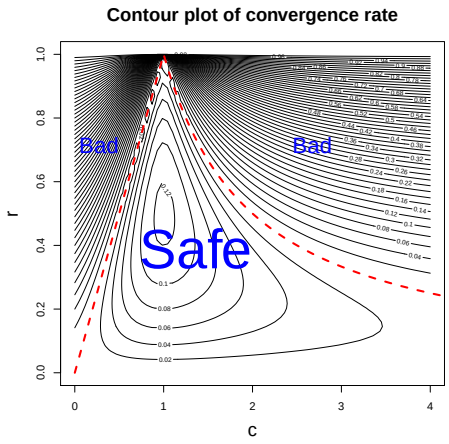


Figure 2.4: Contour plot: the dashed lines correspond to the two regression slopes, $c = r$ and $c = r^{-1}$.

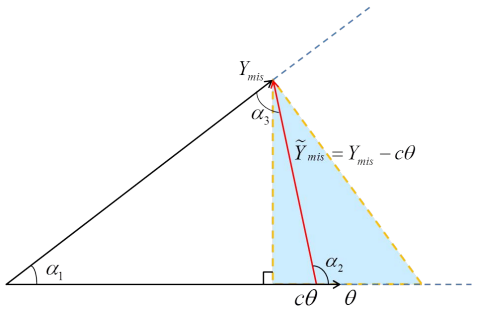


Figure 2.5: Geometric interpretation: the shaded area corresponds to the "safe" zone, where the formed triangle is nonobtuse.

site signs to make the interweaving algorithm stable. This finding is consistent with empirical observations and heuristic arguments reported in Yu and Meng (2011) that the interweaving strategy works by taking advantage of the "beauty and beast" nature of two competing DAs. It may also help us search for similar "safe" interweaving algorithms for more complicated problems.

### 2.2.3 Going Beyond Normality

But before one conjectures generalizations inspired by this simple example, one must contemplate the possibility that, without the normality condition, such a "safe zone" may completely disappear. After all, the aforementioned 1/8 bound for $r_{1\&2}$ depends critically on the triangulation formulation in (2.12), which was possible because maximal correlation is the same as linear correlation (when it is non-negative) under joint normality. One therefore should at least show such a "safe zone" exists

beyond the normality setting.

A common generalization moving beyond normality is to consider a $t$-type of distribution. Here we consider a general class of the so-called "normal/independent" distributions, which includes the $t$ distribution as a special case (see Lange and Sinsheimer (1993)). This class of (univariate or multivariate) distributions model a random variable $Y$ as $Y = Z/W$ (modulo an affine transformation), where $Z$ is (multivariate) normal, and $W$ is univariate and is independent of $Z$ (and hence the "normal/independent" nomenclature). Obviously, choosing $W = \sqrt{\chi^2_v/v}$ gives the $t$ distribution with $v$ degrees of freedom.

With this setup, let us replace the normal model (2.8) by the following conditional normal model. That is, conditioning on a common variable $W$, the posterior distribution of $(\theta, Y_{mis})$ is:

$$\left.\begin{pmatrix} \theta \\ Y_{mis} \end{pmatrix}\right| Y_{obs}, W \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{W^2}\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}\right], \tag{2.13}$$

where $0 \leq r < 1$ is known and is free of $W$ but may depend on $Y_{obs}$. The working parameter remains the same, that is, $\tilde{Y}_{mis} = Y_{mis} - c\theta$, and hence (2.9) remains as well other than adding the conditioning on $W$ and the corresponding multiplicative factor $W^{-2}$ for its covariance matrix. Furthermore, the regression slopes remain the same because

$$E[Y_{mis}|\theta, Y_{obs}] = \mathrm{E}\left\{\mathrm{E}[Y_{mis}|\theta, W, Y_{obs}]\,|\theta, Y_{obs}\right\} = \mathrm{E}\left\{r\theta\,|\theta, Y_{obs}\right\} = r\theta$$

and similarly for $\mathrm{E}[\theta|Y_{mis}, Y_{obs}] = rY_{mis}$.

Without restricting the (posterior) distribution of $W$, we consider in general the maximal correlation between $\theta$ and $Y_{mis}$, which governs the rate of convergence for the DA algorithm for $(\theta, Y_{mis})$. Intuitively, this maximal correlation is determined by two separate sources of dependence, namely the maximal correlation brought in by the common variable $W$ and correlation $r$ between the normal components after conditioning on $W$. Mathematically, this intuition is roughly captured by Lemma 1 of Yu and Meng (2011), which in the current case allows us to establish that

$$\mathcal{R}(\theta, Y_{mis}) \leq r + (1-r)\mathcal{R}(\theta, W)\mathcal{R}(Y_{mis}, W). \tag{2.14}$$

Using this inequality together with (2.3) and the fact that $\mathcal{R}(\theta, W) = \mathcal{R}(Y_{mis}, W)$, we can show (see Appendix) that under (2.13), the rate of convergence for interweaving $Y_{mis}$ and $\tilde{Y}_{mis} = Y_{mis} - c\theta$ satisfies

$$r_{1\&2} \leq [g + (1-g)r][g + (1-g)r_1][g + (1-g)r_2], \tag{2.15}$$

where $g = \mathcal{R}^2(\theta, W)$, and

$$r_1 = \frac{|c - r|}{\sqrt{1 + c^2 - 2cr}} \quad \text{and} \quad r_2 = \frac{|1 - cr|}{\sqrt{1 + c^2 - 2cr}}. \tag{2.16}$$

This leads to the "safe" zone $c \in [r, r^{-1}]$ because within this zone, as shown in the Appendix,

$$r_{1\&2} \leq \frac{1}{8}[1 + g]^3, \tag{2.17}$$

which is strictly less than 1 as long as $g = \mathcal{R}^2(\theta, W) < 1$. Note the bound in

(2.17) again is independent of the value of $r$, and is predetermined by the maximal correlation between a normal/independent variable $Z/W$ and its denominator $W$.

To see how useful the bound in (2.17) can be, let us consider the (bivariate ) $t$ distribution, where $W^2 \sim \chi^2_v/v$. Then $\mathcal{R}(\theta, W)$ is simply the maximal correlation between a $t$ random variable and its denominator, which depends only on the degrees of freedom $v$. We therefore denote it as $\mathcal{R}_v(\theta, W)$ to emphasize this dependence. The analytic calculation of $\mathcal{R}_v(\theta, W)$ seems intractable, but nevertheless we can show that (see Appendix): as $v \to 0$, $\mathcal{R}_v(\theta, W) \to 1$; and as $v \to \infty$, $\mathcal{R}_v(\theta, W) \to 0$. (Incidentally and somewhat ironically, the proof of the latter assertion turns out to be surprisingly difficult, but we were able to establish it by employing a set of well-known theoretical tools for bounding MCMC convergence rate itself.) Therefore (2.17) is a generalization of the $8^{-1}$ bound under normality because, as $v \to \infty$, the $t$ distribution converges to normal, and $\frac{1}{8}[1 + \mathcal{R}^2_v(\theta, W)]^3 \to \frac{1}{8}$.

For an arbitrary degrees of freedom $v$, we generated 100,000 $t$ samples and then used the ACE algorithm of Breiman and Friedman (1985)—as given in the $R$-package *acepack*—to estimate the maximal correlation $\mathcal{R}_v(\theta, W)$. For a given $v$, this process was repeated 50 times to construct (95%) confidence intervals, represented by the "vertical dots" in the left panel of Figure 2.6, which plots the resulting estimated curve of $\mathcal{R}_v(\theta, W)$ as a function of $v$ (on a equal-spaced grid 0 to 10 plus $v = 20$). The right panel plots the corresponding bound in (2.17), using the $g = \mathcal{R}^2_v(\theta, W)$ values displayed in the left panel.

We see from the right panel that as soon as $v \geq 6$, the rate appears to not exceed 1/5. Even for $v = 1$, that is, the Cauchy distribution, the *upper confidence limit* on

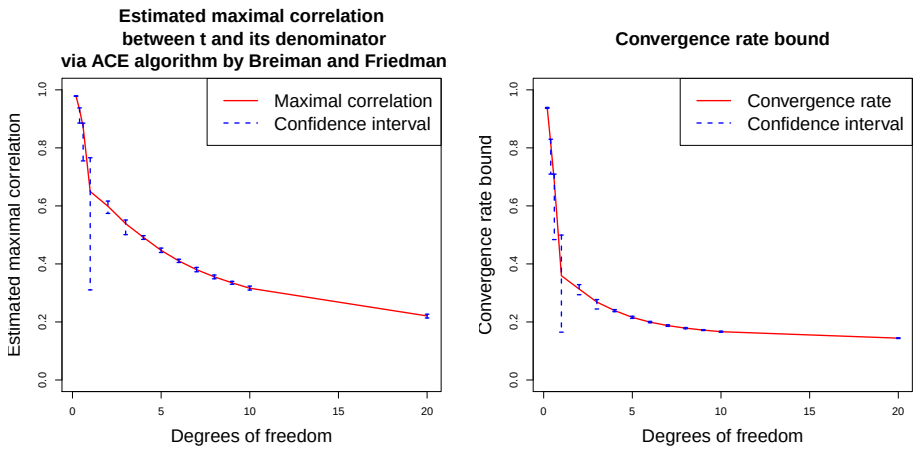


Figure 2.6: Estimated $\mathcal{R}_v(\theta, W)$ and convergence rate bound: the left plot is the estimated $\mathcal{R}_v(\theta, W)$ via ACE; the right plot is the corresponding value given by (2.17), an upper bound for convergence rate.

the upper bound of the convergence rate is only about 1/2. Whereas these bounds are not as good as 1/8 for the normal case, they are far better than adequate for practical purposes, considering most numerical bounds used in practice are above 0.9 and even above 0.99; see for example Hobert (2001a) and Van Dyk and Meng (2001), Rejoinder. Indeed, if we use 0.9 as standard, then unless one fits a $t$ model with tiny fractional degrees of freedom (e.g., $v \leq 0.1$), the interweaving algorithm will be safe as long as $c \in [r, r^{-1}]$.

Regarding the phase transition phenomenon, for the normal model (2.8) we were able to demonstrate it exactly because the chain was reversible and the inequality (2.3) becomes equality under that normal model. For this more general normal/independence model, we currently can only demonstrate such a phenomenon for the bound in (2.15). This is given in Figure 2.7, where the four values of $g$ correspond to four values of the degrees of freedom $v$ in the left panel of Figure 2.6. We see clearly the very similar shape as in Figure 2.3, other than that the function values in the safe

zone increase as $g$ increases. This of course only provides suggestive evidence (and it is only for $t$ distributions), and we certainly hope a more direct demonstration can be found, perhaps via finding a lower bound that shares a similar shape as in Figure 2.7.

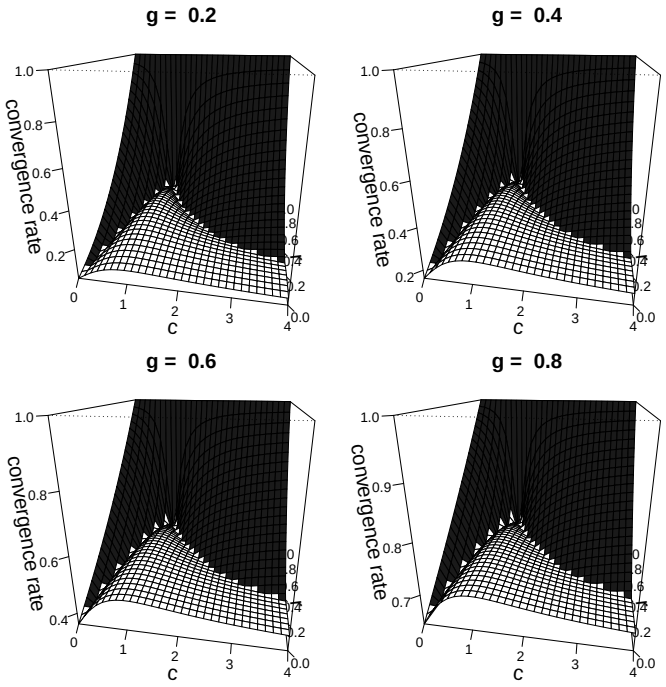


Figure 2.7: Bound in (2.15) as a function of $(c, r)$, with four values of $g = \mathcal{R}^2(Z/W, W)$. Note the different ranges of values on the vertical axis.

Regardless of the extent to which the phase transition phenomenon exists for an arbitrary choice of $W$, it is clear that the choice of $c = 1$ is always safe irrespective of the actual value of $r$ (which depends on the actual data) in the sense that (2.17) always hold for $c = 1$. Indeed, here $c = 1$ can be viewed as a minimax choice because it minimizes the maximal convergence rate (strictly speaking, an upper bound of the rate) against all possible values of $r$.

A couple of remarks are in order before we illustrate the power of applying the interweaving strategy with residual augmentations in a practical setting. First, the

demonstrations above are to provide theoretical insights (e.g., the phase transition phenomenon) and to illustrate theoretical potential (e.g., the upper bound (2.17)); they do not take into account the issue of ease or cost of implementation, an issue that will be investigated in the next section. Second, as a side note, the very wide confidence intervals seen in Figure 2.6 only at $v = 1$ may seem puzzling at the first sight, because one might expect the Monte Carlo error getting progressively worse as $v$ decreases below 1, which corresponds to tail behavior that is even heavier than Cauchy (so much so that a sample mean is more dispersed than a single observation). Whereas we do not have a good explanation for this phenomenon, we suspect it is related to a hidden symmetry in the maximal correlation, that is, $\mathcal{R}(Z/W, W) = \mathcal{R}(W/Z, W)$, with the Cauchy distribution corresponding to the center of symmetry because it is invariant to the reciprocal transformation, and hence its unique properties.

## 2.3 An Empirical Exploration via Probit Regression

### 2.3.1 A Locally Linearized Direct Residual Augmentation

Consider the widely used Probit regression model:

$$Y_{obs,i} = \text{sign}(Y_{mis,i}), \qquad Y_{mis,i}|\theta, X_i \sim N(X_i\theta, 1), \tag{2.18}$$

where $Y_{obs,i}$ is the observed binary ($\pm 1$) outcome, the sign of the latent score $Y_{mis,i}$, $X_i$ is a $1 \times p$ vector of covariates, and $\theta$ is a $p \times 1$ vector of regression coefficients. Write

$Y_{obs} = (Y_{obs,1}, \cdots, Y_{obs,n})^\top$, $Y_{mis} = (Y_{mis,1}, \cdots, Y_{mis,n})^\top$, and $X^\top = (X_1^\top, \cdots, X_n^\top)$. Taking the standard noninformative prior $p(\theta) \propto 1$, we have the well-known full conditional distributions for the standard DA/Gibbs sampler (see Albert (1992); Albert and Chib (1993); Meng and Schilling (1996)):

$$Y_{mis,i}|Y_{obs}, \theta \sim \text{TN}(X_i\theta, 1, Y_{obs,i}); \tag{2.19}$$

$$\theta|Y_{obs}, Y_{mis} \sim N(\hat{\theta}, (X^\top X)^{-1}). \tag{2.20}$$

Here $\hat{\theta} = (X^\top X)^{-1} X^\top Y_{mis}$, and $\text{TN}(\mu, \sigma^2, Y_{obs,i})$ denotes a $N(\mu, \sigma^2)$ distribution truncated to the interval $(0, \infty)$ if $Y_{obs,i} = 1$ and to $(-\infty, 0)$ if $Y_{obs,i} = -1$. The standard DA/Gibbs sampler iterates between (2.19) and (2.20). Though convenient, it can be extremely slow. Several methods, therefore, have been proposed to improve it, including PX-DA (Liu and Wu (1999)) and ASIS (Yu and Meng (2011)). Below we first demonstrate how to implement the residual augmentation, and then we compare it to several existing algorithms.

Given the normal model in (2.18) for $Y_{mis}$, which is univariate (in contrast to $\theta$, which is often multivariate), it is easier to consider the additive DRA $\tilde{Y}_{mis} = Y_{mis} - \text{E}[Y_{mis}|\theta, Y_{obs}]$. In particular, it is known that if we let $H(z) = z + M(z)$ where $M(z) = \frac{\phi(z)}{\Phi(z)}$ is the inverse Mills ratio, then (recall $Y_{obs,i} = \pm 1$)

$$\text{E}[Y_{mis,i}|\theta, Y_{obs}] = \text{E}[Y_{mis,i}|\theta, Y_{obs,i}] = Y_{obs,i} H(Y_{obs,i} X_i \theta), \qquad i = 1, \ldots, n. \tag{2.21}$$

However, since $H(\pm X_i\theta)$ is non-linear in $\theta$, deriving and then sampling from the resulting $p(\theta|Y_{obs}, \tilde{Y}_{mis})$ is rather a difficult task. As a compromise, we seek a locally

linear approximation to $H(z)$ by utilizing its derivative

$$G(z) = H'(z) = 1 - zM(z) - M^2(z).$$

The resulting local residual augmentation has the form

$$\tilde{Y}_{mis,i} = Y_{mis,i} - G(Y_{obs,i}X_i\theta)X_i\theta, \qquad i = 1, \ldots, n. \tag{2.22}$$

However, the corresponding $p(\theta|Y_{obs}, \tilde{Y}_{mis})$ is still hard to employ, because of the $\theta$ inside the non-linear $G(\cdot)$ function.

### 2.3.2 Seeking Compromise via Adaptive Data-Dependent Augmentation

To further simplify the implementation, we adopt the adaptive MCMC idea (see Rosenthal (2011) and references therein). That is, at $(t+1)$st iteration, we adopt a DA scheme that depends on the value of $\theta^{(t)}$:

$$\tilde{Y}_{mis,i} = Y_{mis,i} - b_i^{(t)}X_i\theta, \quad \text{where} \quad b_i^{(t)} = G(Y_{obs,i}X_i\theta^{(t)}). \tag{2.23}$$

It is critical to recognize that (2.22) and (2.23) are different augmentation schemes, even though they share the same conditional distribution $p(\tilde{Y}_{mis}^{(t+1)}|\theta = \theta^{(t)}, Y_{obs})$. Their difference lies in the two different conditional distributions $p(\theta^{(t+1)}|\tilde{Y}_{mis} = \tilde{Y}_{mis}^{(t+1)}, Y_{obs})$. For scheme (2.22), the $\theta$ inside the $G(\cdot)$ function is free or "live", therefore we need to take it into account when deriving $p(\theta|\tilde{Y}_{mis}; Y_{obs})$ for drawing $\theta^{(t+1)}$.

In contrast, for scheme (2.23), the $\theta^{(t)}$ inside the $G(\cdot)$ is fixed or "dead" by iteration $(t+1)$st, so in deriving $p(\theta|\tilde{Y}_{mis}; Y_{obs})$, $b_i^{(t)} = G(Y_{obs,i}X_i\theta^{(t)})$ is just a constant, rending $\tilde{Y}_{mis}$ of (2.23) truly linear in $\theta$.

This "adaptive linearity" on one hand permits an easy implementation, but on the other hand destroys the proper convergence of the resulting Markov chain. This is because the adaptive DA, namely an iteration-dependent DA model $p^{(t)}(\theta, Y_{mis}|Y_{obs})$, can easily destroy the detailed balance condition. Whereas the detailed balance condition is not necessary for MCMC to converge, without it proper convergence can be easily destroyed. As a matter of fact, our empirical checking indicated that our adaptive algorithm does not converge to our desired target, as demonstrated in Figure 2.10 of Section 2.3.5 below.

Fortunately this is a relatively easy problem to resolve, because the reason we invoke the adaptation is to seek a suitable compromise between simplicity and speed. We therefore can run the adaptive algorithm for a burn-in period, say until $t = t^*$, and then fix $b_i^{(t)} = b_i$ for all $t > t^*$ (and all $i$), eliminating adaptation. Here the value $b_i$ can be chosen in many ways by analyzing $\{b_i^{(t)}, t \leq t^*\}$, such as the average of the last (say) 10% of the $\{b_i^{(t)}, t \leq t^*\}$. Another way to motivate this switching strategy is to consider the adaption as a greedy strategy, i.e., it aims to find the best piecewise linear approximation given the $\theta$ drawn at the current iteration. But what we really need is a good approximation given $\theta$ within a reasonable range as determined by its posterior distribution. Therefore, at the end of the adaptive stage, we form a compromise by taking an appropriate summary of $b_i$'s from the adaptive stage. Currently we do not have a general theoretical framework for choosing the optimal

summary. Nor do we believe there is a unique optimal choice here, because such a choice typically entails a trade-off between statistical efficiency and computational efficiency. Nevertheless, we conducted a preliminary empirical investigation of the impact of the choices of $b_i$, as reported is in Section 2.3.5 below.

In contrast to a global residual augmentation such as $\tilde{Y}_{mis} = Y_{mis} - cX\theta$, where $c$ is a scalar working parameter, the adaptation outlined above allows us to search for a more powerful residual augmentation (for our goal to reduce auto-correlations) by taking into account heterogeneity in different components as governed by the actual observed data. Specifically, the adaption leads to a component-wise (direct) residual augmentation in the form of

$$\tilde{Y}_{mis} \equiv \begin{pmatrix} \tilde{Y}_{mis,1} \\ \tilde{Y}_{mis,2} \\ \vdots \\ \tilde{Y}_{mis,n} \end{pmatrix} = \begin{pmatrix} Y_{mis,1} - b_1 X_1 \theta \\ Y_{mis,2} - b_2 X_2 \theta \\ \vdots \\ Y_{mis,n} - b_n X_n \theta \end{pmatrix} = Y_{mis} - BX\theta, \tag{2.24}$$

where $B = \text{diag}\{b_1, \ldots, b_n\}$. What makes (2.24) more powerful than $\tilde{Y}_{mis} = Y_{mis} - cX\theta$ is not only that it permits heterogeneity among the $n$ components, but more importantly the value of individual working parameter $b_i$ *takes into account the information from the actual observed data* because it depends on the value of $Y_{obs,i}$ as seen in (2.23). The idea of data dependent augmentation has been proposed in the previous literature. Specifically Papaspiliopoulos et al. (2003) provided a data dependent partially non-centered algorithm for the normal hierarchical model. They also suggested a general recipe based on a quadratic approximation to the log-likelihood.

In comparison with their approach, we proposed a general methodology of residual augmentations (direct or indirect) which by nature is data dependent. However, exact residual augmentations are rarely easy to implement, therefore we use a (linear) approximation (as in the probit regression). Based on what we have so far, this approximation needs to be constructed case by case, seeking a compromise between implementation simplicity and statistical efficiency; see Section 2.4 for detailed discussions.

### 2.3.3 A Prototype Algorithm

With the setup outlined above, we can carry out (at least) two algorithms. The first is simply a direct DA algorithm using (2.24) as its augmentation scheme, albeit we need to deal with its adaptive nature, as outlined below. The second is to interweave the first with the standard DA based on the original DA $Y_{mis}$ to gain additional benefit. Below we provide the details for the first, as the interweaving one is rather trivial once the first one is in place.

Specifically, the direct (initially) adaptive DA algorithm requires two-stage execution:

I. *Adaptive Stage*: $t = 1, \ldots, t^*$, update $b_i = b_i^{(t)} (i = 1, \ldots, n)$ at each iteration;

II. *Sampling Stage*: Same as Adaptive Stage, except $b_i$ is fixed as $\bar{b}_i$, the average of the last 10% of the $b_i^{(t)}$'s obtained from the Adaptive Stage ($i = 1, \ldots, n$). (See Section 2.3.5 for other choices.)

Operationally, during the Adaptive Stage, we carry out the following (where $\tilde{Y}_{mis} = (\tilde{Y}_{mis,1}, \ldots, \tilde{Y}_{mis,n})^\top$):

- Draw $\tilde{Y}_{mis}^{(t+1)}|\theta^{(t)}, Y_{obs}$:

Step 1 Update $b_i \Leftarrow b_i^{(t)} = G(Y_{obs,i}X_i\theta^{(t)}), \quad i = 1, \ldots, n$;

Step 2 Draw $Y_{mis,i}^{(t+1)} \sim \text{TN}(X_i\theta^{(t)}, 1, Y_{obs,i})$ and then compute $\tilde{Y}_{mis,i}^{(t+1)} = Y_{mis,i}^{(t+1)} - b_i X_i \theta^{(t)}$.

- Draw $\theta^{(t+1)}|\tilde{Y}_{mis}^{(t+1)}, Y_{obs}$:

Step 3 For $i = 1, \ldots, n$, compute $\tilde{X}_i = (1 - b_i)X_i$ and then

$$\hat{\mu} = (\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top \tilde{Y}_{mis}^{(t+1)}, \quad \hat{\Sigma} = (\tilde{X}^\top \tilde{X})^{-1},$$

where $\tilde{X}^\top = (\tilde{X}_1^\top, \ldots, \tilde{X}_n^\top)$;

Step 4 Draw $\theta^{(t+1)}|\tilde{Y}_{mis}^{(t+1)}, Y_{obs} \sim \text{TN}(\hat{\mu}, \hat{\Sigma})$ with the truncation determined by the constraint that $\text{sign}(\tilde{Y}_{mis,i}^{(t+1)} + b_i X_i \theta^{(t+1)}) = Y_{obs,i}$. We implement this step via a nested Gibbs sampler: for each $i$, draw $\theta_i^{(t+1)}|\tilde{Y}_{mis}^{(t+1)}, Y_{obs}, \theta_{-i}^{(t+1)}$, a truncated univariate normal distribution, and repeat it $K$ cycles.

At the Sampling Stage, we simply skip Step 1, that is, we use $b_i = \bar{b}_i$ for all iterations to produce our MCMC samples. We emphasize that this algorithm is by no means optimal in any sense; there should be many ways to improve upon it especially regarding the potentially time consuming nested Gibbs sampler used in Step 4; see Section 2.3.5 for an exploration. Our aim here is to provide the first prototype algorithm, in a real application setting, that builds upon the concept of residual augmentation formulated in Yu and Meng (2011). Nevertheless, our preliminary numerical experiments, as reported below, have shown great potential even for this

prototype algorithm.

### 2.3.4 A Numerical Comparison

To see the effectiveness of our prototype algorithm and of its interweaving with the standard Gibbs sampler, we conducted a numerical experiment using the lupus nephritis data set of Van Dyk and Meng (2001) Table 1, which has $n = 55$ patients and $p = 3$ covariates (including a constant term for the intercept). We compare it with various other algorithms. The algorithms we included in our comparisons are:

**I.** Standard Gibbs sampler given by (2.19)-(2.20). This is also known as the DA algorithm with *Sufficient Augmentation* (SA) $Y_{mis}$ (Yu and Meng (2011)), and hence it is the same as setting $b_i \equiv 0$ in our prototype algorithm for all $i$ (therefore $\tilde{Y}_{mis} = Y_{mis}$, which makes Step 4 the same as (2.20)).

**II.** A marginal augmentation/PX-DA algorithm based on a *multiplicative* working parameter $\tilde{Y}_{mis} = \sigma Y_{mis}$, with Haar working prior $p(\sigma^2) \propto \sigma^{-2}$—see Liu and Wu (1999) and van Dyk and Meng (1999).

**III.** The DA algorithm based on *Ancillary Augmentation* (AA) $\tilde{Y}_{mis} = Y_{mis} - X\theta$ (Yu and Meng (2011)), which is the same as setting $b_i \equiv 1$ in our prototype algorithm for all $i$.

**IV.** The ASIS algorithm (Yu and Meng (2011)) that interweaves SA and AA in (I) and (III) respectively.

**V.** Our prototype DRA algorithm as given in Section 2.3.3.

**VI.** The algorithm that interweaves (I) and (V) (IS-DRA).

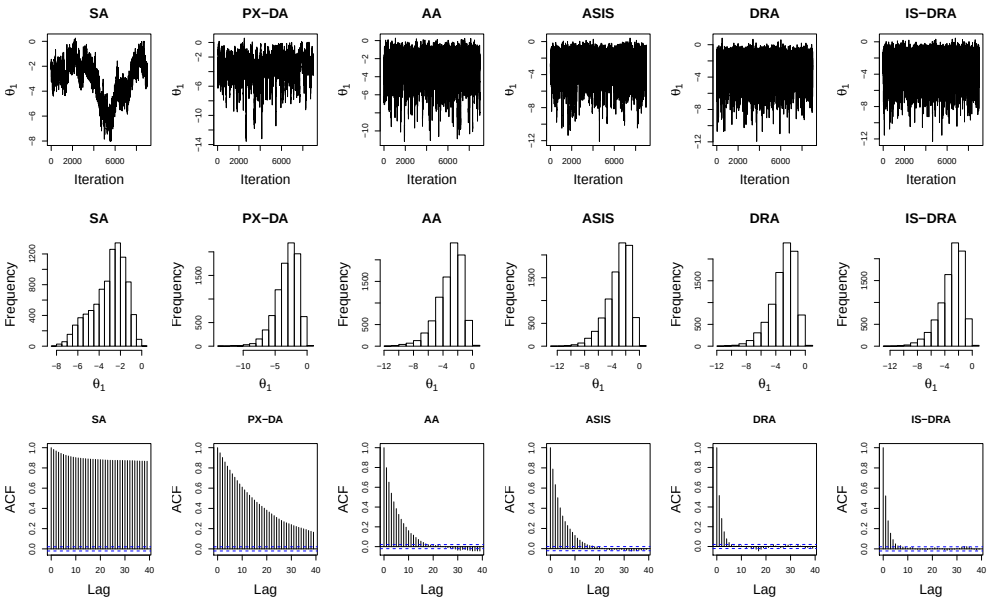


Figure 2.8: Comparing various samplers for the lupus nephritis data: trajectories, histograms and autocorrelations ($K = 30$ for III-VI).

Here the first two are well-known algorithms in the literature, which we employ as benchmarks, even though algorithm II uses a multiplicative working parameter (and hence theoretically it is not directly comparable with those built upon an additive working parameter). The next two are more recent algorithms proposed in Yu and Meng (2011) but without the benefit of tuning $b_i$ according to the actual data because they set $b_i = 1$ for all $i$. The last two are our new prototype algorithms (without and with interweaving), benefiting from allowing data to have a strong influence on choosing $b_i$ $(i = 1, \ldots, n)$. Figure 2.8 displays the trajectories, histograms, and autocorrelations of the draws of $\theta_1$ (the first coefficient) for these six algorithms. We see the algorithms work progressively better, at least in terms of autocorrelations, empirically demonstrating that we are on the right track in our effort to reduce

autocorrelations by using progressively more efficient DA schemes and with the help of the interweaving strategy.

Table 2.1: Comparison of methods after $10,000$ samples averaged over 25 runs. For AA, ASIS, DRA and IS-DRA, the number within the parentheses (e.g. DRA($K$)) denotes the number of iterations in the nested Gibbs sampler. For *relative speed*, we use SA as the reference point, e.g., in this application, IS-DRA(30) is about 110 times faster than SA in terms of the relative speed.

| Method | Mean Time | ESS (min,median,max) | ESS/Time | Relative Speed |
|---|---|---|---|---|
| SA | 20.2 | (4, 16, 81) | 0.8 | 1 |
| PX-DA | 20.6 | (180, 235, 273) | 11.4 | 14 |
| AA(1) | 19.2 | (64, 115, 157) | 6 | 7 |
| AA(10) | 22.5 | (211, 454, 601) | 20.2 | 25 |
| AA(30) | 27.9 | (871, 1025, 1235) | 36.8 | 46 |
| ASIS(1) | 24.9 | (78, 122, 203) | 4.9 | 6 |
| ASIS(10) | 27.3 | (363, 475, 592) | 17.4 | 22 |
| ASIS(30) | 32.8 | (771, 1047, 1337) | 32 | 40 |
| DRA(1) | 20.2 | (199, 259, 366) | 12.8 | 16 |
| DRA(10) | 22.9 | (976, 1233, 1458) | 53.8 | 67 |
| DRA(30) | 28.8 | (2216, 2928, 3968) | 101.6 | 127 |
| IS-DRA(1) | 24.7 | (222, 285, 356) | 11.6 | 14 |
| IS-DRA(10) | 27.5 | (1033, 1294, 1503) | 47.1 | 59 |
| IS-DRA(30) | 33.3 | (2363, 2950, 3286) | 88.6 | 110 |

Clearly autocorrelation measures only (one aspect of the) statistical efficiency. Another important measure is CPU time, an aspect of computational efficiency. Table 2.1 reports the CPU time (in seconds) from 25 replications, together with estimated effective sample size (ESS) obtained from the *R*-package *coda*. We see that in terms of both ESS and Relative Speed, DRA(30)—with or without IS— ranks at the top, about one order of magnitude improvement over PX-DA and about two orders of improvement over SA. However, unlike the statistical efficiency measure ESS, which does not depend on how the algorithm is actually implemented (as long as it is implemented correctly), CPU time depends critically on how the algorithm is implemented,

in what program language(s) it is written, on what machine it is carried out, etc. As a matter of fact, when we initially implemented Step 4 directly via the $R$ language, the drawing from the truncated multivariate normal turned out to be so costly that the gain in statistical efficiency by DRA was outweighed by the lost of computational efficiency. The results in Table 2.1 are from using the $R$-package *tmvtnorm* (by Stefan Wilhelm), which was implemented in Fortran. All the rest of the implementation was done in $R$, except for the drawing from a truncated multivariate uniform as needed by AA and hence also by ASIS (which corresponds to Step 4 of our prototype algorithm; see Section 4.1 of Yu and Meng (2011)). These truncated uniform drawings were also carried out by the same Fortran program *tmvtnorm* with the covariance matrix set to a very large value, so the truncated multivariate normal effectively becomes truncated multivariate uniform. We adopted this strategy to ensure a meaningful comparison of CPU times so the simulation results do not bias toward our DRA; indeed, when we implemented AA and ASIS completely in $R$, their CPU time was much worse than our DRA using *tmvtnorm*, further illustrating how computational efficiency depends critically on the actual implementation, not just the algorithm itself.

We remark here that the substantial increases in ESS as $K$ increases from $K = 1$ to $K = 30$ clearly demonstrate the importance and effectiveness of using data augmentation schemes that are as close to residual augmentations as possible. We also note that in this case the additional gain/protection from using interweaving is rather minor, a consequence of a rather effective DRA for this problem and particular data set. Yu and Meng (2011) provided ample evidence that the performance of any single DA tends to depend on the actual data set much more substantially than those

by interweaving a pair. Our theoretical bounds given in Section 2.3 (albeit they do not apply to the Probit regression problem) provide further suggestive evidence of the robust nature of our interweaving strategy.

### 2.3.5 Seeking Effective Data-Dependent Working Parameter

In Section 2.3.2 we emphasized the importance and potential of allowing the actual data to govern the choice of the working parameter. In the current cases, the working parameters are $\{b_i, i = 1, \ldots, n\}$. In Section 2.3.3 we then mentioned that there are a number of possible choices of $b_i$ for the sampling stage based on working parameter values obtained during the adaptive stage: $\{b_i^{(t)}, t = 1, \ldots, t^*\}$. As a preliminary assessment of the impact of the choice of data-dependent working parameters on ESS, Figure 2.9 displays the box-plots of ESS for six choices of $b_i$'s. They are:

1. Last: Set $b_i = b_i^{(t*)}$, the last updated value of $b_i$ from the adaptive stage;
2. Mean: Set $b_i = \bar{b}_i$, the average of the last 10% $b_i^{(t)}$'s from the adaptive stage;
3. Median: Set $b_i = \text{med}\{b_i\}$, the median of the last 10% $b_i^{(t)}$'s from the adaptive stage;
4. Mode: Set $b_i = \text{mode}\{b_i\}$, the mode of an estimated density (using a kernel method) of the last 10% $b_i^{(t)}$'s from the adaptive stage;
5. Mode2: Set $b_i = G(Y_{obs,i} X_i \hat{\theta})$ (see (2.23)), where $\hat{\theta}$ is the mode of an estimated density (using a kernel method) of the last 10% of $\theta^{(t)}$'s from the adaptive stage;

6. MLE: Set $b_i = G(Y_{obs,i} X_i \hat{\theta}_{MLE})$, where $\hat{\theta}_{MLE}$ is the MLE of $\theta$ under the Probit model. (This last choice of $b_i$ does not require the adaptive stage, and it is included as a benchmark.)

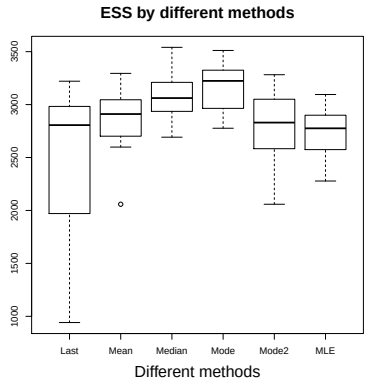


Figure 2.9: Comparing various choices of the data-dependent working parameter based on 25 simulations.

As one may expect, using only the last value from the adaptive stage creates too much variation, although it is the least costly in terms of CPU time and easiest to implement. The latter is true for using the MLE, which eliminates the adaptive stage altogether. Unfortunately these two methods also have the lowest ESS on average. The best performer seems to be using the mode of $b_i^{(t)}$, but it involves a kernel estimation (and mode finding), which can be more costly timewise, albeit for our simulation this was not a problem. For our Table 2.1, we adopted the mean choice because we conjectured that it would represent a practical compromise between statistical efficiency and computational efficiency. Figure 2.9 suggests, however, that the median perhaps is an even better compromise. Further research obviously is needed to assess whether using the median (or another method) is a good compromise in general.

We did, however, find a partial indication for the better performance of the median when we attempted to address both reviewers' question about how we could trust $b_i$'s from the adaptive stage, where the draws from our prototype algorithm itself cannot be trusted. The answer lies in the fact that we are not seeking the theoretically optimal choice of $b_i$, but rather any reasonable choice of it that would result in an algorithm with acceptably satisfactory efficiency. Recall the choice of $b_i$ does not affect the validity of our prototype algorithm as long as it is fixed during the sampling stage. Furthermore, although in the adaptive stage the draws of $\theta$ from our algorithm follow a different distribution than the one for the draws from the sampling stage, the two distributions apparently are close enough that their corresponding distributions for $b_i = G(Y_{obs,i}X_i\theta)$ do not provide significantly different summary statistics, especially for the robust ones such as medians.

To illustrate this point, Figure 2.10 displays the Q-Q plots of the samples of the three components of $\theta$ from the sampling stage against their counterparts from the adaptive stages. We see clearly that although the plots show some visible differences between the two distributions, the differences lie primarily in their tails.

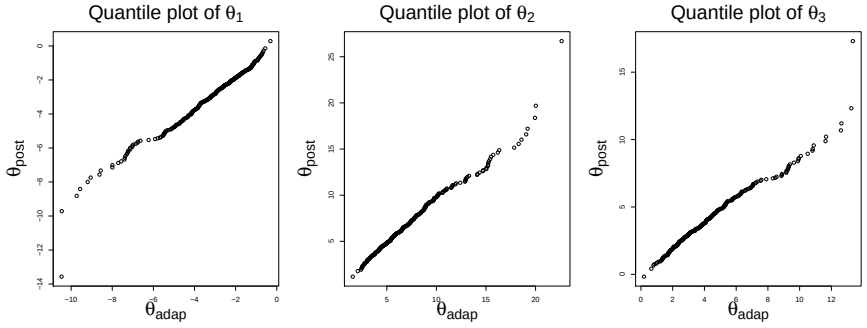


Figure 2.10: Q-Q plots for samples of $\theta_i, i = 1, 2, 3$: sampling stage verses the adaptive stage.

Figure 2.11 displays the corresponding Q-Q plot of samples of randomly selected four (out of $n = 55$) $b_i$ from the sampling stage against those from the adaptive stage. Again we see the major differences lie in tails. To see the impact of adaptation numerically, let $\bar{b}_i^{[S]}$ and $\bar{b}_i^{[A]}$ be the sample means of $b_i$ from the sampling stage and the adaptive stage respectively. Then for the same data underlying Figure 2.10 and Figure 2.11, we have (recall $n = 55$ for our lupus nephritis data set)

$$\frac{1}{55}\sum_{i=1}^{55}\left|\bar{b}_i^{[S]} - \bar{b}_i^{[A]}\right| = 0.00303, \tag{2.25}$$

which is only one third (0.00303/0.918=0.0033) of a percent compared with $\sum_{i=1}^{55}\bar{b}_i^{[S]}/55 = 0.918$. If we replace the sample means in (2.25) by their sample median counterparts, then the absolute difference will be even smaller: 0.00226. The corresponding relative difference compared with the average of the sample medians is $0.00226/0.926 = 0.0024$, only one quarter of a percent. We therefore have rather good empirical verification that the lack of proper convergence during the adaptive stage had non-significant impact on our overall findings.

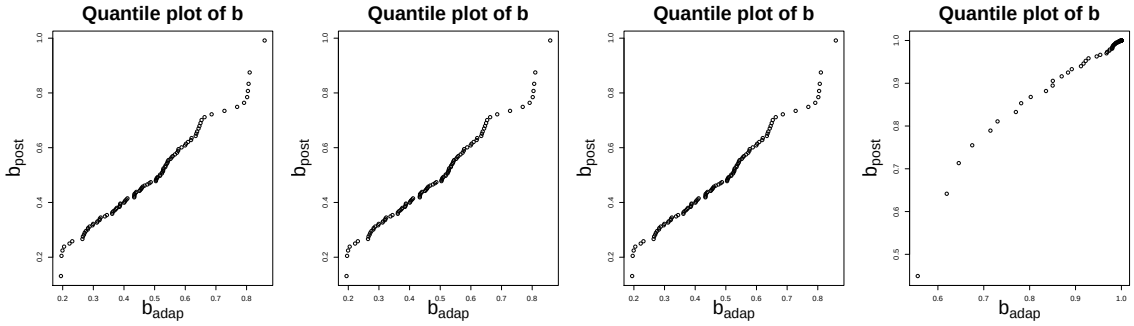


Figure 2.11: Q-Q plot: samples of $b_i$ (for four different $i$'s) from the sampling and adaptive stages.

It is also worthy to point out the trade-off between the number of iterations during the adaptive stage and the sampling stage. Practically, the length of the adaptive stage is likely to be positively correlated with the quality of our choices of $b_i$'s for the sampling stage. However, the lengths of the two stages obviously compete with each other for a total given computational budget. There is hence a trade-off between a longer sampling stage with a less effective choice of $b$ versus a shorter sampling stage but with a more effective $b$. We still need to develop a good practical guideline for such a trade-off, as well as for a number of other trade-offs discussed in the next Section.

## 2.4 Concluding remarks

The primary purpose of this article was to provide initial evidence of the potential of residual augmentations, proposed recently (Yu and Meng (2011)). On the theoretical side, we demonstrated the possibility of establishing numerically rather tight universal bounds (e.g., 1/8 or 1/5) by utilizing a unique "competing" nature of the interweaving strategy, as well as the existence of the "safe zone" and the resulting convenient minimax choice ASIS. At the same time, we uncovered a somewhat unexpected phase transition behavior, which makes the issue of robustness-efficiency trade-off particularly critical and tricky. Mathematically, the situation reminds us of AR(1) type of time series models (see Peña et al. (2001)), where the unit root serves as the only boundary between a stationary region and an explosive region. It is well known that uncertainties in identifying a unit root could lead to rather different statistical properties from what were intended (e.g., Meng and Xie (2013)). Analogously

errors in computing regression slopes for residual augmentations could mean the difference between delivering nearly independently and identically distributed draws and producing almost identical draws (because of extremely poor mixing)!

Therefore, much needs to be done in order to identify situations where we can push the data-dependent residual schemes to achieve their maximal efficiency, and where it is too dangerous to do so and hence we should stay with robust "data-free" schemes such as ASIS. Whereas we succeeded in establishing such results for a class of normal/independent models, we nevertheless benefited from the conditional normality inherited in such a class of models and the symmetric nature of $(\theta, Y_{mis})$ as in (2.13). We imagine the task is rather challenging in general because without normality of some sort (marginal or conditional), the analytic manipulation of maximal correlations is typically intractable. Furthermore, the three maximal correlations in (2.3) generally cannot be mapped into the same triangle because (for example) the function (i.e., transformation) of $\theta$ that leads to its maximal correlation with $Y_{mis}$ may not be the same function for maximizing its correlation with $\tilde{Y}_{mis}$. This would render the geometric expression (2.12) inapplicable, at least not directly. Nevertheless, given the general difficulties in establishing useful bounds for convergence rates for MCMC (see various chapters in Brooks et al., 2011), we are encouraged by the preliminary theoretical results reported in Section 2.3.

On the algorithmic side, as we have seen from the Probit models, there are at least two issues we need to deal with effectively in order to fully realize the potential of residual augmentations, with or without interweaving. The first is how to find a good compromise between statistical efficiency, which requires us to stay as closely as

possible to the theoretically optimal residuals (under whatever criterion adopted), and implementation/computational efficiency, which often requires simple approximations to the optimal residual for effective execution of the $\tilde{Y}_{mis} \to \theta$ step. The second is that even when we know how to carry out the $\tilde{Y}_{mis} \to \theta$ step in theory, its actual implementation can have a significant impact on the overall competitiveness of the resulting algorithm. As seen in Section 2.3.4, different implementations of the nested Gibbs sampler have led to very different algorithmic efficiency.

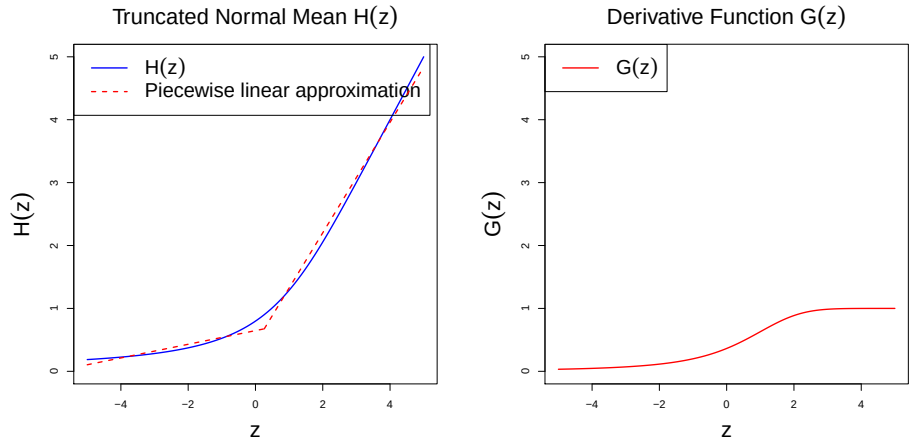


Figure 2.12: The left plot is the $H$ function with a two-piecewise linear approximation; the right plot is the derivative function $G(z) = H'(z)$.

We are also working on finding a better approximation to the actual residual than our current linear approximation. As is evident from the left panel in Figure 2.12, the conditional mean function $H(z)$ can be approximated well by a two-piecewise linear function. That is, we can find two suitable derivative values (see the right panel) $G(z)$ as the $b_i$'s for our residual augmentations $\tilde{Y}_{mis,i} = Y_{mis,i} - b_i X\theta$ depending on the value of $X\theta$ (from the left panel, choosing 0 as the connecting point of the two linear pieces seems to be both effective and convenient). However better approximations do not

transfer to better algorithms unless the added computational burden does not unduly offset the gain in statistical/algorithmic efficiency.

Furthermore, for our probit regression we have constructed only DRA, mainly because in this case DRA is simpler in construction than IRA due to the fact that $\mathrm{E}[Y_{mis,i}|\theta, Y_{obs}] = \mathrm{E}[Y_{mis,i}|\theta, Y_{obs,i}]$, permitting component-wise (as indexed by $i$) calculations (see (2.21)), which is not the case for $\mathrm{E}[\theta|Y_{mis}, Y_{obs}]$. However, for each component $i$, we notice that the probit model depends on $\theta$ only through $X_i\theta$. Hence it is possible to consider forming component-wise IRA in the form of $\tilde{Y}_{mis,i} = X_i\theta - \mathrm{E}[X_i\theta|Y_{mis,i}, Y_{obs,i}]$, which would still render *component-wise zero posterior correlation*:

$$\mathrm{Cov}(\tilde{Y}_{mis,i}, Y_{mis,i}|Y_{obs,i}) = 0, \quad i = 1, \ldots, n. \tag{2.26}$$

Although component-wise derivation/calculation will render implementation simplicity, we are likely giving up some statistical efficiency because (2.26) does not achieve the actual zero posterior correlation $\mathrm{Cov}(\tilde{Y}_{mis}, Y_{mis}|Y_{obs}) = 0$; note (2.26) implies neither $\mathrm{Cov}(\tilde{Y}_{mis,i}, Y_{mis,j}|Y_{obs,i}) = 0$ for any $i \neq j$, nor $\mathrm{Cov}(\tilde{Y}_{mis,i}, Y_{mis,i}|Y_{obs}) = 0$ for any $i$.

We are currently investigating a number of such trade-off issues between statistical efficiency and computational efficiency (e.g., implementation simplicity). There are many challenges ahead, and what is reported above are only those from our initial investigation. At the same time, we have so many options to explore, from forming DRA and IRA to many of their approximations and variations (e.g., component-wise residuals), and with or without interweaving. In our general pedagogical effort, explaining the difference between regressing $Y$ on $X$ and regressing $X$ on $Y$, to those

who are ingrained in deterministic thinking of a functional relationship, has not been a trivial task. But it is the very existence of these two regression lines that offers us a unified theme to explore and construct MCMC algorithms which come closer to realizing the sweet 3-S dream, a dream we invite all readers to share.

# Chapter 3

# Resting state fMRI data analysis

## 3.1 Introduction

Exploring the human brain structure and understanding how it works has always been a fascinating but extremely difficult challenge. Among the existing methods, functional magnetic resonance imaging (fMRI) is an MRI procedure that measures brain activity by detecting associated changes in blood flow. Specifically, we focus on fMRI using blood-oxygen-level-dependent (BOLD) contrast. In BOLD fMRI, neuronal activity levels are captured by blood flow in the brain. Ever since the 1990's, fMRI has become one major tool in neuroimaging research because it does not require people to undergo shots, surgery, etc. (Huettel et al. (2009)).

Within this field, there are two general approaches researchers take: task based design and resting state design. For task or stimulus based design, the scientific goal is to study which regions or cortex are associated with a specific function(such as vision, hearing, etc.). Subjects are usually lying in the MRI machine, treated in the

control state (doing nothing) or the experimental state (given a specific stimulus) in a pre-fixed order. Data are collected during those two stages and then corresponding statistical analysis can be conducted from there. Some concrete examples include Bullmore and Bassett (2011) and references therein. Fox and Raichle (2007) also provides a simple example of a paradigm that requires subjects to open and close their eyes at fixed time intervals. Please see Figure 3.1 for an example.

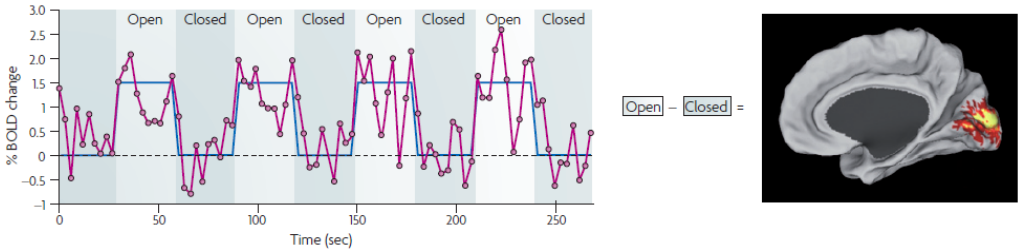


Figure 3.1: **Traditional fMRI analysis and BOLD noise**. Unaveraged BOLD time course from a region in the primary visual cortex during a simple task paradigm that requires subjects to open and close their eyes. The signal intensity difference between the eyes-open condition and the eyes-closed condition is highlighted in the right picture (Fox and Raichle (2007)).

However, there is also spontaneous modulation of the signal which can not be attributed to the experiment. The signal values are usually averaged over time blocks to minimize this spontaneous "noise" in those task based studies. On the other hand, these spontaneous brain activities are themselves of interest for various reasons, among which two are most important: 1. The extra energy consumed by task experiments is not much more than that consumed by the resting state (usually less than 5% more (Raichle and Mintun (2006)). Therefore, in order to study the function of human brain, we need to take into account this major component of energy consumption. 2. The resting state data provides a way of measuring the functional connectivity (which will be defined in the next section) of different regions of the

brain, which can be used to construct a network of the brain. It is an extremely interesting and active research area now to study this neural network, because it provides insights into how the human brain operates as well as certain indications of mental diseases.

In contrast to task based fMRI, in resting state fMRI subjects are usually required to lie in the MRI machine without any stimulus or active thinking. Data are collected during the resting state in the experiment.

## 3.2 Functional connectivity

One key concept that neuroscience researchers are particularly interested in is the functional connectivity. Functional connectivity is the relationship between the neuronal activation patterns of anatomically separated brain regions. Intuitively speaking, people want to explore the underlying network of the human brain, which can be understood as interactions between different regions of the brain. Naturally, functional connectivity should be measuring this interaction or network. For instance, pairwise correlations (correlation matrix) are proposed to measure the linear dependence between different regions; partial correlations (precision matrix, Salvador et al. (2005)) are also used to adjust linear dependence on other parts; moreover, measurements such as mutual information (Bassett and Bullmore (2009)) aim to capture higher order dependence. In this chapter, we only focus on using correlation matrix or partial correlation matrix as the measurement of functional connectivity.

First of all, what does the typical data set look like for calculating the correlation matrix? As emphasized earlier, subjects are not given any task or stimulus

during the process. Data are collected on the voxel (usually a cubic with 3mm length sides) level. However, for various reasons, we won't directly focus on the voxel level analysis. Instead, the brain is usually divided into several regions, according to some anatomical or non-anatomical criterion, such as automated anatomical labeling (AAL, Tzourio-Mazoyer et al. (2002)), Eickhoff-Zilles (EZ, Eickhoff et al. (2005)), independent component analysis (ICA, De Luca et al. (2006)), etc. The fMRI signal of each region of interest (ROI) is obtained by averaging over signals of all voxels within the region. The data set we choose uses AAL criterion. AAL is an automated parcellation method, which projects the divisions in the brain atlas onto brain-shaped volumes of functional data. On average it has 115 ROIs. fMRI signals are collected in discrete time (e.g. every 20s). Therefore, the raw data set for each subject consists of a data matrix with one dimension being the number of ROIs and the other dimension being the number of repeated measurements across time.

In the experiment, there may be issues that are unexpected or out of the experimenter's control (e.g. subject's movement); therefore data are usually preprocessed before analysis. Preprocessing steps usually include motion correction, slice-timing correction, spatial filtering, etc. (see Tanabe et al. (2002) and references therein). Ideally, preprocessing steps can reduce the effects of both instrumental and physiological noise, therefore increasing the signal-to-noise ratio. However, improper preprocessing may destroy signals, e.g. too much smoothing. How to optimize preprocessing itself is an important research topic but beyond the scope of this chapter (see Churchill et al. (2012)).

## 3.3 Question of interest

A lot of the analysis is based on the brain network constructed with the fMRI data. After that the network structure is obtained by thresholding (or non-thresholding) the measures of functional connectivity between ROIs. So first and most importantly, how do researchers define a "correct" and "precise" measurement of the functional connectivity? Pairwise correlations have been widely used in the literature. However, there is one issue that has not been addressed much, to the best of our knowledge: suppose we have observations of a bivariate random variable $(X, Y) : (x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. The standard results hold for estimating their correlation $\rho = \text{corr}(X, Y)$ using the sample correlation when those samples are i.i.d.. In the case of an autocorrelated sample (as often seen in time series), the asymptotic variance of the sample correlation is not clear. Furthermore, the sample correlation is not (at least asymptotically) the most efficient estimator when there exists autocorrelation. The structure of the rest of this chapter is as follows: in Section 3.4.1, 3.4.2 and 3.4.3 we review the separable covariance model, the VAR model and their relationship. Then we propose fitting the separable covariance model to the fMRI data to adjust the temporal correlation. The maximum likelihood estimator outperforms the method of moments, which is the standard method being used in the literature, both in theory and our simulation studies. In the end of the chapter, we compared the fitted results of two estimators to the fMRI data set.

## 3.4 Separable covariance model

### 3.4.1 Review of Kronecker product

The notion of Kronecker product is extremely important in this chapter. Therefore, below is a brief review of the mathematical operation of Kronecker product. Denoted by $\otimes$, Kronecker product is an operation on two matrices, as a generalization of the outer product. Specifically, let $A$ be a $p \times q$ matrix and $B$ be an $m \times n$ matrix, then the Kronecker product $A \otimes B$ is the $pm \times qn$ block matrix:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1q}B \\ \vdots & \ddots & \vdots \\ a_{p1}B & \cdots & a_{pq}B \end{bmatrix}$$

The Kronecker product is bilinear and associative:

$$A \otimes (B + C) = A \otimes B + A \otimes C, \tag{3.1}$$

$$(A + B) \otimes C = A \otimes C + B \otimes C, \tag{3.2}$$

$$(kA) \otimes B = A \otimes (kB) = k(A \otimes B), \tag{3.3}$$

$$(A \otimes B) \otimes C = A \otimes (B \otimes C). \tag{3.4}$$

(3.3) will be used to illustrate the non uniqueness of the maximum likelihood estimator.

### 3.4.2 Separable covariance model

In multivariate statistics, the matrix normal distribution has a very close relationship with the Kronecker product. Let $X$ be an $n \times p$ random matrix, $\Omega_{p\times p}$ and $\Sigma_{n\times n}$ be two positive definite matrices respectively. $X$ is said to have the matrix normal distribution if the probability density function is:

$$p(X|M,\Omega,\Sigma) = \frac{\exp(-\frac{1}{2}\text{tr}[\Omega^{-1}(X-M)^T\Sigma^{-1}(X-M)])}{(2\pi)^{np/2}|\Omega|^{n/2}|\Sigma|^{p/2}}. \tag{3.5}$$

If we stack all the columns of $X$ together, we end up with a vector $\text{vec}(X)$, which follows a multivariate normal distribution. The connection between the matrix normal distribution and multivariate normal distribution is as follows:

$$X \sim MN_{n\times p}(M,\Omega,\Sigma) \Longleftrightarrow \text{vec}(X) \sim N_{np}(\text{vec}(M),\Omega\otimes\Sigma).$$

Matrix normal distribution is one of the most important matrix distributions in statistics and it has been widely used in spatial-temporal modeling such as geographical problem. In those problems, usually one dimension is space and the other dimension is time. The Kronecker product assumes that the spatial covariance is independent of the temporal covariance. Therefore, the covariance between any two observations can be factored as the product of their spatial covariance and temporal covariance, i.e.

$$\text{Cov}(X_{t,i}, X_{t',i'}) = \Sigma_{t,t'}\Omega_{i,i'}. \tag{3.6}$$

In this sense, the covariance matrix is separable into the spatial and temporal

covariance matrices. One most attractive property of this model is it significantly reduces the number of parameters. Often but not always the separability assumption is plausible.

### 3.4.3 Relationship between separable covariance model and VAR model

The vector autoregression (VAR) model is another widely used multivariate statistical model. It is often used to capture the linear dependence among multiple time series. One natural question then rises: what is the relationship between VAR model and the separable covariance model? Does either one model include the other? Or do they have any overlapping? In order to answer these questions, let's first briefly review the basics of VAR model. A $p$-th order VAR is:

$$Y_t = \mu + A_1 Y_{t-1} + \cdots + A_p Y_{t-p} + \epsilon_t,$$

where $Y_t$ is a $p \times 1$ random vector, $\mu$ is a $p \times 1$ constant vector, $A_k$ is a $p \times p$ matrix and $\epsilon_t$ is a $p \times 1$ vector of error term satisfying:

1. $\mathrm{E}[\epsilon_t] = 0$;
2. $\mathrm{E}[\epsilon_t \epsilon_t^T] = V$, where $V$ is a $p \times p$ covariance matrix of $\epsilon_t$;
3. $\mathrm{E}[\epsilon_t \epsilon_{t'}^T] = 0$, whenever $t \neq t'$.

Naturally, if a process satisfies the assumption of VAR model, there exists both spatial and temporal covariances, so it is interesting to find the connection between

these two models. Generally speaking, neither one model is contained in the other one, but they do have some overlapping:

1. Separable covariance model is not contained in VAR model. VAR model has the assumption of weak stationarity. It implies that the marginal variances are the same across time, and also the autocorrelation only depends on the lag. While in the separable covariance model, it is not necessarily true. In the latter case, the temporal covariance structure is more flexible: there could be different marginal covariances across time, i.e. $\mathrm{Var}(Y_t) \neq \mathrm{Var}(Y_{t'})$, or the autocorrelation is not necessarily stationary, i.e. $\mathrm{Corr}(Y_t, Y_{t+1}) \neq \mathrm{Corr}(Y_{t-1}, Y_t)$.

2. VAR model is not contained in separable covariance model. In separable covariance model, one fundamental property is that the temporal correlation structure is the same across space (for different regions). However, in VAR model

$$Y_t = AY_{t-1} + \epsilon_t,$$

   the correlation structure across space is not necessarily the same unless $A$ is multiple of an identical matrix $rI_p$. For example, if $A = \begin{pmatrix} r_1 & 0 \\ 0 & r_2 \end{pmatrix}$ $(r_1 \neq r_2)$, then the covariance is not separable.

3. However, there exists some special case where VAR and separable covariance models overlap. Suppose $A$ is a multiple of an identical matrix (i.e. $rI_p$) and for simplicity we assume the mean is 0, then the VAR model is:

$$Y_t = rY_{t-1} + \epsilon_t.$$

In this case,

$$Y_j = r^{j-i}Y_i + r^{j-i-1}\epsilon_{i+1} + \cdots + \epsilon_j,$$

when $j > i$. From the stationarity, we have $\mathrm{Cov}(Y_j) = r^2\mathrm{Cov}(Y_i) + V$, which implies that $\mathrm{Cov}(Y_i) = \frac{1}{1-r^2}V$. Meanwhile,

$$\mathrm{Cov}(Y_j, Y_i) = r^{j-i}\mathrm{Cov}(Y_i) = \frac{r^{j-i}}{1-r^2}V.$$

Therefore, this particular VAR(1) model corresponds to a separable covariance model with $\Omega = V$ and $\Sigma = \{\Sigma_{ij}\}$, where $\Sigma_{ij} = \dfrac{r^{|j-i|}}{1-r^2}$. In Section 3.4.6 we will show the fitted results of a similar model to this, with the only change of AR(1) temporal correlation to AR(2) temporal correlation.

### 3.4.4 Estimation of covariance matrices

#### Method of moment and MLE

Now we have set up the separable covariance model. The next question is how we can estimate the parameters. In our fMRI problem, we are essentially interested in the spatial correlation matrix so the temporal correlation matrix is nuisance parameters. Let $X_1, \cdots, X_N$ be $N$ i.i.d. samples from the matrix normal distribution $MN_{n\times p}(M = 0, \Omega, \Sigma)$. In the lack of temporal correlation (i.e. $\Sigma = I_n$), the maximum likelihood

estimator of $\Omega$ is the usual sample covariance estimator:

$$\hat{\Omega}_{MLE} = \frac{1}{N}\sum_{k=1}^{N} X_k^T X_k.$$

However, when there exists temporal correlation, this estimator is no longer the MLE. It is still unbiased (the method of moments), but may not be asymptotically most efficient. The MLE of $\Omega$ has been proposed in different papers (Lu and Zimmerman (2004) and therein), but has not been used in the literature of resting state fMRI analysis, to the best of our knowledge. In the next sections, we propose using the maximum likelihood estimator instead of the method of moments. Section 3.4.4 compares the theoretical asymptotic efficiency of these two estimators in a simple but non trivial example; Section 3.4.5 includes simulation studies to confirm the results; Section 3.4.4 reviews the algorithm to calculate the MLE, which is presented by Dutilleul (1999); Section 3.4.6 includes the fitted results of the fMRI data set.

**Asymptotic efficiency of MLE and MoM**

To compare the asymptotic efficiency of two estimators, we consider a simple but non trivial example. Suppose the spatial dimension $p = 2$ and $A = rI_2$. As seen in Section 3.4.3, the following VAR model is also variance separable:

$$\begin{pmatrix} Y_{t,1} \\ Y_{t,2} \end{pmatrix} = \begin{pmatrix} rY_{t-1,1} \\ rY_{t-1,2} \end{pmatrix} + \begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \end{pmatrix},$$

where

$$\begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \end{pmatrix} \sim N\Big( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \Big).$$

It is equivalent to the separable covariance model with

$$\Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and

$$\Sigma = \frac{1}{1-r^2} \begin{pmatrix} 1 & r & r^2 & \cdots & r^{n-1} \\ r & 1 & r & \cdots & r^{n-2} \\ r^2 & r & 1 & \cdots & r^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r^{n-1} & r^{n-2} & r^{n-3} & \cdots & 1 \end{pmatrix}.$$

Suppose $r$ is known and we are interested in estimating $\rho$. Let's consider the following two estimators.

1. Method of moments:

$$\hat{\rho}_{MoM} = \frac{(1-r^2)}{n} \sum_{k=1}^{n} Y_{k,1} Y_{k,2}.$$

2. Maximum likelihood estimator:

$$\hat{\rho}_{MLE} = \text{argmax}_{\rho} l(\rho),$$

where

$$\begin{aligned} l(\rho) &= -\frac{n}{2}\log(1-\rho^2) - \frac{1-r^2}{2(1-\rho^2)}(Y_{1,1}^2 + Y_{1,2}^2 - 2\rho Y_{1,1}Y_{1,2}) \\ &\quad -\frac{1}{2(1-\rho^2)}\sum_{k=2}^{n}\{(Y_{k,1}-rY_{k-1,1})^2 + (Y_{k,2}-rY_{k-1,2})^2\} \\ &\quad +\frac{\rho}{1-\rho^2}\sum_{k=2}^{n}(Y_{k,1}-rY_{k-1,1})(Y_{k,2}-rY_{k-1,2}). \end{aligned}$$

Both estimators are asymptotically unbiased, but their variances differ. The asymptotic variance of $\hat{\rho}_{MoM}$ is $\frac{1+\rho^2}{n^2}(n + \sum_{i\neq j} r^{2(j-i)})$. The asymptotic variance of $\hat{\rho}_{MLE}$ is $\frac{(1-\rho^2)^2}{n(1+\rho^2)}$, which does not depend on $r$. Therefore, the relative efficiency is the ratio

$$\frac{\mathrm{Var}(\hat{\rho}_{MoM})}{\mathrm{Var}(\hat{\rho}_{MLE})} = (\frac{1+\rho}{1-\rho})^2 \frac{n + \sum_{i\neq j} r^{2(j-i)}}{n}. \tag{3.7}$$

The relative efficiency depends on the autocorrelation $|r|$. It coincides with our intuition: as $|r|$ increases, $\hat{\rho}_{MoM}$ becomes less efficient by ignoring this autocorrelation. More importantly, when the autocorrelation $r$ is fixed, this ratio monotonically increases as the spatial dimension $n$ increases, meaning that the MLE is more efficient in high dimensional problems. Other than this simple model, the analytical form of the relative efficiency of these two estimators is beyond the scope of this chapter. However, in Section 3.4.5, a comprehensive simulation study is conducted to investigate this question: is $\hat{\Omega}_{MLE}$ more efficient than $\hat{\Omega}_{MoM}$? If so, by how much? But before the simulation study, let's first see how the MLE of $\Omega$ can be obtained in a general separable covariance model.

**Algorithm to calculate MLE**

The analytical form of the MLE is not available, but we can use an existing algorithm to calculate it. Several different versions of the algorithm were derived by Mardia and Goodall (1993), Dutilleul (1999), and Brown et al. (2001). Here we chose the one by Dutilleul (1999).

First of all, it is worth noting that $\Omega \otimes \Sigma = \frac{1}{a}\Omega \otimes a\Sigma (a \neq 0)$. Therefore, the MLE exists up to a normalizing constant: if $(\hat{\Omega}, \hat{\Sigma})$ is the MLE, then $(\frac{1}{a}\hat{\Omega}, a\hat{\Sigma})$ yields to the same likelihood for any positive $a$. Suppose $X_1, X_2, \cdots, X_N$ are i.i.d. samples from the matrix normal distribution $MN_{n\times p}(M, \Omega, \Sigma)$, the log-likelihood function is:

$$l = -\frac{Nnp}{2}\log(2\pi) - \frac{Np}{2}\log(|\Sigma|) - \frac{Nn}{2}\log(|\Omega|) - \frac{1}{2}\sum_{k=1}^{N} tr\{\Sigma^{-1}(X_k - M)\Omega^{-1}(X_k - M)^T\}. \tag{3.8}$$

Dutilleul (1999) presented an iterative algorithm to calculate the MLE $\hat{\Omega}$ and $\hat{\Sigma}$. Let $\hat{M}, \hat{\Omega}$ and $\hat{\Sigma}$ be the MLE of $M, \Omega$ and $\Sigma$ respectively. Obviously,

$$\hat{M} = \frac{1}{N}\sum_{k=1}^{N} X_k = \bar{X}. \tag{3.9}$$

$\hat{\Omega}$ and $\hat{\Sigma}$ are computed by iterating the following process:

$$\begin{cases} \hat{\Omega} = \frac{1}{Nn}\sum_{k=1}^{N}(X_k - \bar{X})^T\hat{\Sigma}^{-1}(X_k - \bar{X}); \\ \hat{\Sigma} = \frac{1}{Np}\sum_{k=1}^{N}(X_k - \bar{X})\hat{\Omega}^{-1}(X_k - \bar{X})^T. \end{cases}$$

Moreover, Dutilleul (1999) also derived the necessary and sufficient condition for

when the MLE exists:

$$N \geq \max\{\frac{n}{p}, \frac{p}{n}\} + 1.$$

In our data set, each subject corresponds to a data matrix $X_k$ and there are several subjects in one study. Because the condition requires $N \geq 2$, therefore we can not estimate the MLE for each subject respectively. Instead, we can only estimate the common covariance matrices if we assume $X_k$ is generated independently from the same distribution $MN_{n\times p}(M, \Omega, \Sigma)$. This is a draw back of the MLE in this problem. However, in Section 3.4.6 we present an empirical Bayes method, where the individual covariance matrix of $X_k$ can be estimated respectively. Moreover, for discussions on the convergence and stability of the MLE, see Lu and Zimmerman (2004).

### 3.4.5 Simulation studies on the comparison of MoM and MLE

The theoretical comparison of the two estimators for high dimensional correlation matrices is beyond the scope of this chapter. However, we can still compare their empirical efficiency by simulation studies.

The first simulation study compares the performance of two estimators under different set up of the spatial and temporal correlation matrices. Suppose $X_k \sim MN_{n\times p}(0, \Omega_{p\times p}, \Sigma_{n\times n}), k = 1, 2, \cdots, N$. We chose $n = p = 100$ to mimic the fMRI data. The performance was evaluated based on the mean squared error (MSE) of the estimator, i.e. $||\hat{\Omega} - \Omega||_2$. For each combination of $\Omega, \Sigma$, we simulated $N$ independent samples from the distribution $MN_{n\times p}(0, \Omega, \Sigma)$ and calculated $\hat{\Omega}_{MoM}, \hat{\Omega}_{MLE}$ and then

the MSE. This process was repeated 500 times to calibrate the variability. Below are the four combinations of $\Omega, \Sigma$ we chose:

1. $\Sigma_1 = I_n$, e.g. there is no temporal correlation. $\Omega_1$ has an AR(1) structure, i.e. $\text{Corr}(X_{i,j,k}, X_{i,j,k'}) = |\rho|^{|k-k'|}$, where $\rho$ is the coefficient parameter in $\Omega_1$.

2. Both $\Omega_2$ and $\Sigma_2$ have an AR(1) structure, with coefficients $\rho_1$ and $r_1$ respectively.

3. $\Sigma_3 = \Sigma_2$ still has an AR(1) structure. To relax the assumption on the spatial correlation, we let $\Omega_3$ be a random matrix generated from $W^{-1}(\Omega_2, \nu = 200)$, where $W^{-1}(V, \nu)$ is the inverse Wishart distribution with parameters $(V, \nu)$.

4. Both $\Omega_4$ and $\Sigma_4$ are random matrices: $\Omega_4 = W^{-1}(\Omega_2, \nu = 200)$ and $\Sigma_4 = W^{-1}(\Sigma_2, \nu = 200)$.

The results are summarized in Figure 3.2: in the first case, where there is no temporal correlation $\Sigma = I_n$, both estimators have the same efficiency asymptotically. However, $\hat{\Omega}_{MoM}$ outperforms $\hat{\Omega}_{MLE}$ when sample size is small ($N < 10$). In the latter three cases, where there is temporal correlation, $\hat{\Omega}_{MLE}$ outperforms $\hat{\Omega}_{MoM}$, especially when sample size is small. When both $\Omega$ and $\Sigma$ are structured – AR(1), $\hat{\Omega}_{MLE}$ is more efficient than $\hat{\Omega}_{MoM}$, by a factor between 2.2 and 3.9; when $\Omega$ and $\Sigma$ are random matrices, $\hat{\Omega}_{MLE}$ is more efficient than $\hat{\Omega}_{MoM}$ by only a factor between 1.1 and 1.3. Note that our focus is to improve the efficiency of the estimator especially when sample size is small (ideally $N = 1$), therefore the reduction in the MSE is still non negligible.

The second simulation study involves the spatial and temporal dimensions $n$, $p$. We fixed sample size $N$ to be 5 and let $n = p$ vary. The comparison of the estimators

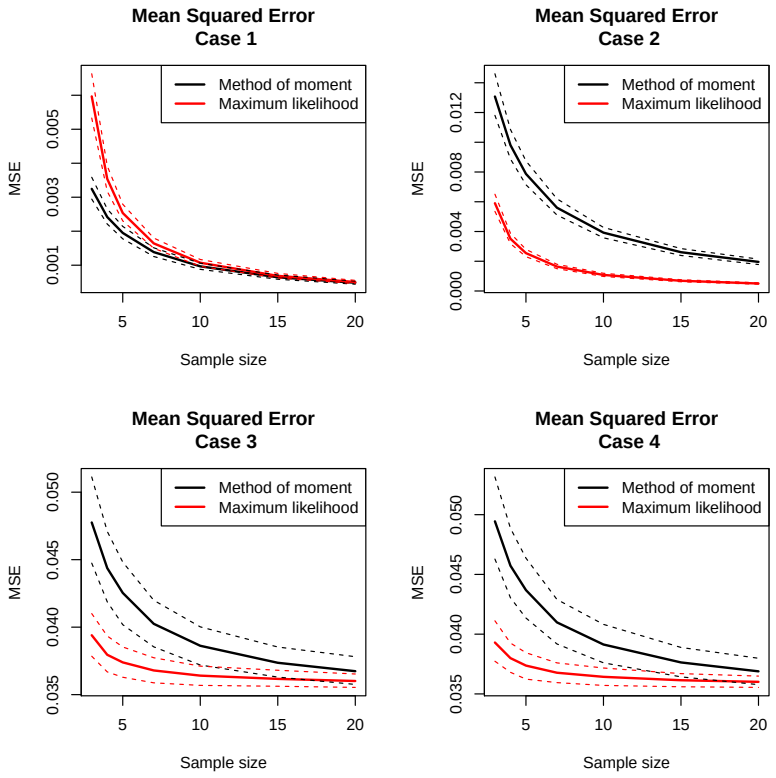


Figure 3.2: Mean squared error of $\hat{\Omega}_{MoM}$ and $\hat{\Omega}_{MLE}$ in different scenarios.

is still based on the MSE. AR(1) structured correlation matrices are used for both $\Omega$ and $\Sigma$. The results are summarized in Figure 3.3.

Clearly, the relative efficiency increases as the dimension ($n = p$) increases. This is inline with our theoretical comparison (3.7) in the simple model. In the fMRI problem, the dimension is usually between 100 and 300, therefore the improvement in the statistical efficiency is around the factor of 3.

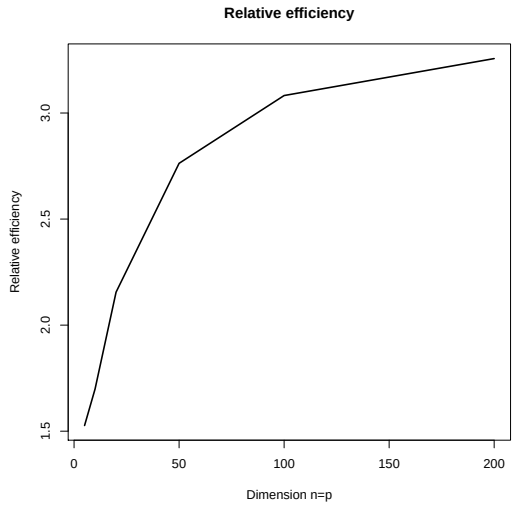


Figure 3.3: Relative efficiency of $\hat{\Omega}_{MLE}$ to $\hat{\Omega}_{MoM}$ under different dimensions.

### 3.4.6 Application to the resting state fMRI data

#### Introduction to the data set

Now we have seen that when the assumption holds, the maximum likelihood estimator is asymptotically more efficient than the method of moments. In the resting state fMRI, time series data are collected for ROIs. Due to factors such as physiological cycles, machine noise, etc., there exists temporal correlation in the data. In the next section, we shall fit the separable covariance model to obtain $\hat{\Omega}_{MLE}$ and compare it with $\hat{\Omega}_{MoM}$.

The data sets we used are released by Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC (2013)). Seven data sets containing pre-processed fMRI brain images are included:

- The Kennedy Krieger Institute, Baltimore, MD, USA ("KKI")
- The Donders Institute, Nijmegen, The Netherlands ("NeuroIMAGE")

- New York University Medical Center, New York, NY, USA ("NYU")
- Oregon Health and Science University, Portland, OR, USA ("OHSU")
- Peking University, Beijing, P. R. China ("Peking_1", "Peking_2", "Peking_3")

AAL criterion was chosen (116 ROIs) to aggregate the voxel level data. There are 458 participants aged between 7 and 21 in total in these seven data sets. These participants can be diagnosed as different levels of attention deficit-hyperactivity disorder (ADHD). For simplicity, we divide them into ADHD and non-ADHD. The initial goal is to estimate the spatial correlation structure (as a measurement of functional connectivity) for ADHD and non-ADHD participants respectively and to see if there is any significant difference between them.

**Separable covariance model with AR(2) and damped oscillator temporal correlation**

The separable covariance model assumes the same temporal correlation structure across different ROIs, therefore we first plot the autocorrelation (ACF) plots (Figure 3.4) for selected regions in both groups.

Clearly, the temporal correlation exists and could potentially inflate the variance of the estimator. There is also one consistent feature among these different ROIs: the autocorrelation is periodic and decaying. The explanation involves the physiological cycles of human brains and other factors and therefore is not straightforward. But we can capture this pattern by a simple time series model, for example AR(2) model – AR(1) model does not suffice here because of the periodic autocorrelation. More specifically:

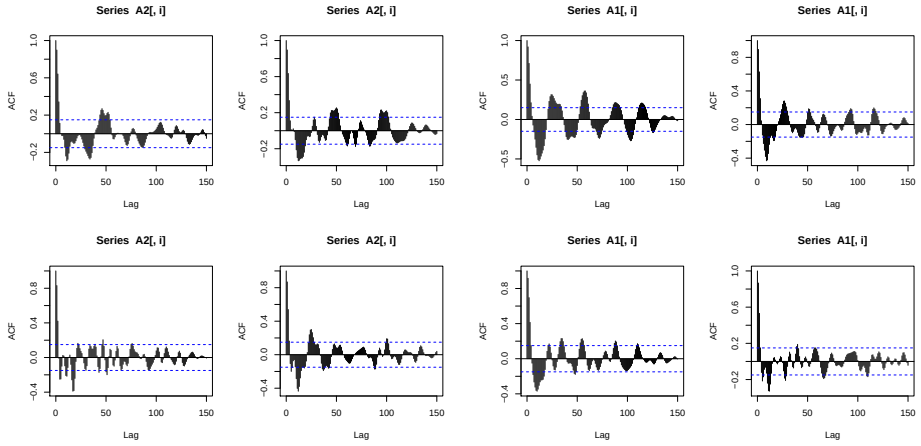


Figure 3.4: Autocorrelation plots from ADHD and normally developed children.

$$X_k \sim MN_{n\times p}(M, \Omega, \Sigma)$$

and here we assume $\Sigma$ is the autocorrelation matrix from the AR(2) model

$$Y_t = r_1 Y_{t-1} + r_2 Y_{t-2} + \epsilon_t.$$

Therefore $\Sigma$ can be viewed as a function of $r_1, r_2$, i.e. $\Sigma(r_1, r_2)$. A structured temporal correlation is preferred here as compared with an unstructured one because of the following reasons: 1. highly reduced number of parameters; 2. easier interpretation of the model; 3. the structured temporal correlation has a good fit to the data. Since we assume the AR(2) model in the temporal correlation, the algorithm to compute the MLE is slightly different from Dutilleul's (designed for unstructured ones). The full description of the algorithm is as follows:

1. Calculate $\hat{M} = \bar{X} = \frac{1}{N}\sum_{k=1}^{N} X_k$.

2. Given $\hat{r}_1, \hat{r}_2$, calculate $\hat{\Sigma} = \Sigma(\hat{r}_1, \hat{r}_2)$. Then calculate $\hat{\Omega} = \frac{1}{Nn} \sum_{k=1}^{N} (X_k - \bar{X})^T \hat{\Sigma}^{-1} (X_k - \bar{X})$.

3. Given $\hat{\Omega}$, calculate $\hat{\Sigma}$. In this case, if we do Cholesky decomposition $\hat{\Omega}^{-1} = LL^T$ and define $Z_k = X_k L$, then we have removed the spatial correlation, which essentially means we have $p$ independent observations for each $Z_k$, and in total $Np$ samples from the same AR(2) process with coefficients $r_1, r_2$. As an approximation, the conditional MLE of $\hat{r}_1, \hat{r}_2$ can be obtained by ordinary linear regressions.

We repeat step 2 and 3 until the algorithm converges. Among the seven data sets, the NYU data set has the largest sample size 139. Therefore we will present the fitted results to the NYU data set in the rest of this chapter. Similar results are obtained on the other data sets. Using the above algorithm, the estimated spatial correlation matrix $\hat{\Omega}_{MLE}$ together with $\hat{\Omega}_{MoM}$ are shown in Figure 3.5:

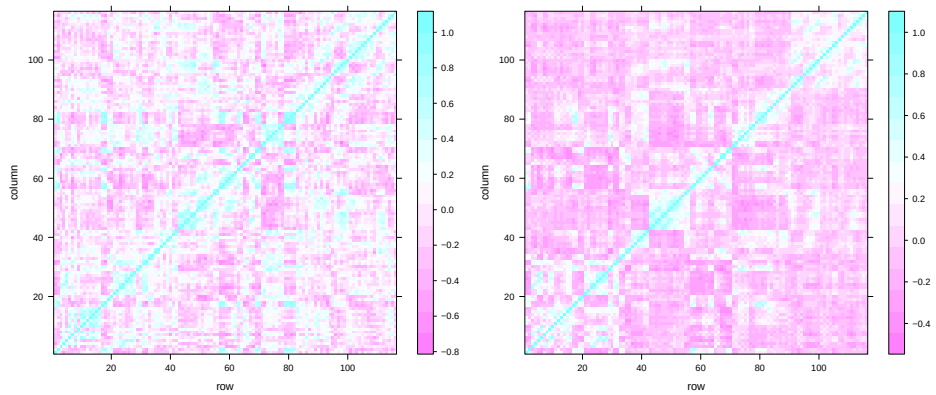


Figure 3.5: $\hat{\Omega}_{MoM}$ and $\hat{\Omega}_{MLE}$ for normally developed children in NYU data set: the left plot is $\hat{\Omega}_{MoM}$ and the right plot is $\hat{\Omega}_{MLE}$.

The difference in these two plots lies in two aspects: 1. there are less extreme values in the correlations in $\hat{\Omega}_{MLE}$ than in $\hat{\Omega}_{MoM}$, especially for very negative values; 2. $\hat{\Omega}_{MLE}$ is more structured than $\hat{\Omega}_{MoM}$. On the other hand, the AR(2) model, though simple, does not provide a good enough fit to the actual autocorrelation function. In order to capture the periodicity and decaying pattern, we considered a damped oscillator function to model the temporal correlation function. The damped oscillator function has the form $e^{-ah}b\sin(ch+d)$, where $a, b, c, d$ are four parameters and $h$ is the lag, i.e. two observations with a lag of $h(h > 0)$ have correlation $e^{-ah}b\sin(ch+d)$. This is very similar to our previous model, except replacing the AR(2) temporal correlation structure by this damped oscillator function. The comparison of the estimated spacial correlation matrices under these two models is in Figure 3.6:

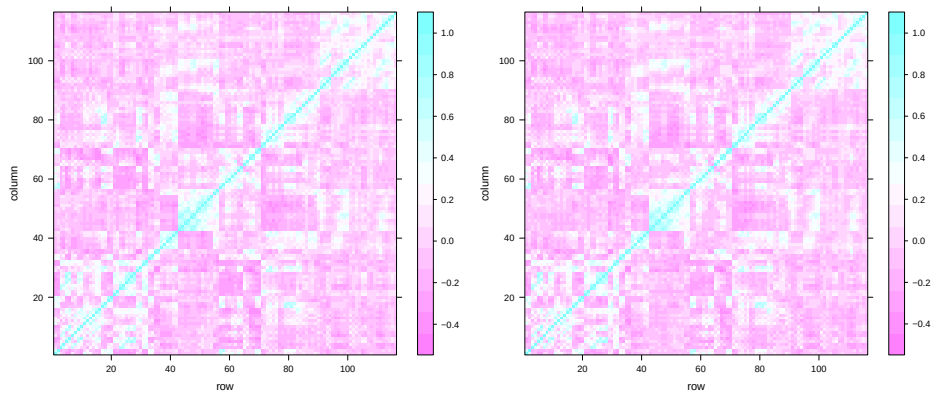


Figure 3.6: $\hat{\Omega}_{MLE}$ under the AR(2) and damped oscillator model for normally developed children in NYU data set: the left plot is $\hat{\Omega}_{MLE}$ under the AR(2) model and the right plot is $\hat{\Omega}_{MLE}$ under the damped oscillator model.

The difference is not as big as the comparison of $\hat{\Omega}_{MLE}$ with $\hat{\Omega}_{MoM}$. However, if we look at the goodness of fit in temporal correlation, we can see that the damped oscillator model provides better fit than the AR(2) model (see Figure 3.7).

Now a natural question is: we have these two estimators $\hat{\Omega}_{MoM}, \hat{\Omega}_{MLE}$, which one is better in this problem? Unlike in the simulation studies, where the ground truth is known, the comparison of these two estimators in the resting state fMRI data set is not easy. One reason is that little is well understood in the area; another reason is that fMRI data is quite noisy, therefore a small improvement of the estimator may not be clearly reflected in any test. Nevertheless, there is some consensus from previous studies. This will be described in details in the next section, but simply speaking, the correlations among some groups of ROIs are consistently stronger. For example, the primary visual cortex is specialized in processing information about static and moving objects. It consists of four ROIs in our data set and these ROIs are consistently reported as highly correlated in the literature. As a result, our estimation of correlations in this cortex may give us some indication on the goodness of fit. From $\hat{\Omega}_{MoM}$, the estimated correlations among the cortex has mean of 0.51 and minimum of 0.22; while from $\hat{\Omega}_{MLE}$, the estimated correlations among the cortex has mean of 0.62 and minimum of 0.42. Similar results are obtained for the other cortexes. This comparison of course can not directly prove that our estimator is better than the method of moments, but at least provides us more confidence since our results are inline with the consensus in the field.

**An empirical Bayes model**

However, as mentioned before, there is one common draw back of the MLE under these models: the condition that the sample size $N > 1$ is required, therefore only the common spatial correlation matrix $\Omega$ can be estimated instead of each individual $\Omega^k$.

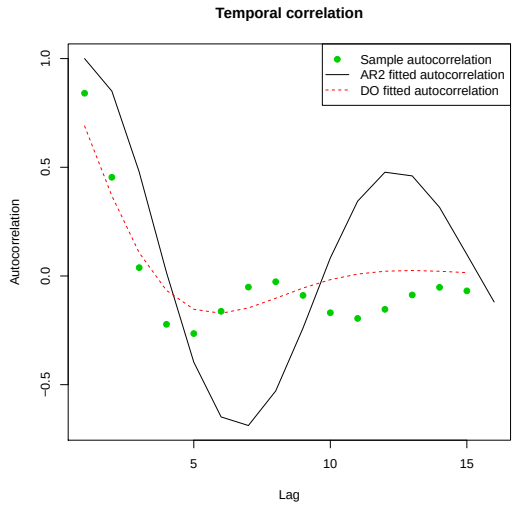


Figure 3.7: Sample ACF plot and fitted autocorrelation functions with AR(2) and Damped Oscillator model: the dots represent sample autocorrelations, the solid line represents the fitted autocorrelation function by the AR(2) model and the dashed line represents the fitted autocorrelation function by the Damped Oscillator model.

In this section, we present an empirical Bayes model, which not only can estimate the individual $\Omega^k$, but also incorporates the prior knowledge from previous studies.

There are quite a few previous studies in this field, among which there is some consensus. van den Heuvel and Hulshoff Pol (2010) reports that:

" A number of group resting-state studies have consistently reported the formation of functionally linked resting-state networks during rest. These studies, although all using different groups of subjects, different methods (e.g. seed, ICA or clustering) (Beckmann et al., 2005, Biswal et al., 1995, Damoiseaux et al., 2006, De Luca et al., 2006 and Salvador et al., 2005a; Van den Heuvel et al., 2008a) and different types of MR acquisition protocols, show large overlap between their results, indicating the robust formation of functionally linked resting-state networks in the brain during rest."

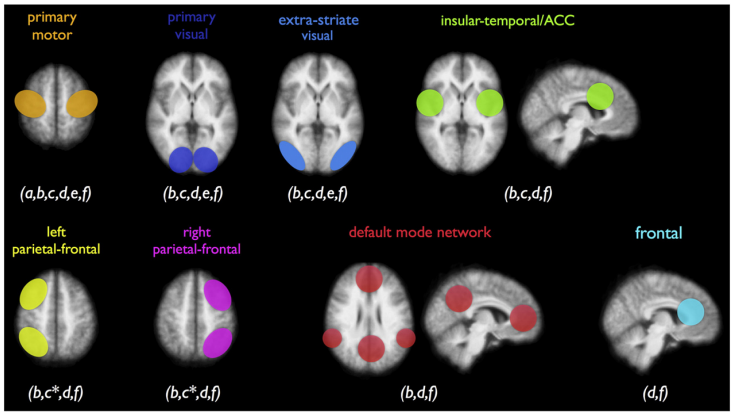


Figure 3.8: Eight subnetworks consistently reported by different studies (van den Heuvel and Hulshoff Pol (2010)).

And they also included Figure 3.8 highlighting eight most consistent reported resting state subnetworks. Intuitively speaking, the interactions within subnetworks are generally higher than those between subnetworks. In order to incorporate this domain knowledge, we assumed a blocked spatial correlation matrix. Within each subnetwork, there is one parameter $\rho_i$ as the correlations within the subnetwork, between different subnetworks, there is one single "background" correlation parameter $\rho_0$. So there are nine parameters in total and the correlation matrix has the following form:

$$\Omega_0 = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \rho_0 & \cdots & \rho_0 & \rho_0 \\ \rho_1 & 1 & \rho_1 & \rho_0 & \cdots & \rho_0 & \rho_0 \\ \rho_1 & \rho_1 & 1 & \rho_0 & \cdots & \rho_0 & \rho_0 \\ \rho_0 & \rho_0 & \rho_0 & 1 & \cdots & \rho_0 & \rho_0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_0 & \rho_0 & \rho_0 & \rho_0 & \cdots & 1 & \rho_8 \\ \rho_0 & \rho_0 & \rho_0 & \rho_0 & \cdots & \rho_8 & 1 \end{pmatrix}$$

It is hard to find relevant information on those parameters $\rho_i$. Therefore, we adopted an empirical Bayes method: the nine correlation parameters are first fitted from the data set and then used as the prior. More specifically, we first computed the unstructured MLE $\hat{\Omega}_{MLE}$, and then find $\hat{\rho}_0, \cdots, \hat{\rho}_8$ such that $||\hat{\Omega}_{MLE} - \Omega_0(\hat{\rho}_0, \cdots, \hat{\rho}_8)||_2$ is minimized. For computational simplicity, we use the conjugate prior for the spatial covariance matrix – inverse Wishart distribution, with parameters $(\nu_0\Omega_0, \nu_0)$. Then the Bayesian model is:

$$\begin{aligned} \Omega|\nu_0, \Omega_0 &\sim W^{-1}(\nu_0\Omega_0, \nu_0) \\ \pi(M, r_1, r_2|\Omega) &\propto 1 \\ X|M, \Omega, r_1, r_2 &\sim MN_{n\times p}(M, \Omega, \Sigma(r_1, r_2)) \end{aligned}$$

A Gibbs sampler to sample from the posterior distribution is:

- Sample $\Omega|M, \Sigma, X \sim W^{-1}(\Omega_N, \nu_N)$, where $\nu_N = \nu_0 + Nn$ and $\Omega_N = \nu_0\Omega_0 + S_N$.

$$S_N = \sum_{i=1}^{N}(X_i - M)^T\Sigma^{-1}(X_i - M).$$

- Sample $M|\Omega, \Sigma, X$: $\text{vec}(M) \sim N_{np}(\text{vec}(\bar{X}), \frac{\Omega\otimes\Sigma}{N})$.

- Sample $\Sigma|\Omega, M, X$, i.e. sample $r_1, r_2|\Omega, M, X$. If we let $Z_k = (X_k - M)L^T$, where $LL^T = \Omega$ is the Cholesky decomposition, then we have removed spatial correlations. In other words, each $Z_k$ can be treated as $n$ independent AR(2) processes, with mean 0 and common coefficients $r_1, r_2$. Exact conditional distributions can be drawn, however, if we ignore the contribution of the first two samples in each process, the approximation is close enough. In other words, we can draw $r_1, r_2|\Omega, M, X$ from Bayesian regression model as a good approximation.

Notice that $\nu_0$ represents the equivalent sample size in the prior. If we chose $\nu_0$ to be far smaller than $Nn$, we essentially put most weight on our data and therefore the result is close to the MLE. More importantly, the Bayesian method provides a way of estimating the spatial correlation matrix for each subject individually. If we let each individual have its own parameters $M^k, \Sigma^k$ and $\Omega^k$, then we can calculate the posterior mean of $\Omega^k$ respectively.

In order to let the data speak, a relatively small $\nu_0 = 10$ was chosen. For each subject $k$ in the data set, we calculated the posterior mean $\hat{\Omega}^k$. From the results we can see that there is a lot of variability across different subjects. For instance, $\hat{\Omega}^k_{12}(k = 1, 2, \cdots, N)$ has a mean of 0.45 and a standard deviation of 0.24. The variability across subjects suggest that assigning a common covariance matrix may not be appropriate here and our empirical Bayes model does adjust for that.

Another question we can potentially answer with this model is whether any difference in the spatial correlation matrix can be found between ADHD and non-ADHD children. Among these 139 subjects, 97 are normally developed children and the other 42 are diagnosed with ADHD. If we treat each element of the spatial correlation matrix as a feature, there will be $\frac{p(p-1)}{2} = 6670$ features in total. If we apply two sample t tests to all the features, we may be able to find the important features that can distinguish the two groups. Below is a histogram of the t test statistics for all the features:

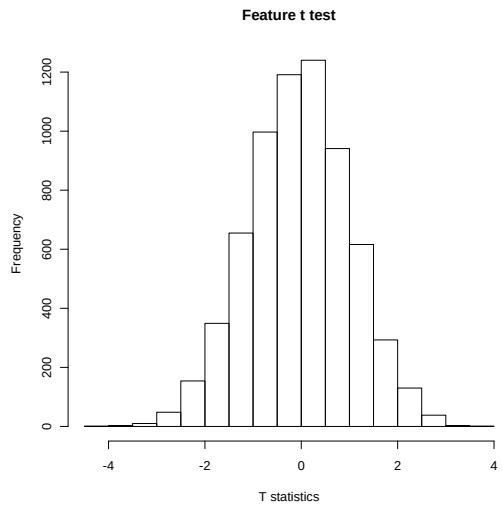


Figure 3.9: Two sample t test statistics for all 6670 features.

Right now it is not clear to us if these differences are just sampling error or there

are informative features that can predict the state of the subject (ADHD or non-ADHD). If the latter were true, that means we can diagnose (with some probability) ADHD by examining the correlations among some ROIs.

## 3.5 Concluding remarks

Resting state fMRI is an active research area, where a lot of studies focus on building the brain networks with functional connectivity. Therefore it is of great interest and importance to improve the estimate of functional connectivity, which is our focus in this chapter. To the best of our knowledge, most studies use the standard estimator $\hat{\Omega}_{MoM}$ from the independent samples by ignoring the temporal correlation. By imposing the separable covariance model, we proposed using $\hat{\Omega}_{MLE}$ instead of $\hat{\Omega}_{MoM}$. And the statistical efficiency of $\hat{\Omega}_{MLE}$ is demonstrated both in theoretical comparison and simulation studies. More importantly, we also present an empirical Bayes model, under which the individual spatial correlation matrix can be estimated.

Much work needs to be done in this problem. For example, theoretical comparison and simulation studies verify the superiority of the MLE when the separability assumption holds. But how robust is this result? What is its performance when the assumption is violated? More work needs to be done before we claim our proposed estimator is better in similar data sets. Another problem is that we assume a structured temporal correlation in the model but not the spatial correlation. One key reason for that is researchers currently do not have a good understanding of the spatial correlation yet, so we choose to "let the data speak". However, the high dimension nature ($p \approx 200$) does not encourage us to do so. A structured correlation matrix may lead

to better efficiency.

Moreover, we can further consider fitting a hierarchical model

$$\begin{aligned}
\Omega|\Omega_0 &\sim \pi(\Omega|\Omega_0) \\
\Omega^k|\Omega,\nu_0 &\sim W^{-1}(\nu_0\Omega,\nu_0) \\
\pi(M^k,r_1^k,r_2^k|\Omega^k) &\propto 1 \\
X_k|M^k,\Omega^k,r_1^k,r_2^k &\sim MN_{n\times p}(M^k,\Omega^k,\Sigma(r_1^k,r_2^k)).
\end{aligned}$$

to not only allow estimation for each individual correlation matrix but also take into account of information together. As for now, how to choose the conditional distribution $\pi(\Omega|\Omega_0)$ to allow easy posterior distribution sampling remains a difficult part.

More importantly, estimating the spatial correlation matrices is only the intermediate step in the fMRI study. A lot of problems can be attacked and solved from there. For example, with the tool of brain imaging data, many researchers have devoted to using fMRI or other imaging data to facilitate medical diagnosis. Craddock et al. (2009) has shown the significant difference in the brain networks among healthy people and clinically depressed patients. Our data sets consist of ADHD and normally developed children, therefore it is of great interest to see if any classifier can be built to predict their states. In this case, the individual spatial correlation matrix $\hat{\Omega}_k$ is treated as features. With the empirical Bayes model, we are hoping to construct better features than the standard methods in terms of classification accuracy.

# Appendix A

# Appendix

## Proof of condition (1.16) for AR(1) process

In this appendix, we prove that $\{U_n\}$ satisfies the mixing condition (1.16) given by the generalization of Billingsley's theorem for an AR(1) process.

Proof: Recall that we want to prove for $\{U_n \triangleq \Phi(X_n)\}$, the $\alpha$ mixing coefficient satisfies the condition $\sum_{n=1}^{\infty} n^2 \alpha_n^{\frac{1}{4}} \leq \infty$. First note that $\{\Phi(X_n)\}$ has the same $\alpha$ mixing coefficient as $\{X_n\}$. This is because $\Phi(\cdot)$ is a one-to-one transformation. Then we only need to prove the mixing condition (1.16) for $\{X_n\}$.

The $\rho$ mixing coefficient $\rho(n)$ of the sequence $\{X_n\}$ is equal to $\rho^n$. This follows from the fact that the maximal correlation between a bivariate normal distribution is their correlation. In other words, if $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, then $\rho(\sigma(X), \sigma(Y)) = \rho$ (Lancaster (1957)). In the AR(1) process, the correlation of $X_0$ and $X_n$ is $\rho^n$ and hence $\rho(\sigma(X_0), \sigma(X_n)) = \rho^n$. Therefore $\rho(n) = \rho^n$.

Provided with the fact $\alpha(n) \leq \rho(n)$, we know $\alpha(n) \leq \rho^n$ for $\{\Phi(X_n)\}$. Plugging

this upper bound, $\sum_{n=1}^{\infty} n^2 \alpha_n^{\frac{1}{4}} \leq \sum_{n=1}^{\infty} n^2 \rho^{\frac{n}{4}} < \infty$ (for $-1 < \rho < 1$).

# Proof of the $8^{-1}$ bound for the normal model

**Proof**: Let $g(r, c)$ be the function defined by the right-hand side of (2.10). Then when $r \leq c \leq r^{-1}$,

$$g(r,c) = \frac{r(c-r)(1-rc)}{1+c^2-2rc} \leq \frac{1}{8} \quad \Leftrightarrow \quad (1+8r^2)c^2 - 2r(4r^2+5)c + 1 + 8r^2 \geq 0 \tag{A.1}$$

But for the quadratic form (in $c$) on the right-hand side, the discriminant $\Delta = 4(2r+1)^2(2r-1)^2(r^2-1) \leq 0$. This establishes our claim that when $c$ is in the safe zone $[r, r^{-1}]$, $r_{1\&2} \leq 8^{-1}$. (As mentioned before, a geometric proof is to use (2.12). It can also be viewed as a special case of the $t$ model with infinite degrees of freedom, discussed below.)

# Proof of the bound (2.17)

To prove this bound we need the notion of *partial maximal correlation* (Yu and Meng (2011)) defined for three random variables $X, Y, Z$ as the following:

$$\mathcal{R}_Z(X,Y) = \sup_{f,h \in L^2} \frac{\mathrm{Cov}(f(X) - \mathrm{E}[f(X)|Z], h(Y) - \mathrm{E}[h(Y)|Z])}{\sqrt{\mathrm{V}(f(X) - \mathrm{E}[f(X)|Z])\mathrm{V}(h(Y) - \mathrm{E}[h(Y)|Z])}}. \tag{A.2}$$

We also need the notion of *conditional maximal correlation*, which is defined as

$$\begin{aligned}\mathcal{R}(X,Y|Z) &= \sup_{f,h\in L^2} \mathrm{Corr}(f(X),h(Y)|Z) &\text{(A.3)}\\ &= \sup_{f,h\in L^2} \frac{\mathrm{Cov}(f(X),h(Y)|Z)}{\sqrt{\mathrm{V}(f(X)|Z)\mathrm{V}(h(Y)|Z)}}. &\text{(A.4)}\end{aligned}$$

The difference is that $\mathcal{R}(X,Y|Z)$ plays the role of conditional correlation, which is a function of $Z$; while $\mathcal{R}_Z(X,Y)$ plays the role of partial correlation, which is averaged over $Z$. But they obey the following inequality:

$$\mathcal{R}_Z(X,Y) \leq \sup_z \mathcal{R}(X,Y|Z=z). \tag{A.5}$$

This can be proved by first noticing that for any triple $\{X,Y,Z\}$,

$$\mathrm{Cov}(f(X)-\mathrm{E}[f(X)|Z], h(Y)-\mathrm{E}[h(Y)|Z]) = \mathrm{E}\Big[\mathrm{Cov}(f(X),h(Y)|Z)\Big] \tag{A.6}$$

whenever the needed moments exist. Applying (A.6) to both numerator and denominator of (A.2) (for the two parts in the denominator, we take $f=h$ in (A.6)), we obtain

$$\begin{aligned}\mathcal{R}_Z(X,Y) &= \sup_{f,h\in L^2} \frac{\mathrm{E}\Big[\mathrm{Cov}(f(X),h(Y)|Z)\Big]}{\sqrt{\mathrm{E}\Big[\mathrm{V}(f(X)|Z)\Big]\mathrm{E}\Big[\mathrm{V}(h(Y)|Z)\Big]}}, &\text{(A.7)}\\ &\leq \sup_z \mathcal{R}(X,Y|Z=z) \times \sup_{f,h\in L^2} \frac{\mathrm{E}\Big[\sqrt{\mathrm{V}(f(X)|Z)\mathrm{V}(h(Y)|Z)}\Big]}{\sqrt{\mathrm{E}[\mathrm{V}(f(X)|Z)]\mathrm{E}[\mathrm{V}(h(Y)|Z)]}} &\text{(A.8)}\\ &= \sup_z \mathcal{R}(X,Y|Z=z), &\text{(A.9)}\end{aligned}$$

where the last equality follows from the Cauchy-Schwartz inequality, which becomes equality when $f = h$.

Now applying (A.5) to (2.13), we have

$$\mathcal{R}_W(\theta, Y_{mis}) \leq \sup_w \mathcal{R}(\theta, Y_{mis}|W = w) = r. \tag{A.10}$$

By the Lemma 1 of Yu and Meng (2011),

$$\mathcal{R}(\theta, Y_{mis}) \leq \mathcal{R}_W(\theta, Y_{mis}) + (1 - \mathcal{R}_W(\theta, Y_{mis}))\mathcal{R}(\theta, W)\mathcal{R}(Y_{mis}, W). \tag{A.11}$$

Noting that under (2.13), $\mathcal{R}(Y_{mis}, W) = \mathcal{R}(\theta, W)$, we see from (A.10)-(A.11) that

$$\mathcal{R}(\theta, Y_{mis}) \leq g + (1 - g)r, \quad \text{where} \quad g = \mathcal{R}^2(\theta, W). \tag{A.12}$$

Letting $\tilde{Y}_{mis} = Y_{mis} - c\theta$, it is easy to see that the above derivation also applies to $\mathcal{R}(\tilde{Y}_{mis}, \theta)$ and $\mathcal{R}(\tilde{Y}_{mis}, Y_{mis})$, except with $r$ replaced respectively by

$$r_1 \equiv \mathcal{R}(\tilde{Y}_{mis}, \theta|W) = \frac{|c - r|}{\sqrt{1 + c^2 - 2cr}} \quad \text{and} \quad r_2 \equiv \mathcal{R}(\tilde{Y}_{mis}, Y_{mis}|W) = \frac{|1 - cr|}{\sqrt{1 + c^2 - 2cr}}, \tag{A.13}$$

where the calculation of $\mathcal{R}(\tilde{Y}_{mis}, \theta|W)$, for example, can be directly read off from the covariance matrix in (2.9) (the missing multiplicative factor $W^{-2}$ is not relevant for the correlation calculation). Consequently, from (2.3), we have

$$\begin{aligned} r_{1\&2} &\leq \mathcal{R}(\theta, Y^{mis})\mathcal{R}(\theta, \widetilde{Y}^{mis})\mathcal{R}(\widetilde{Y}^{mis}, Y^{mis}) && \text{(A.14)} \\ &\leq [g + (1 - g)r][g + (1 - g)r_1][g + (1 - g)r_2] \equiv F(r, c, g), && \text{(A.15)} \end{aligned}$$

Now we prove that in the "safe" zone, where $r \le c \le r^{-1}$,

$$F(r,c,g) \le \frac{1}{8}(1+g)^3.$$

We prove this in two steps.

1. For fixed $0 < r, g < 1$, $F(r,c,g)$ is maximized at $c = 1$. Because

$$\frac{\partial F}{\partial c} = \frac{(1-r^2)(1-g)[r+(1-r)g]}{(1+c^2-2cr)^2}[g\sqrt{1+c^2-2cr}+(1-g)(1+c)](1-c), \tag{A.16}$$

we have

$$\frac{\partial F}{\partial c}\begin{cases} > 0, \text{ if } c < 1 \\ = 0, \text{ if } c = 1 \\ < 0, \text{ if } c < 1. \end{cases}$$

Therefore, we see $F(r,c,g)$ is maximized at $c = 1$ for any $r$ and $g$.

2. For $c = 1$ and fixed $g$, $F(r,1,g)$ is maximized at $r = \frac{1}{2}$. Because

$$\frac{\partial F(r,1,g)}{\partial r} = \frac{(1-g)[(1-g)\sqrt{(1-r)/2}+g]\left[1-g+\frac{g}{\sqrt{2(1-r)}+1}\right]}{\sqrt{2(1-r)}}(1-2r), \tag{A.17}$$

we see that

$$\frac{\partial F(r,1,g)}{\partial r}\begin{cases} > 0, \text{ if } r < \frac{1}{2} \\ = 0, \text{ if } r = \frac{1}{2} \\ < 0, \text{ if } r < \frac{1}{2}. \end{cases}$$

Hence $F(r,1,g)$ is maximized when $r=\frac{1}{2}$ for any fixed $g<1$.

As a result, $F(r,c,g)\leq F(\frac{1}{2},1,g)=\frac{1}{8}(1+g)^3$.

# Proof of the limits of $\mathcal{R}_v(\theta,W)$

Now we prove that when $W^2\sim\chi^2_v/v$, $\mathcal{R}_v(\theta,W)\to 1$ as $v\to 0$, under the model (2.13). Consider two functions $g,h$ : $g(\theta)=|\theta|^{v/4}$ and $h(W)=W^{-v/4}$. Because $g(\theta)$ and $h(W)$ both have finite variances, their linear correlation is a lower bound for $\mathcal{R}_v(\theta,W)$. Thus it is sufficient to show that this linear correlation goes to one as $v\to 0$. Direct calculation shows

$$\text{Corr}(|\theta|^{v/4},W^{-v/4}) \;=\; 2^{\frac{v}{4}}\Gamma[\frac{4+v}{8}]\sqrt{\frac{\Gamma[\frac{v}{4}]\Gamma[\frac{v}{2}]-\Gamma[\frac{3v}{8}]^2}{2\pi\Gamma[\frac{v}{2}]^2-2^{\frac{v}{2}}\Gamma[\frac{3v}{8}]^2\Gamma[\frac{4+v}{8}]^2}}. \qquad \text{(A.18)}$$

By the fact that $\Gamma[v]\Gamma[1-v]=\dfrac{\pi}{\sin(\pi v)}$ when $0<v<1$, we know $\lim_{v\to 0}v\Gamma[v]=1$. Together with the fact that $\lim_{v\to 0}\Gamma[\frac{4+v}{8}]=\Gamma[\frac{1}{2}]=\sqrt{\pi}$, we deduce the right-hand side of (A.18) converges to

$$\sqrt{\pi}\sqrt{\frac{8-\dfrac{64}{9}}{2\pi\times 4-\dfrac{64}{9}\pi}}=1,$$

which completes our proof.

The proof for $\mathcal{R}_v(\theta,W)\to 0$ as $v\to\infty$ turns out to be much more involved, even though the result seems obvious because as $v\to\infty$, $W$ converges almost surely to the constant 1, and hence it should be independent of any random variable. The

trouble is that there is no theory to automatically guarantee that $\mathcal{R}_v(\theta, W)$ is a continuous function of $v$. In general, it is a rather complex task to establish even such a continuity with respect to a simple linear combination weight because in general it is not true (see Bryc and Dembo (2005)). We therefore take an indirect route, by considering a two-step Gibbs sampler alternating between sampling $\theta|W$ and $W|\theta$, whose $L^2$ convergence rate is $\mathcal{R}_v^2(\theta, W)$. It was shown in Roberts and Tweedie (2001) that, for a time reversible Markov chain (such as a two-step Gibbs sampler), its $L^2$ geometric rate is equivalent to its $L^1$ rate. By definition, geometric ergodicity in $L^1$ means that the total variation distance to the target distribution can be bounded by an exponentially decaying function. The bounds in Jones and Hobert (2004) yield precisely such functions, from which we can read off bounds on the geometric rate. Therefore, we can establish the desired result by proving that the $L^1$ rate converges to zero as $v \to \infty$.

To prove this, we first consider an equivalent two-step Gibbs sampler that alternates between $\theta|Y$ and $Y|\theta$, where $Y = vW^2 \sim \chi_v^2$. Clearly, to draw from $\theta|Y$, we only need to draw $Z \sim N(0,1)$ independently of $Y$, and then form $\theta = Z/\sqrt{Y/v}$. To draw $Y|\theta$, we note the identity $Y = (Y + Z^2)/(\theta^2/\nu + 1)$, and the fact that $1/(\theta^2/\nu + 1) = Y/(Y + Z^2)$ has a beta distribution and is independent of $Y + Z^2$, which has a $\chi_{\nu+1}^2$ distribution. Hence we simply draw $G \sim \chi_{\nu+1}^2$ independently of $\theta$ and let $Y = G/(\theta^2/\nu + 1)$. Combining the two steps we may represent one iteration of the $Y$ margin by

$$Y \to Y^{new} \equiv \frac{G}{1 + Z^2/Y}, \tag{A.19}$$

where $Z^2 \sim \chi_1^2$, $G \sim \chi_{\nu+1}^2$, and $Y$, $Z^2$, $G$ are independent. The Markov chain

(A.19) is irreducible (with respect to Lebesgue measure), aperiodic and positive Harris recurrent with $\chi^2_\nu$ as its invariant distribution. Therefore, to bound its $L^1$ rate we can establish suitable minorization and drift conditions and appeal to Rosenthal (1995)'s result as stated by Jones and Hobert (2004), Theorem 3.1.

Assume $\nu > 6$ and define

$$V(y) = \nu \left( \frac{\nu - 6}{y} - 1 \right)^2, \quad y > 0.$$

Direct calculation using moments of the inverse $\chi^2$ distribution yields

$$\mathrm{E}[V(Y^{new})|Y] = \gamma V(Y) + b, \quad \text{where } \gamma = \frac{3}{(\nu-1)(\nu-3)} \text{ and } b = \frac{2\nu^2}{(\nu-1)(\nu-3)}.$$

Let $d_R > 4$ be a constant, and suppose $\nu$ is large enough so that $\nu > d_R > 2b/(1-\gamma)$. Define the set $C = \{y > 0 : \ V(y) \leq d_R\}$, which is simply the interval

$$y \in [y_*, y^*], \quad \text{where } y_* = \frac{\nu - 6}{1 + \sqrt{d_R/\nu}} \text{ and } \ y^* = \frac{\nu - 6}{1 - \sqrt{d_R/\nu}}.$$

Let $\epsilon = \sqrt{y_*/y^*}$. For any fixed $y \in C$ the density of $Z^2/y$ is bounded below by $\epsilon$ times the density of $Z^2/y^*$, because

$$\sqrt{\frac{y}{2\pi x}} e^{-yx/2} \geq \sqrt{\frac{y_*}{2\pi x}} e^{-y^* x/2}, \quad x > 0.$$

It follows that, if we denote the distribution of $G/(1 + Z^2/y)$ by $P(y, \cdot)$ (i.e., $P(y, \cdot)$

is the transition kernel of (A.19)), then

$$P(y, \cdot) \geq \epsilon P(y^*, \cdot), \quad y \in C.$$

Specifically, one can sample from $P(y, \cdot)$ by setting $Y^{new} = G/(1 + Z^2/y^*)$ with probability $\epsilon$ and using another transition rule with probability $1 - \epsilon$.

We have now verified all conditions of Theorem 3.1 of Jones and Hobert (2004) and can conclude that the $L^1$ rate of (A.19) is bounded above by $\max\{(1-\epsilon)^r,\ U^r/\alpha^{1-r}\}$, where

$$\alpha = \frac{1 + d_R}{1 + 2b + \gamma d_R}, \quad U = 1 + 2(\gamma d_R + b)$$

and $r \in (0, 1)$ is an arbitrary constant. However, for fixed $d_R$, as $\nu \to \infty$ we have $b \to 2,\ \gamma \to 0,\ \epsilon \to 1$, and this upper bound tends to $5^r/((1 + d_R)/5)^{1-r}$. By choosing an arbitrarily large $d_R$ we can make this limiting upper bound arbitrarily small. Hence the $L^1$ rate must tend to zero as $\nu \to \infty$.

# Derivation of the asymptotic variance of $\hat{\rho}_{MoM}$ and $\hat{\rho}_{MLE}$

The variance of $\hat{\rho}_{MoM}$ is:

$$
\begin{aligned}
& \mathrm{Var}(\frac{(1-r^2)}{n}\sum_{k=1}^{n} Y_{k,1}Y_{k,2}) \\
= \; & \frac{(1-r^2)^2}{n^2}\{n\mathrm{Var}(Y_{1,1}Y_{1,2}) + 2\sum_{1\le i<j\le n}\mathrm{Cov}(Y_{i,1}Y_{i,2}, Y_{j,1}Y_{j,2})\} \\
= \; & \frac{(1-r^2)^2}{n^2}\{\frac{n(1+\rho^2)}{(1-r^2)^2} + 2(1+\rho^2)\sum_{1\le i<j\le n}\frac{r^{2(j-i)}}{(1-r^2)^2}\} \\
= \; & (n + 2\sum_{1\le i<j\le n} r^{2(j-i)})\frac{1+\rho^2}{n^2}
\end{aligned}
$$

The asymptotic variance of $\hat{\rho}_{MLE}$ is:

$$
\begin{aligned}
l'(\rho) = & \frac{1}{1-\rho^2}\{n\rho + (1-r^2)Y_{1,1}Y_{1,2} + \sum_{k=2}^{n}(Y_{k,1} - rY_{k-1,1})(Y_{k,2} - rY_{k-1,2})\} \\
& - \frac{\rho}{(1-\rho^2)^2}\{(1-r^2)(Y_{1,1}^2 + Y_{1,2}^2 - 2\rho Y_{1,1}Y_{1,2}) \\
& + \sum_{k=2}^{n}[(Y_{k,1} - rY_{k-1,1})^2 + (Y_{k,2} - rY_{k-1,2})^2 - 2\rho(Y_{k,1} - rY_{k-1,1})(Y_{k,2} - rY_{k-1,2})]\} \\
l''(\rho) = & \frac{2\rho}{(1-\rho^2)^2}\{n\rho + (1-r^2)Y_{1,1}Y_{1,2} + \sum_{k=2}^{n}(Y_{k,1} - rY_{k-1,1})(Y_{k,2} - rY_{k-1,2})\} \\
& + \frac{n}{1-\rho^2} - \frac{1+3\rho^2}{(1-\rho^2)^3}\{(1-r^2)(Y_{1,1}^2 + Y_{1,2}^2 - 2\rho Y_{1,1}Y_{1,2}) \\
& + \sum_{k=2}^{n}[(Y_{k,1} - rY_{k-1,1})^2 + (Y_{k,2} - rY_{k-1,2})^2 - 2\rho(Y_{k,1} - rY_{k-1,1})(Y_{k,2} - rY_{k-1,2})]\} \\
& + \frac{2\rho}{(1-\rho^2)^2}\{Y_{1,1}Y_{1,2}(1-r^2) + \sum_{k=2}^{n}(Y_{k,1} - rY_{k-1,1})(Y_{k,2} - rY_{k-1,2})\}
\end{aligned}
$$

The expected Fisher information is:

$$
E[-l''(\rho)] = n\frac{1+\rho^2}{(1-\rho^2)^2},
$$

and therefore

$$\mathrm{Var}(\hat{\rho}_{MLE}) \approx \frac{(1-\rho^2)^2}{n(1+\rho^2)}.$$

# Bibliography


Aguirre, G. K., Zarahn, E., et al. (2005), "A critique of the use of the Kolmogorov-Smirnov (KS) statistic for the analysis of BOLD fMRI data," *Magnetic Resonance in Medicine*, 39, 500–505.

Albert, J. H. (1992), "Bayesian estimation of normal ogive item response curves using Gibbs sampling," *Journal of Educational and Behavioral Statistics*, 17, 251–269.

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American statistical Association*, 88, 669–679.

Bassett, D. S. and Bullmore, E. T. (2009), "Human brain networks in health and disease," *Current opinion in neurology*, 22, 340.

Bayley, G. and Hammersley, J. (1946), "The "effective" number of independent observations in an autocorrelated time series," *Supplement to the Journal of the Royal Statistical Society*, 8, 184–197.

Billingsley, P. (2009), *Convergence of probability measures*, vol. 493, Wiley-Interscience.

Bradley, R. C. (2005), "Basic properties of strong mixing conditions. A survey and some open questions," *Probability surveys*, 2, 37.

Breiman, L. and Friedman, J. H. (1985), "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, 80, 580–598.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2010), *Handbook of Markov Chain Monte Carlo: Methods and Applications*, Chapman & Hall.

Brown, P., Kenward, M., and Bassett, E. (2001), "Bayesian discrimination with longitudinal data," *Biostatistics*, 2, 417–432.

Bryc, W. and Dembo, A. (2005), "On the maximum correlation coefficient," *Theory of Probability & Its Applications*, 49, 132–138.

Bullmore, E. T. and Bassett, D. S. (2011), "Brain graphs: graphical models of the human brain connectome," *Annual review of clinical psychology*, 7, 113–140.

Churchill, N. W., Oder, A., Abdi, H., Tam, F., Lee, W., Thomas, C., Ween, J. E., Graham, S. J., and Strother, S. C. (2012), "Optimizing preprocessing and analysis pipelines for single-subject fMRI. I. Standard temporal motion and physiological noise correction methods," *Human brain mapping*, 33, 609–627.

Craddock, R. C., Holtzheimer III, P. E., Hu, X. P., and Mayberg, H. S. (2009), "Disease state prediction from resting state functional connectivity," *Magnetic resonance in Medicine*, 62, 1619–1628.

De Luca, M., Beckmann, C., De Stefano, N., Matthews, P., Smith, S. M., et al. (2006), "fMRI resting state networks define distinct modes of long-distance interactions in the human brain," *Neuroimage*, 29, 1359–1367.

Deo, C. M. (1973), "A note on empirical processes of strong-mixing sequences," *The Annals of Probability*, 1, 870–875.

Doob, J. L. (1949), "Heuristic approach to the Kolmogorov-Smirnov theorems," *The Annals of Mathematical Statistics*, 20, 393–403.

Dutilleul, P. (1999), "The MLE algorithm for the matrix normal distribution," *Journal of statistical computation and simulation*, 64, 105–123.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., and Zilles, K. (2005), "A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data," *Neuroimage*, 25, 1325–1335.

Fox, M. D. and Raichle, M. E. (2007), "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging," *Nature Reviews Neuroscience*, 8, 700–711.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient parametrisations for normal linear mixed models," *Biometrika*, 82, 479–488.

— (1996), "Efficient parametrizations for generalized linear mixed models," *Bayesian statistics*, 5, 48–74.

Gelfand, A. E. and Smith, A. F. (1990), "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, 85, 398–409.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 721–741.

Hobert, J. (2001a), "Discussion of paper by van Dyk and Meng," *Journal of Computational and Graphical Statistics*, 10, 59–68.

Hobert, J. and Roman, J. (2011), "Discussion of paper by Yu and Meng," *Journal of Computational and Graphical Statistics*, 20, 571–580.

Hobert, J. P. (2001b), "Stability relationships among the Gibbs sampler and its subchains," *Journal of Computational and Graphical Statistics*, 10, 185–205.

Huettel, S. A., Song, A. W., and McCarthy, G. (2009), "Functional Magnetic Resonance Imaging, Sunderland, MA: Sinaur Associates," *Inc., January*, 16.

Jones, G. L. and Hobert, J. P. (2004), "Sufficient burn-in for Gibbs samplers for a hierarchical random effects model," *The Annals of Statistics*, 32, 784–817.

Kolmogoroff, A. (1933), "Sulla determinatione empirica di una legge di distributione," *Giornale dell'Istituto Italiano degli Attuari*, 4, 83–91.

Lancaster, H. O. (1957), "Some properties of the bivariate normal distribution considered in the form of a contingency table," *Biometrika*, 289–292.

Lange, K. and Sinsheimer, J. S. (1993), "Normal/independent distributions and their applications in robust regression," *Journal of Computational and Graphical Statistics*, 2, 175–198.

Liu, J. S. and Wu, Y. N. (1999), "Parameter expansion for data augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

Lu, N. and Zimmerman, D. (2004), "On likelihood-based inference for a separable covariance matrix," *Statistics and Actuarial Science Dept., Univ. of Iowa, Iowa City, IA, Tech. Rep*, 337.

Mallows, C. (1967), "Linear processes are nearly Gaussian," *Journal of Applied Probability*, 313–329.

Mardia, K. V. and Goodall, C. R. (1993), "Spatial-temporal analysis of multivariate environmental monitoring data," *Multivariate Environmental Statistics*, 6, 76.

Meng, X.-L. and Schilling, S. (1996), "Fitting full-information item factor models and an empirical investigation of bridge sampling," *Journal of the American Statistical Association*, 91, 1254–1267.

Meng, X.-L. and Van Dyk, D. (1998), "Fast EM-type implementations for mixed effects models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 559–578.

— (2002), "The EM Algorithman Old Folk-song Sung to a Fast New Tune," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59, 511–567.

Meng, X.-L. and Van Dyk, D. A. (1999), "Seeking efficient data augmentation schemes via conditional and marginal augmentation," *Biometrika*, 86, 301–320.

Meng, X.-L. and Xie, X. (2013), "I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb?" *Econometric Reviews*, to appear.

NITRC (2013), `http://www.nitrc.org/`, accessed: 11/03/2013.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003), "Non-centered parameterisations for hierarchical models and data augmentation (with discussion)," in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, eds. Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., Oxford Univ. Press, New York, pp. 307–326.

— (2007), "A general framework for the parametrization of hierarchical models," *Statistical Science*, 22, 59–73.

Peña, D., Tiao, G. C., and Tsay, R. S. (2001), *A course in time series analysis*, J. Wiley.

Raichle, M. E. and Mintun, M. A. (2006), "Brain work and brain imaging," *Annu. Rev. Neurosci.*, 29, 449–476.

Roberts, G. O. and Tweedie, R. L. (2001), "Geometric L2 and L1 convergence are equivalent for reversible Markov chains," *Journal of Applied Probability*, 37–41.

Rosenthal, J. S. (1995), "Minorization conditions and convergence rates for Markov chain Monte Carlo," *Journal of the American Statistical Association*, 90, 558–566.

— (2011), "Optimal proposal distributions and adpative MCMC," in *Handbook of Markov Chain Monte Carlo*, eds. Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., Chapman & Hall/CRC, Boca Raton, pp. 93–110.

Salvador, R., Suckling, J., Coleman, M. R., Pickard, J. D., Menon, D., and Bullmore, E. (2005), "Neurophysiological architecture of functional magnetic resonance images of human brain," *Cerebral Cortex*, 15, 1332–1342.

Smirnov, N. (1939), "Ob uklonenijah empiriceskoi krivoi raspredelenija," *Recueil Mathematique*, 6, 3–26.

Tanabe, J., Miller, D., Tregellas, J., Freedman, R., and Meyer, F. G. (2002), "Comparison of detrending methods for optimal fMRI preprocessing," *NeuroImage*, 15, 902–907.

Tanner, M. A. and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American statistical Association*, 82, 528–540.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., et al. (2002), "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage*, 15, 273–289.

van den Heuvel, M. P. and Hulshoff Pol, H. E. (2010), "Exploring the brain network: a review on resting-state fMRI functional connectivity," *European Neuropsychopharmacology*, 20, 519–534.

Van Dyk, D. A. and Meng, X.-L. (2001), "The art of data augmentation," *Journal of Computational and Graphical Statistics*, 10, 1–50.

— (2010), "Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: A graphical guide book," *Statistical Science*, 25, 429–449.

Weiss, M. S. (1978), "Modification of the Kolmogorov-Smirnov statistic for use with correlated data," *Journal of the American Statistical Association*, 73, 872–875.

Yu, Y. and Meng, X.-L. (2011), "To Center or Not to Center: That Is Not the QuestionAn Ancillarity–Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Efficiency," *Journal of Computational and Graphical Statistics*, 20, 531–570.

Yue, S. and Wang, C. (2004), "The Mann-Kendall test modified by effective sample size to detect trend in serially correlated hydrological series," *Water Resources Management*, 18, 201–218.