INFERENCE FOR CORRELATION MATRICES FOR LONGITUDINAL AND ORDERED DATA

By

YANPIN WANG

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2012
To my parents, my son, and my husband
ACKNOWLEDGMENTS

I place my deepest appreciation to my advisor, Professor Micheal Daniels. Without his guidance and persistent help, I will never be able to reach this far.

I am thankful to my doctoral committee members, Prof. Malay Ghosh, Prof. Xiaomin Lu, and Prof. Ronald Shorr, who have provided plentiful support and valuable insight during my dissertation. I owe a debt of gratitude to all of my teachers at the Department of Statistics for providing an outstanding education. I would like to give my special thanks to Tina Greenly, Summer Layton, Robyn Crawford, and Marilyn Saddler for their extra help in every possible way.

My special thanks will go to my husband, Jiangto Luo, who encourages me all time, and my son, Binjie Luo, who cheers the every step I made. I will place my deepest gratitude to my parents, who give me unconditionally and constantly love. I deeply oblige the sacrifices my parents had to make to provide me the opportunity to have better education.

Last but not least, I would like to thank all of my friends at the University of Florida. I will never forget their humor, friendship, and unwavering support.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>9</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>12</td>
</tr>
<tr>
<td>2 BAYESIAN MODELING OF THE DEPENDENCE IN LONGITUDINAL DATA</td>
<td>27</td>
</tr>
<tr>
<td>via Partial Autocorrelations and Marginal Variances</td>
<td></td>
</tr>
<tr>
<td>2.1 Background</td>
<td>27</td>
</tr>
<tr>
<td>2.1.1 Brief literature Review for Estimating a Correlation Matrix</td>
<td>27</td>
</tr>
<tr>
<td>2.1.2 Review of Partial Autocorrelations and Modeling Correlation Matrices</td>
<td>28</td>
</tr>
<tr>
<td>2.1.3 Outline of this Chapter</td>
<td>29</td>
</tr>
<tr>
<td>2.2 Models for the Covariance Matrix</td>
<td>30</td>
</tr>
<tr>
<td>2.2.1 Partial Autocorrelations</td>
<td>30</td>
</tr>
<tr>
<td>2.2.2 Marginal Variances</td>
<td>31</td>
</tr>
<tr>
<td>2.3 Priors for $\gamma$</td>
<td>31</td>
</tr>
<tr>
<td>2.3.1 Review of Priors for Unstructured Partial Autocorrelations</td>
<td>32</td>
</tr>
<tr>
<td>2.3.2 An alternative Prior for Unstructured Partial Autocorrelations</td>
<td>32</td>
</tr>
<tr>
<td>2.3.3 Proposed Prior on $\gamma$</td>
<td>33</td>
</tr>
<tr>
<td>2.3.4 Extension to Unit-level Covariates</td>
<td>36</td>
</tr>
<tr>
<td>2.3.5 Connection to g-priors</td>
<td>36</td>
</tr>
<tr>
<td>2.3.6 Prior for $\eta$</td>
<td>37</td>
</tr>
<tr>
<td>2.3.7 Posterior Propriety</td>
<td>38</td>
</tr>
<tr>
<td>2.4 Simulations</td>
<td>39</td>
</tr>
<tr>
<td>2.4.1 Models</td>
<td>39</td>
</tr>
<tr>
<td>2.4.2 Results</td>
<td>40</td>
</tr>
<tr>
<td>2.5 Data Example: Schizophrenia Trial</td>
<td>41</td>
</tr>
<tr>
<td>2.6 Discussions</td>
<td>43</td>
</tr>
<tr>
<td>3 ESTIMATING LARGE CORRELATION MATRICES BY BANDING THE PARTIAL</td>
<td>52</td>
</tr>
<tr>
<td>AUTOCORRELATION MATRIX</td>
<td></td>
</tr>
<tr>
<td>3.1 Background</td>
<td>52</td>
</tr>
<tr>
<td>3.2 Review of Partial Autocorrelations</td>
<td>53</td>
</tr>
<tr>
<td>3.3 Banding the Partial Autocorrelation Matrix</td>
<td>54</td>
</tr>
<tr>
<td>3.3.1 Estimating the Partial Autocorrelations for each Band</td>
<td>55</td>
</tr>
<tr>
<td>3.3.1.1 Statement of needed results</td>
<td>56</td>
</tr>
<tr>
<td>3.3.1.2 Results for computing the mle for each band</td>
<td>57</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Posterior means of $\beta$</td>
<td>45</td>
</tr>
<tr>
<td>2-2</td>
<td>Summary measures for the simulation</td>
<td>49</td>
</tr>
<tr>
<td>2-3</td>
<td>Summary count from the simulations</td>
<td>50</td>
</tr>
<tr>
<td>2-4</td>
<td>Descriptive summaries of schizophrenia trial</td>
<td>51</td>
</tr>
<tr>
<td>2-5</td>
<td>Posterior summaries of schizophrenia</td>
<td>51</td>
</tr>
<tr>
<td>3-1</td>
<td>Risk simulations based on metal and rock data</td>
<td>65</td>
</tr>
<tr>
<td>3-2</td>
<td>Risk simulation based on AR(1) correlation matrices</td>
<td>68</td>
</tr>
<tr>
<td>3-3</td>
<td>Risk simulations of 4 band partial autocorrelation matrices</td>
<td>69</td>
</tr>
<tr>
<td>3-4</td>
<td>Estimated number of bands for rock and metal of sonar data</td>
<td>69</td>
</tr>
<tr>
<td>4-1</td>
<td>Simulation records for AR(1) structure</td>
<td>83</td>
</tr>
<tr>
<td>4-2</td>
<td>Simulation records for model 2</td>
<td>84</td>
</tr>
<tr>
<td>4-3</td>
<td>Simulation records for model 3</td>
<td>84</td>
</tr>
<tr>
<td>4-4</td>
<td>Simulation records for model 4</td>
<td>86</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Marginal priors</td>
<td>46</td>
</tr>
<tr>
<td>2-2</td>
<td>Trajectories of sample size 30, 100, 400</td>
<td>47</td>
</tr>
<tr>
<td>2-3</td>
<td>Observed trajectories and fitted trajectories</td>
<td>48</td>
</tr>
<tr>
<td>3-1</td>
<td>Images of sample correlation and estimated correlation matrices</td>
<td>67</td>
</tr>
<tr>
<td>4-1</td>
<td>True partial autocorrelation curves for models</td>
<td>85</td>
</tr>
<tr>
<td>4-2</td>
<td>Sample partial autocorrelations and smoothed curves within each band of the metal data</td>
<td>87</td>
</tr>
<tr>
<td>4-3</td>
<td>Sample partial autocorrelations and smoothed curves within each band of the rock data</td>
<td>88</td>
</tr>
<tr>
<td>4-4</td>
<td>Intensity plots for the metal and rock data</td>
<td>89</td>
</tr>
</tbody>
</table>
Many parameters and positive-definiteness are two major obstacles in estimating and modelling a correlation matrix for longitudinal data. In addition, when longitudinal data is incomplete, incorrectly modelling the correlation matrix often results in bias in estimating mean regression parameters (Little and Rubin (2002) and Daniels and Hogan (2008)). Although the sample covariance matrix is an unbiased estimator of the covariance matrix of a Gaussian random vector, it has poor properties if the dimension (p) is large (Stein (1975)). Besides, covariance matrices are often sparse for large p. Recently, estimating large covariance matrices has seen an upsurge in practical and theoretical approaches due to a plethora of high dimensional data.

In my dissertation, we introduce a flexible class of regression models for a covariance matrix parameterized using marginal variances and partial autocorrelations. We propose a class of priors for the regression coefficients and examine the importance of correctly modeling the correlation structure on estimation of longitudinal (mean) trajectories via simulations. The regression approach is illustrated on data from a longitudinal clinical trial. In addition, we propose a computationally efficient approach to estimate (large) p-dimensional correlation matrices of ordered data based on an independent sample of size n. This approach is considerably faster than many existing methods and only requires inversion of $k (k << p)$ dimensional covariance matrices. The resulting estimator is guaranteed to be positive definite as long as $k \leq n - 2$ (even
when $n < p$). We evaluate our estimator via extensive simulations and compare it to the Ledoit-Wolf estimator. We also illustrate the approach using high-dimensional sonar data. Finally, we introduce a nonparametric estimator of a correlation matrix based on smoothing the partial autocorrelations within bands. An asymptotic optimal bandwidth is derived based on a Toeplitz condition, and a plug-in bandwidth selector is suggested. The improvement is demonstrated by simulations. The simulations suggest that the Toeplitz condition on partial autocorrelation matrices is not restrictive.
CHAPTER 1
INTRODUCTION

Estimation of covariance matrices is important for the analysis of multivariate data. Based on a population covariance matrix, one can estimate principal components and eigenvalues, construct linear discriminant functions and confidence intervals (confidence bounds) on linear functions, and establish dependencies and conditional dependencies. Longitudinal data, which is multivariate data ordered in time, possesses features of both general multivariate data and time series data. Time series analysis is concerned with statistical inference from data which are not necessarily independent and identically distributed. Time series analysis often assume stationarity which does not holds true. In longitudinal studies, we have repeated (across units) time series.

Modeling the dependence structure carefully is important in making inference from longitudinal data. Its importance is magnified in incomplete data since incorrectly modeling the dependence structure often results in a biased estimate of the mean parameters (Little and Rubin (2002); Daniels and Hogan (2008)). Moreover, covariance matrices are often sparse for large p and have special features under ordered (longitudinal) data, such as $Y_j$ and $Y_k$ being closer to independence or conditional independence as $|j - k|$ increases. Finding a way to estimate covariance matrices is crucial as p is large, especially when sample size $n < p$.

The main contribution of my dissertation includes to develop a framework for Bayesian modeling of the dependence structure for longitudinal data via partial autocorrelations and marginal variances, and estimating large covariance matrices using the partial autocorrelation parametrization. We will briefly review concepts and literatures in the followings.

1. **Definition of Partial Autocorrelations**

   We parameterize a correlation matrix $R = (\rho_{jk})$ of multivariate random vector $Y = (Y_1, Y_2, ..., Y_p)$ to a partial autocorrelation matrix. An element in the partial
autocorrelation matrix $\Pi = (\pi_{jk})$, where $\pi_{jj} = 1$ and for $1 \leq j < k \leq m$, $\pi_{jk}$ is the correlation between $Y_j$ and $Y_k$ adjusted for the intervening variables, $\{Y_{j+1}, \ldots, Y_{k-1}\}$.

Unlike $R$, the off-diagonal elements of $\Pi$ can independently vary in $(-1, 1)$ while the corresponding correlation matrix, $R$, remaining positive-definite.

There is a simple relationship between the elements of $R$ and $\Pi$. The partial autocorrelation $\pi_{j,j+1} = \rho_{j,j+1}$ and $\pi_{j,j+k}$, $(2 \leq k \leq m - 1)$ is

$$\pi_{j,j+k} = \frac{\rho_{j,j+k} - r_1(j,k)R_2^{-1}(j,k)r_3^T(j,k)}{[1 - r_1(j,k)R_2^{-1}(j,k)r_3^T(j,k)]^{1/2}[1 - r_3(j,k)R_2^{-1}(j,k)r_3^T(j,k)]^{1/2}}, \quad (1-1)$$

with $r_1(j,k) = (\rho_{j,j+1}, \ldots, \rho_{j,j+k-1})$, $r_3(j,k) = (\rho_{j+k,j+1}, \ldots, \rho_{j+k,j+k-1})$, and $R_2(j,k)$ contains the middle $k - 1$ rows and columns of $R[j:j+k]$, i.e.

$$R[j:j+k] = \begin{pmatrix} 1 & r_1(j,k) & \rho_{j,j+k} \\ r_1^T(j,k) & R_2(j,k) & r_3^T(j,k) \\ \rho_{j+k,j} & r_3(j,k) & 1 \end{pmatrix} \quad (1-2)$$

Correspondingly, $\rho_{j,j+k}$ can be written as a function of $\pi_{j,j+k}$,

$$\rho_{j,j+k} = r_{jk} + \pi_{j,j+k}A_{jk}, \quad (1-3)$$

where

$$r_{jk} = r_1(j,k)R_2^{-1}(j,k)r_3^T(j,k)$$

$$A_{jk}^2 = [1 - r_1(j,k)R_2^{-1}(j,k)r_3^T(j,k)][1 - r_3(j,k)R_2^{-1}(j,k)r_3^T(j,k)].$$

A partial autocorrelation matrix will often be sparse (Dempster (1972); Friedman et al. (2008)) since the correlations in it are conditional (on the intervening variables) correlations. For example, an AR(1) correlation matrix corresponds to a partial autocorrelation matrix with only one non-zero band. More generally a partial autocorrelation matrix will have $k$ non-zero bands under a $k$-th order ante-dependence model (Zimmerman and Nunez-Anton (2010); Gabriel (1962))
2. Modeling the Dependence Structure

2.1 Matrix Decomposition

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed random vectors of dimension $p$ following a multivariate normal distribution with mean $\mu$ and variance-covariance matrix $\Sigma$. We review several common decomposition $\Sigma$ that can be useful for parsimoniously modeling the dependence structure.

1. Spectral Decomposition (Eigenvalue decomposition)

\[ \Sigma = PP' \]  

(1–4)

where $\Lambda$ is the diagonal matrix of eigenvalues which specify its shape and the column vectors of $P$ are the corresponding eigenvectors. Models using this decomposition (Eq.1–4) have been developed in Leonard and Hsu (1992) and Chiu et al. (1996).

2. Modified Cholesky Decomposition of Covariance Matrices

\[ T\Sigma T' = B \]  

(1–5)

where $T$ is a unique unit lower triangular matrix with 1s as diagonal entries and unique diagonal matrix $B$ with positive diagonal entries. This decomposition provides a nice interpretation under a longitudinal setting. The below-diagonal elements of $T$ in Eq.1–5 are the negatives of the coefficients of $\hat{Y}_t = \mu_t + \sum_{j=1}^{t-1} \phi_{t,j}(Y_j - \mu_j)$, the linear least-square predictors of $Y_t$ based on its predecessors $Y_{t-1}, \ldots, Y_1$. And the diagonal entries of $B$ are the prediction error variances $\sigma^2_t = \text{var}(Y_t - \hat{Y}_t)$, for $1 \leq t \leq n$. The positive-definite condition of covariance matrices can be easily satisfied by assuming $\sigma^2_t > 0$ (Pourahmadi (1999); Daniels and Pourahmadi (2002)).

3. Variance/Correlation Decomposition

\[ \Sigma = DRD \]  

(1–6)
where \( D \) is diagonal matrix with standard deviations on its diagonal and \( R = (\rho_{jk}) \) is a correlation matrix.

4. Reparameterize \( R = (\rho_{jk})_{p \times p} \) in Eq. 1–6 to partial autocorrelation matrix \( \Pi = (\pi_{jk}) \).

The partial autocorrelation \( \pi_{jk} \) is defined as:

- the lag-1 partial autocorrelations
  \[ \pi_{j,j+1} = \rho_{j,j+1}, j = 1, \cdots, p - 1 \]

- and the partial autocorrelations
  \[ \pi_{jk} = \rho_{jk|j+1,\cdots,k-1}, k - j \geq 2. \] (1–7)

Furthermore, we can use Fisher z-transformation \( z = -\frac{1}{2} \log \frac{1+\pi}{1-\pi} \) transforming \( \pi \in (-1, 1) \) to \( z \in (-\infty, +\infty) \), which motivates the generalized linear model framework with link function \( g(\pi) = -\frac{1}{2} \log \frac{1+\pi}{1-\pi} \).

The lag \( k \) partial Autocorrelations are based on conditional regressions where the conditioning sets have the same number of elements. They are *exchangeable* in some sense (all conditioning on \( k - 1 \) intervening variables). The ante-dependence models introduced by Gabriel (1962) have close relation to this reparameterization. We review it next.

Observations \((Y_1, \ldots, Y_p)\) whose joint distribution is multivariate normal are \( s \)-th-order ante-dependence if \( Y_j \) and \( Y_{j+k+1} \) are independent given the intervening observations \( Y_{j+1}, \ldots, Y_{j+k}, \) for all \( j = 1, \ldots, p - k - 1 \) and all \( k \geq s \) (Gabriel (1962)), i.e.

\[ \text{Cov}(Y_j, Y_{j+k+1}|Y_{j+1}, \ldots, Y_{j+k}) = 0. \]

Hence, we can model these random variables as follows:

\[
\begin{cases}
  y_1 = x_1^T \beta + \epsilon_1 \\
  y_j = x_j^T \beta + \sum_{k=1}^{s^*} \phi_{jk}(y_{j-k} - x_{j-k}^T \beta) + \epsilon_j \quad (j = 2, \ldots, p).
\end{cases}
\] (1–8)
where \( s^* = \min(s, j - 1) \), the \( \epsilon_j \)s are independent normal random variables with zero mean and variance \( \sigma_j^2 \), and the \( \phi_{jk} \)s are unrestricted parameters. Zimmerman and Nunez-Anton (2010) compare several models, including (1) the unstructured covariance model, (2) unstructured ante-dependence models, (3) structured ante-dependence models, (4) autoregressive integrated moving average and similar models, and (5) random coefficients models.

### 2.2 Priors for Covariance Matrices

The most common prior for a covariance matrix is Inverse Wishart distribution, a conjugate prior for a normal covariance matrix. However, this prior lacks flexibility, allowing only one precision parameter for all \( p(p-1)/2 \) elements and requires specification of a "mean" matrix (Daniels and Kass (1999)). Yang and Berger (1994) develop a reference prior for a covariance matrix based on a spectral decomposition of the covariance matrix. Smith and Kohn (2002) provide a prior for large data that allows zero elements in the strict lower triangle of the Cholesky decomposition (Eq.1–5) of the inverse covariance matrix to obtain a parsimonious representation of the covariance matrix. Daniels and Pourahmadi (2002) propose shrinking elements of the T matrix in Eq.1–5 toward specific structure. Barnard et al. (2000) model the correlations given constraints in Eq.1–4 and propose a joint uniform prior and marginal independent uniform priors on conditional distribution of \((R|D)\). Liechty et al. (2004) place a prior on \( R \) with pdf \( f(R|\mu, \sigma^2) \sim \prod_{j<k} \exp\{-(\rho_{jk} - \mu)^2/(2\sigma^2)\}I\{R \in \mathbb{R}^p\} \), where \( \mathbb{R}^p \) is the sample space of \( R \), the indicator function \( I\{} \) ensures that the correlation matrix is positive definite and introduces dependence among the \( \rho_{jk} \)s. Based on Eq.1–7, Daniels and Pourahmadi (2009) propose an independent transformed \( Beta(-1,1)(a_k, a_k) \) on \( \pi_{ij+k} \) with \( a_k = \frac{p-k-1}{2} \) and \( k = 1, 2, ..., p - 1 \) which corresponds to joint uniform prior of \( \rho \) on \( p(p - 1)/2 \) dimensional hyper-cube.

### 3. Estimating Large Covariance Matrices
Estimating large covariance matrices (p is comparable or larger than the sample size n) has gained increased interest recently since high dimensional data is so common in current applications, such as climate data, financial data, functional data, gene expression data, and functional Magnetic Resonance Imaging (fMRI) data. Many approaches can be found in the recent literature. Typically, there are a lot of zeros in entries of the variance-covariance matrix (correlation or conditional correlation) as p grows. The simplest estimate of $\Sigma$ is the sample variance-covariance matrix $
abla \Sigma = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^T$. $\hat{\Sigma}$ is optimal and unbiased estimate of $\Sigma$ when $p < n$ and fixed (Dempster (1972)). And $\hat{\Sigma}$ is a consistent estimator of $\Sigma$ at rate of convergence $n^{-\frac{1}{2}}$. But estimation is poor as p increases, especially $p > n$ (Bickel and Levina (2008b); Bickel and Levina (2008a); Johnstone (2001); and Karoui (2009)). Developing methods in this situation are desired. Recent approaches include 1) Lasso regression, and 2) Banding. We review these next.

3.1 Lasso Regression Based on Likelihood Function

Lasso regression can be defined using a log likelihood function or a loss function. Suppose $(y_i, x_i)$, $i = 1, 2, ..., n$, where $x_i = (x_{i1}, ..., x_{iq})^T$ is the predictor of response variable $y_i$. For simplification, suppose $y_i$ is centered.

Let $Y_i = x_i^T \beta + \epsilon$, where $\beta = (\beta_1, ..., \beta_q)^T$ and $\epsilon \sim N(0, \sigma^2)$.

Definition 1: Based on log likelihood function

Suppose $\ell(\beta)$ is the log likelihood function of parameter $\beta_{q \times 1}$. The lasso estimate $\hat{\beta}$ is defined by

$$\hat{\beta} = \arg \min_{\beta} \ell(\beta)$$

subject to $\sum |\beta_i| \leq s$, where $s \geq 0$ (Tibshirani (1996)).

Definition 2: Based on Squared Loss
Under squared error loss, the lasso estimate \( \hat{\beta} \) is defined as

\[
\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - \sum_j \beta_j x_{ij})^2 \right\} \tag{1-10}
\]

subject to \( \sum_j |\beta_j| \leq s \), where \( s \geq 0 \) (Tibshirani (1996)).

Fu (1998) shows that Eq.1–10 is equivalent to

\[
\hat{\beta} = \arg\min \left\{ \sum_{i=1}^{n} (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j| \right\} \tag{1-11}
\]

which is a penalized regression, for some \( \lambda \).

The property of the \( L_1 \) norm allows Lasso regression to capture zero entries of variance-covariance or correlation matrix. Huang et al. (2006) apply the lasso penalty to the modified Cholesky coefficients. Rothman and Zhou (2008) use a nested lasso which replaces \( L_1 \) penalty by a nested Lasso penalty. Applying group lasso, Bigot et al. (2009) select a sparse set of basis functions in the dictionary used to approximate the process. Most of these approaches deal directly with a covariance matrix and do not adapt easily to a correlation matrix.

### 3.2 Banding on Variance-covariance or Correlation Matrix

To construct a spare estimator of a covariance matrix, Bickel and Levina (2008b) band the sample covariance matrix and the regression parameters of the Cholesky decomposition of covariance matrix, while Wu and Pourahmadi (2003) band the regression parameters after a Cholesky decomposition of inverse covariance matrix.

To determine the number of bands \( k \), Bickel and Levina (2008b) minimized the following risk,

\[
R_k = E||\hat{\Sigma}_k - \Sigma||_{(1,1)}
\]

and

\[
k_0 = \arg\min_k \{ R_k \}
\]
where \( \|M\|_{(1,1)} \equiv \sup\{\|Mx\|_1 : \|x\|_1 = 1\} = \max_j \sum_i |m_{ij}| \), for \( M = (m_{ij}) \).

They propose a resampling scheme to estimate the risk \( R_k \) and \( k_0 \) by randomly dividing the original sample into two data set \((\nu_1, \nu_2)\). \( \nu_1 \) is called the "target" data set whose sample size is chosen to be \( \lfloor \frac{n}{3} \rfloor \), then they estimate the risk by

\[
\hat{R}_k = \frac{1}{N} \sum_{\nu=1}^{N} \|B_k(\Sigma_1(\nu)) - \Sigma_2(\nu)\|_{(1,1)}
\]

where \( N \) is number of resampling, and \( B_k(M) \) is k-band matrix of \( M = (m_{ij}) \) defined as:

\[
B_k(M) = [m_{ij}I(|i-j| \leq k)]
\]

Then \( k \) is selected as

\[
\hat{k} = \arg\min_k \hat{R}(k).
\]

Wu and Pourahmadi (2003) minimize an AIC-type criterion defined as,

\[
AIC(d) = n \sum_{i=1}^{p} \log \tilde{\sigma}_t^2(d) + 2d,
\]

where \( d = 0, 1, ..., \lfloor p^{\frac{1}{2}} \rfloor \), and \( \tilde{\sigma}_t^2 \) is defined as

\[
\tilde{\sigma}_t^2(d) = \begin{cases} 
  n^{-1} \sum_{i=1}^{n} \{y_{t,i} - (\phi_{t-1,i}y_{t-1,i} + ... + \phi_{t-d,i}y_{t-d,i})\}^2 & d \geq 1 \\
  n^{-1} \sum_{i=1}^{n} (y_{t,i} - \bar{y}_t)^2 & d = 0
\end{cases}
\]

4. Models for Longitudinal Data

Models for longitudinal data can be classified into two groups, 1) directly specified (marginal) models; 2) indirectly specified (conditional) models. In a marginal model, the marginal expectation \( E(Y_{ij}) \) is directly modeled as a function of covariates \( x_{ij} \), e.g. \( E(Y_{ij}) = \mu_{ij} = \chi_{ij}^T \beta \) (Liang and Zeger (1986); Heagerty (1999); and Heagerty (2002)). In indirectly specified models, the effect of covariates on responses is modeled conditionally (condition on a random effects or previous history of responses) (Zeger and Karim (1991); Hedeker and Gibbons (1994)).
Here, we review two marginal models, one for continuous responses and one for binary responses. The first one captures dependence via a covariance matrix, and the second via a correlation matrix.

### 4.1 Marginal Models for Longitudinal Studies

1) **Multivariate normal model:**

Let \( Y_i = (Y_{i1}, \ldots, Y_{ip})^T \) be a response random vector and \( x_{it} \) be a \( q \times 1 \) vector of covariates for \( i = 1, \ldots, n \). We assume \( Y_i \sim N(\mu_i, \Sigma_i) \), where

\[
E(Y_i|x_i) = x_i^T \beta_i.
\]

We consider models for \( \Sigma_i \) in Chapter 2.

2) **Multivariate probit model:**

Let \( Y_i = (Y_{i1}, \ldots, Y_{ip})^T \) be a multivariate binary response vector, and \( Z_i = (Z_{i1}, \ldots, Z_{ip})^T \) be a latent variable with \( Z_i \sim N(x_i^T \beta, \Sigma_i) \). Suppose multivariate binary responses \( Y_i \) satisfy:

\[
Y_{ij}(Z_{ij}) = \begin{cases} 
0 & \text{if } Z_{ij} < 0 \\
1 & \text{if } Z_{ij} \geq 0
\end{cases}
\]

where \( j = 1, 2, \ldots, p \). So

\[
P(Y_i = y_i|\beta, \Sigma_i, x_i) = \int_{B_{y_1}} \cdots \int_{B_{y_p}} \phi_p(Z_i|x_i^T \beta, \Sigma_i) dZ_i
\]

where \( B_{y} \) is the interval \((0, +\infty)\), if \( y_{ij} = 1 \) and \((-\infty, 0)\), otherwise. And \( \phi_p(\bullet) \) is the pdf of a \( p \) dimension multivariate normal with mean \( x_i^T \beta \) and covariance matrix \( \Sigma_i \). For identifiability, we set the diagonal elements of \( \Sigma_i \) to be 1. (So \( \Sigma_i \) is a correlation matrix).

We will explore ways to put flexible structure on \( \Sigma_i \) via partial autocorrelation models and ways to construct parsimonious estimates of large covariance matrices using partial autocorrelations in Chapter 2, Chapter 3, and Chapter 4.

### 4.2 Missing Data Analysis
Missing data is very common in longitudinal studies. We review key concepts next.

4.2.1 Models for longitudinal missing data structure

Suppose the full data response vector is \( Y_i = (Y_{i1}, \ldots, Y_{ip})^T \), \( X_i \) is a \( p \times q \) matrix of covariates, and the vector \( R_i = (R_{i1}, \ldots, R_{ip})^T \) indicates which components are observed, with \( R_{ij} = 1 \) if \( Y_{ij} \) is observed, and \( R_{ij} = 0 \) if \( Y_{ij} \) is missing.

The response vector \( Y_i \) can be divided into two parts: 1) observed data (\( Y_{obs} \)) and 2) missing data (\( Y_{mis} \)), i.e. \( Y_i = (Y_{obs}, Y_{mis}) \).

We define the full data response model as

\[
  f(y|x, \theta).
\]

We define the full data model as

\[
  f(y, r|x, \omega).
\]

The relationship between full data response model and full data model is:

\[
  f(y|x, \theta(\omega)) = \sum_{r \in \mathcal{R}} f(y, r|x, \omega)
\]

where \( \mathcal{R} \) is the sample space of \( R \) (\( r \) is the realization of \( R \)).

There are several ways to factor the full data model. For example,

1) Extrapolation Factorization (Daniels and Hogan (2008));

\[
  f(y, r|x, \omega) = f(y_{obs}, y_{mis}, r|x, \omega) = f(y_{mis}|y_{obs}, r, x, \omega) \times f(y_{obs}, r|x, \omega).
\]

Clearly, there is no information in the data about the extrapolation distribution \( f(y_{mis}|y_{obs}, r, x, \omega) \).

2) Selection Model factorization:

\[
  f(y, r|x, \omega) = f(y|x, \theta(\omega)) \times f(r|y, x, \psi(\omega)).
\]
The first factor is the full data response model. We call the second factor \( f(r|y, x, \psi(\omega)) \) as the missing data mechanism (MDM), we will discuss different ways the MDM is classified next.

4.2.2 Missing data mechanism

Starting from Rubin (1976) and Little and Rubin (2002), missingness was classified into three categories:

1) Missing Completely At Random (MCAR)

\[
p(R|Y, x) = p(R|x).
\]

2) Missing At Random (MAR).

\[
p(R|Y, x) = p(R|Y_{\text{obs}}, x).
\]

3) Missing Not At Random (MNAR)

\[
p(R|Y_{\text{obs}}, Y_{\text{mis}}, x) \neq p(R|Y_{\text{obs}}, Y_{\text{mis}}', x).
\]

for \( Y_{\text{mis}} \neq Y_{\text{mis}}' \).

In a Bayesian setting, it is often more natural to classify missingness as ignorable or non-ignorable, we define ignorable next.

A missing data mechanism is called ignorable, if

1) The missing data mechanism is MAR,
2) The full data parameter \( \omega \) can be decomposed as \( \omega = (\theta, \psi) \),
   a) \( \theta \) indexes the full-data response model \( f(Y|\theta) \),
   b) \( \psi \) indexes the missing data mechanism \( f(R|Y, \psi) \).
3) The parameters \( \theta \) and \( \psi \) are a priori independent;
   i.e. \( f(\theta, \psi) = f(\theta)f(\psi) \).
Under ignorable missingness mechanism, the likelihood function can be factored over \( \theta \) and \( \psi \)

\[
L(\theta, \psi|y_{obs}, R, x) = L_1(\psi|R, y_{obs}, x)L_2(\theta|y_{obs}, x)
\]

and observed-data posterior also can be factored as

\[
p(\theta, \psi|y_{obs}, R, x) = \{f(\psi)L_1(\psi|R, y_{obs}, x)\}\{f(\theta)L_2(\theta|y_{obs}, x)\}.
\]

Therefore, the observed-data posterior for \( \theta \), the parameters of the full data response model, is

\[
p(\theta|y_{obs}, R, x) = f(\theta)L_2(\theta|y_{obs}, x)
\]

which does not contain \( \psi \). So the missing data mechanism does not need to be explicitly modeled. In the following chapters, we assume missing data is ignorable.

5. Nonparametric Regression

Assume we have \( n \) pairs of observations \((Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)\). The response variable \( Y \) is related to covariant \( x \) with following regressions

\[
Y_i = r(x_i) + \epsilon_i, \quad i = 1, 2, \ldots, n
\]  

(1-14)

where \( r \) is a regression function and \( \epsilon_i \) is a mean zero random process. This specification implies \( r(x) = E(Y|X = x) \) and we want to estimate the function \( r \).

5.1 Kernel Estimation

Nadaraya (1964) and Watson (1964) introduce an estimator of \( r(x) \) which is the special case of fitting a constant locally at any point \( x_0 \). The Nadaraya-Waston estimator of \( r(x) \) is defined as follows,
\[
\hat{r}(x) = \frac{\sum_{|x_i-x_0|<h(x_i)} Y_i W(x_i-x_0)}{\sum_{|x_i-x_0|<h(x_i)} W(x_i-x_0)}
\]  

(1–15)

where \( w(x_i-x_0) = \frac{K((x_i-x_0)/h(x_i))}{h(x_i)} \), \( h(x_i) \) is a smoothing parameters, and \( K(x) \) is a kernel function which satisfies the following conditions,

- \( K(x) \geq 0 \)
- \( \int K(x)dx = 1, \int xK(x)dx = 0, \) and \( \sigma_k^2 \equiv \int x^2K(x)dx > 0. \)

Commonly used kernels include,

- **Gaussian Kernel**: \( K(x) = \frac{1}{2\pi} \exp\{ -\frac{x^2}{2} \} \),
- **Boxcar Kernel**: \( K(x) = \frac{1}{2} I(|x| \leq 1) \)
- **Epanechnikov Kernel**: \( K(x) = \frac{3}{4} (1-x^2) I(|x| \leq 1) \)
- **Tricube Kernel**: \( K(x) = \frac{70}{81} (1-|x|^3)^2 I(|x| \leq 1) \)

We define an estimator \( \hat{r}(x) \) of \( r(x) \) as the one which minimizes a selected loss function.

The most commonly used loss function is \( L_p \) loss which is defined as follows,

\[
L_p = \left\{ \int |r(x) - \hat{r}(x)|^p dF(x) \right\}^{\frac{1}{p}}.
\]  

(1–16)

\( L_1 \) loss is resistant to outliers and invariant under 1-1 transformations. However, \( L_2 \) loss (squared error loss) gives many nice properties. For example, under squared error loss, the Bayesian estimator is the posterior mean \( \hat{\theta} = E(\theta | Y) \), and mean square error is sum of variance and squared bias \( E(\hat{Y} - Y)^2 = (E(\hat{Y} - E(Y)))^2 + E(Y - E(Y))^2. \)

Under squared error loss,

\[
\arg\min_{\hat{r}(x)} E(r(x) - \hat{r}(x))^2 = \arg\min_{\hat{r}(x)} \{ \text{Bias}^2 + \text{Variance} \} \equiv \arg\min_h \{ \text{Bias}^2 + \text{Variance} \}
\]

Clearly, \( \hat{r}(x) \) is a function of smooth parameter \( h(x_i) \).
The decrease in Bias is accompanied by increasing variance and we need a balance between these two to minimize squared error loss (or mean integrated squared error loss).

There is an extensive literature on how to choose a smoothing parameter $h$. We outline some of the more relevant literature next.

### 5.2 Choose Smooth Parameter

Commonly methods to estimate the smoothing parameter include

- cross-validation method (Stone (1974)).
- data driven plug-in method (Gasser et al. (1991); Ruppert et al. (1995)).

Leave-one-out cross-validation score is defined as:

$$CV(h) = \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{r}_{(-i)}(x_i))^2,$$

(1–17)

where $\hat{r}_{(-i)}$ is the estimator obtained by omitting the $i^{th}$ pair $(Y_i, X_i)$. We choose $h$ to minimize $CV(h)$. Leave-one-out cross validation is the simplest method for model selection but it is computationally expensive, does not provide a consistent estimator, and results serious problem as data are clustered (Craven and Wahba (1979)).

There are many modified versions to overcome these deficiency, such as biased cross-validation (Scott and Terrell (1987)), partitioned cross validation (Marron (1987)), and generalized cross validation (Craven and Wahba (1979)). Generalized cross validation has received much attention. Golub et al. (1979) it modifies CV by introducing a factor $\frac{1}{(1-hw_0/n)^2}$, i.e.

$$GCV(h) = \hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \frac{(Y_i - \hat{r}_{(-i)}(x_i))^2}{(1 - hw_0/n)^2}$$

$h$ is chosen to minimize $GCV(h)$.

Data driven plug-in rules have received more attention recently (Gasser et al. (1991); Ruppert et al. (1995)). The optimal bandwidth is calculated by estimating some
related quantities, such as variance ($\sigma^2$) and curvature of an unknown function ($r''(x)$). It is computationally efficient compared to cross-validation methods. However, the pilot bandwidth, which is minimizer of mean square error of estimating $\sigma^2$ and $r''(x)$ with an arbitrary specification may result in oversmoothing and/or missing important features (Loader (1999); Park and Marron (1990a)).

Generally, cross validation is more robust, while the plug-in rule is more efficient but depends on some stronger assumptions.

6. Outline of the Dissertation

My dissertation has three parts. First, we develop Bayesian methods to model dependence in longitudinal data via partial autocorrelations and marginal variances. Second, we propose a method to estimate large covariance matrices by banding the partial autocorrelation matrix. Third, we explore nonparametric smoothing of a partial autocorrelation matrix within bands.

**Bayesian Modeling dependence structure:** Many parameters and positive-definiteness are two major obstacles in estimating and modelling a correlation matrix for longitudinal data. In addition, when longitudinal data is incomplete, incorrectly modelling the correlation matrix often results in bias in estimating mean regression parameters. In Chapter 2, we introduce a flexible class of regression models for a covariance matrix parameterized using marginal variances and partial autocorrelations. We propose a class of priors for the regression coefficients and examine the importance of correctly modeling the correlation structure on estimation of longitudinal (mean) trajectories via simulations. The regression approach is illustrated on data from a longitudinal clinical trial.

**Estimating large covariance matrix:** In Chapter 3, we propose a computationally efficient approach to estimate (large) p-dimensional correlation matrices of ordered data based on an independent sample of size $n$ by banding the partial autocorrelation matrix. The number of bands ($k$) is chosen by an exact multiple hypothesis testing procedure.
This approach is considerably faster than many existing methods and only requires inversion of \( k \) dimensional covariance matrices. In addition, the resulting estimator is guaranteed to be positive definite as long as \( k \leq n - 2 \) (even when \( n < p \)). We evaluate our estimator via extensive simulations and compare it to the Ledoit-Wolf estimator. We also illustrate the approach using high-dimensional sonar data.

**Nonparametric modeling of the correlation structure:** In Chapter 4, we introduce a nonparametric estimator of a correlation matrix based on smoothing the partial autocorrelations within bands. We improve upon the estimator of a correlation matrix from Chapter 3. An asymptotic optimal bandwidth is derived based on Toeplitz condition on a banded partial autocorrelation matrix, and a plug-in bandwidth selector is suggested. The improvement is demonstrated by simulation and the method is applied to two high dimension data sets.
Many parameters and positive-definiteness are two major obstacles in estimating and modelling a correlation matrix for longitudinal data. In addition, when longitudinal data is incomplete, incorrectly modelling the correlation matrix often results in bias in estimating mean regression parameters. In this paper, we introduce a flexible class of regression models for a covariance matrix parameterized using marginal variances and partial autocorrelations. The partial autocorrelations can freely vary in the interval $(-1, 1)$ while maintaining positive definiteness of the correlation matrix so the regression parameters in these models will have no constraints. We propose a class of priors for the regression coefficients and examine the importance of correctly modeling the correlation structure on estimation of longitudinal (mean) trajectories via simulations. The regression approach is illustrated on data from a longitudinal clinical trial.

2.1 Background

Longitudinal data, measurements on the same subject over time, arise in many areas, from clinical trials to environmental studies. In such studies, to draw valid inference, the covariance between repeated observations on the same individuals needs to be properly modeled. In particular, in incomplete longitudinal data, mis-modeling the covariance matrix can result in biased estimates of fixed effect mean parameters (Little and Rubin (2002); Daniels and Hogan (2008)). Two major obstacles for modeling covariance matrices are 1) the (potentially) high dimensionality, and 2) positive-definiteness.

2.1.1 Brief literature Review for Estimating a Correlation Matrix

Many approaches have been proposed for estimating a covariance matrix more efficiently, whether by shrinking eigenvalues to obtain more stability (Yang and Berger (1994); Efron and Morris (1976)) or reducing the dimension via structure (Leonard and Hsu (1992); Chiu et al. (1996); Pourahmadi (1999, 2000); Daniels and Zhao (2003)).
There has also been research on shrinkage to introduce stability in structured ways (Daniels and Kass (1999); Daniels and Pourahmadi (2002)) or without structure (Liechty et al. (2004); Wong et al. (2003)). These approaches can often be thought of in terms of specific decompositions of a covariance matrix. Our approach will focus on the variance/correlation decomposition, used recently by Barnard et al. (2000), which decomposes the covariance matrix $\Sigma$ into $\Sigma = DRD$, where $R$ is correlation matrix and $D$ is diagonal matrix of standard deviations. Our approach here will rely on this decomposition and a further decomposition of the correlation matrix $R$ into partial autocorrelations which we review next.

### 2.1.2 Review of Partial Autocorrelations and Modeling Correlation Matrices

Consider a $p \times p$ correlation matrix $R$ with $(j, j+k)$th element, $\rho_{j+j+k} = \text{Cor}(Y_j, Y_{j+k})$. The matrix $R$ can be re-parameterized using partial autocorrelations, $\pi_{j+j+k} = \text{Cor}(Y_j, Y_{j+k} | Y_i, j < l < j+k)$. To provide an expression for them as a function of the marginal correlations, we first define some notation.

Let $R[j : j+k] = R[j : j+k, j : j+k]$ be the $k + 1 \times k + 1$ submatrix of $R$ which takes elements from $jth$ row to $(j+k)th$ row and $jth$ column to $(j+k)th$ column. Then we partition $R[j : j+k]$ as follows

$$R[j : j+k] = \begin{pmatrix}
1 & r_1(j, k) & \rho_{j,j+k} \\
 r_{1}^T(j, k) & R_2(j, k) & r_{3}^T(j, k) \\
 \rho_{j+k,j} & r_3(j, k) & 1
\end{pmatrix},$$

where $r_1(j, k) = (\rho_{j,j+1}, \ldots, \rho_{j,j+k-1})$, $r_3(j, k) = (\rho_{j+k,j+1}, \ldots, \rho_{j+k,j+k-1})$, and $R_2(j, k)$ contains the middle $k - 1$ rows and columns of $R[j : j+k]$.

The partial autocorrelations have the following form as a function of the marginal correlations,

$$\begin{cases}
\pi_{j,j+k} = \rho_{j,j+k} | j+1, \ldots, j+k-1 = \frac{\rho_{j,j+k} - r_1(j,k)R_2^{-1}(j,k)r_{3}^T(j,k)}{\sqrt{1 - r_1(j,k)R_2^{-1}(j,k)r_{3}^T(j,k)}} \frac{1}{\sqrt{2}} \quad 2 \leq k \leq p - j
\end{cases}$$

(2–1)
The marginal correlations, $\rho_{j,j+k}$ can also be written as a simple function of the partial autocorrelations,

$$\rho_{j,j+k} = r_{jk} + \pi_{j,j+k}A_{jk}, \quad (2-2)$$

where $r_{jk} = r_1(j, k)R_{1}^{-1}(j, k)r_{T}^T(j, k)$ and

$$A_{jk}^2 = [1 - r_1(j, k)R_{2}^{-1}(j, k)r_{T}^T(j, k)][1 - r_3(j, k)R_{2}^{-1}(j, k)r_{T}^T(j, k)]. \quad (2-3)$$

One of advantages of this parameterization is that $\pi_{jk}$ can vary independently in $(-1, 1)$ while maintaining positive definiteness of $R$, unlike $\rho_{jk}$ (see, e.g., Joe (2006)). Based on reparameterizing the marginal correlations into partial autocorrelations, Daniels and Pourahmadi (2009) introduce a prior for $R$ induced by independent uniform priors on the partial autocorrelations, i.e., $p(\pi) = 2^{-p(p-1)/2}$.

We reparameterize $R = (\rho_{ij})$ in terms of the entries of the partial correlation matrix $\Pi = (\pi_{jk})$, with the partial autocorrelations defined above and $\pi_{jj} \equiv 1$. We will transform these partial autocorrelations using Fisher’s z transformation mapping $\Pi$ to $\tilde{\Pi}$ where the off diagonal elements of the latter take values in entire real line $(-\infty, \infty)$. Moving from a constrained $R$ to a real symmetric matrix $\tilde{\Pi}$ gives us a link function framework similar to the theory of generalized linear models in McCullagh and Nelder (1989). These models will extend recent models from the literature for correlation matrices including the multivariate probit (Czado (2000)) and related models (Daniels and Normand (2006)).

2.1.3 Outline of this Chapter

This chapter is arranged as follows. In Section 2.2, we introduce regression models for the partial autocorrelations and marginal variances. We derive and investigate priors for the regression parameters for the partial autocorrelation and marginal variance parameters in Section 2.3. We provide details on posterior computations in Section 2.4. Results of a simulation study to investigate correlation structure misspecification are
given in Section 2.5. Application of the models to a schizophrenia clinical trial is given in Section 2.6. Section 2.7 provides conclusions and extensions.

2.2 Models for the Covariance Matrix

Let \( Y_i : i = 1, \ldots, n \) be a \( p \times 1 \) vector of longitudinal responses measured (without loss of generality) at times \( 1, \ldots, p \) with distribution

\[
Y_i \sim N_p(x_i^T \beta, \Sigma_i),
\]

(2–4)

where \( \beta \) is a vector of (mean) regression parameters with dimension \( p_\beta \times 1 \), \( x_i \) is a \( p_\beta \times p \) covariate matrix, and \( \Sigma_i = D_i R_i D_i^T \). We build regression models for \( R_i \) via the partial autocorrelations and \( D_i \), via the marginal variances in the following subsections.

2.2.1 Partial Autocorrelations

Consider the following regression model for \( \pi_{i,j,k} \), the \( jk \)th partial autocorrelation for subject \( i \),

\[
z(\pi_{i,j,k}) = w_{i,j,k}^* \gamma,
\]

(2–5)

where \( z(\cdot) \) is Fisher’s z-transform, \( z(\pi) = \frac{1}{2} \log \frac{1 + \pi}{1 - \pi} \) and \( w_{i,j,k}^* \) is a \( 1 \times q \) vector of covariates to model structure and subject-level covariates; \( \gamma \) is unconstrained in q-dimension real space \( \mathbb{R}^q \). Given that the partial autocorrelations are correlations between longitudinal observations, conditional on intermediate ones, we might expect higher order ones to be zero. For example, we might specify \( w_{i,j,k}^* = I(|k - j| = 1) \) corresponding to an AR(1) structure with all lag partial autocorrelations bigger than one equal to zero. \( w_{i,j,k}^{*T} = (1, |k - j|) \) implies that \( \Pi \) has a Toeplitz form with z-transform of the element on each subdiagonals having a linear relationship in lag. For related structures for the parameters of the modified Choleski decomposition (see Pourahmadi (1999); Pourahmadi and Daniels (2002)).
2.2.2 Marginal Variances

We assume the logarithm of the marginal standard deviations, \( \sigma_{i,j} \) (i.e., the \( j \)th diagonal element of \( D_i \)) follow the regression models,

\[
\log(\sigma_{i,j}) = A_{i,j} \eta,
\]

where \( A_{i,j} \) is a \( 1 \times q_0 \) vector of covariates to model structure and unit-level covariates. For example,

\[
A_{i,j} = (I(j = 1), I(j > 1))
\]

induces a structure of equal variance except for time 1. The following,

\[
A_{i,j} = (1, j)
\]

corresponds to the marginal variances that are log linear in time. Verbyla (1993) proposed models for the marginal variances (residual variances) in terms of unit level covariates (i.e., heterogeneity) in the setting of independent responses.

2.3 Priors for \( \gamma \)

Standard diffuse priors for \( \gamma \) in Eq. 2–5, e.g., an improper uniform prior on \( \mathbb{R}^q \) or a diffuse normal prior, result in most of the mass for the partial autocorrelations, \( \pi_{i,jk} \) being put at \(-1\) and \(+1\). This happens in many settings with diffuse priors on transformed spaces, e.g., coefficients in logistic regression (see Agresti and Hitchcock (2005)). These are not sensible prior beliefs. In the next two subsections, we will review a prior proposed for the partial autocorrelations from Daniels and Pourahmadi (2009) and propose an alternative one that both avoid this behavior for an unstructured \( \Pi \). We then propose a way to use these priors, which are both within the class of independent Beta priors, to construct priors for \( \gamma \) and point out their connections to g-priors. We also construct a similar prior for \( \eta \) in Eq. 2–6.
2.3.1 Review of Priors for Unstructured Partial Autocorrelations

Independent uniform priors, a special case of a transformed $Beta(-1,1)(a, b)$ distribution with shape parameters $a=1$ and $b=1$, on partial autocorrelations induce desirable behavior for longitudinal (ordered) data by shrinking higher lag marginal correlations toward zero (Daniels and Pourahmadi (2009)). The behavior can be understood by examining the Jacobian from $R$ to $\pi$,

$$J(R \rightarrow \pi) = \prod_{k=1}^{p-1} \prod_{j=1}^{p-k} (1 - \pi_{j,k}^2)^{(p-1-k)/2}.$$ 

As lag increases, more mass is placed toward zero. This is not surprising since most priors on $R$ do not use information on potential ordering of the responses and induce this prior form on partial correlations to obtain identical marginal priors for the marginal correlations.

This behavior of the $Beta(1, 1)$ prior is consistent with serial correlation often seen in longitudinal and ordered data. However, it does not favor positive correlations as we typically see in longitudinal data.

2.3.2 An alternative Prior for Unstructured Partial Autocorrelations

Here we introduce a prior on the partial autocorrelations that favors positive correlations to negative correlations. We propose independent priors on $\pi_{jk}$ with pdf’s, $p(\pi_{jk}) = \frac{1+\pi_{jk}}{2}$, which is a transformed $Beta(-1,1)(a, b)$ distribution with parameters $a=2$ and $b=1$; we refer to these as triangular priors given their shape. The implied marginal priors for $\rho_{jk}$ are given in Figure 2-1B. The priors have decreasing mass close to 1 as lag $(|j - k|)$ increases. This is consistent with serial correlation often seen in longitudinal data and favors more mass on positive correlations than the $Beta(1, 1)$ priors.

In the following section, we will try to use these two priors as a starting point to construct a prior for the regression coefficients, $\gamma$ in Eq. 2–5. In the remaining, all Beta priors will be specified on the interval $(-1, 1)$, but we just denote them as $Beta(a, b)$. 
2.3.3 Proposed Prior on $\gamma$

In the following, when not needed we will drop the subscripts on $\pi_{jk}$. We start by deriving the distribution of $z(\pi_{jk})$ when the $\pi_{jk}$ follow independent $Beta(1, 1)$ priors. For this prior on $\pi$, $z(\pi) = z = \frac{1}{2} \log \frac{1+\pi}{1-\pi}$, with pdf

$$f(z) = \frac{2e^{2z}}{(1 + e^{2z})^2},$$ \hspace{1cm} (2–9)

where $z \in (-\infty, +\infty)$. This is the pdf of a logistic distribution, $z \sim \text{logistic}(0, \frac{1}{2})$. It is well known that the logistic distribution can be approximated with a t-distribution (Albert and Chib (1993)). However, the easy to use construction of the multivariate t-distribution as a gamma mixture of normals has t-marginals but they are not independent as we need based on our original specification of independent Beta’s. As a result, we will use a normal approximation to the logistic distribution, whose multivariate version does have independent marginals, $z \sim N(0, \frac{\pi}{12})$; that is, the random vector $z \sim N(0, \frac{\pi}{12}I_{T \times T})$, where $T = \frac{p(p-1)}{2}$.

Figure 2-1A shows how well the normal prior approximates the original uniform prior on the hypercube (i.e., the independent Beta(1,1) priors) in terms of the marginal correlations. The upper triangular elements represent the marginal priors of $\rho_{jk}$ from the original uniform prior and the lower triangular elements represent the marginal priors of $\rho_{jk}$ from the prior based on the normal approximation. The approximate prior appears to behave sufficiently similarly.

Now, we show how this prior can be used to construct a prior for $\gamma$ in Eq. 2–5. We first focus on the case of $z(\pi_{i,jk}) = z(\pi_{jk})$ and for ease of notation, let $z_{jk} = z(\pi_{jk})$ and $z = (z_{12}, ..., z_{1p}, z_{23}, ...z_{2p}, ...z_{p-1p})^T$. Consider the full rank linear transformation $z = w\gamma$, where

$$w = \left(\begin{array}{c} w^* \\ w^\perp \end{array}\right).$$
\( w^* \) is a \( T \times q \) \((T \geq q)\) full column rank matrix corresponding to the regression in Eq. 2–5. The matrix \( w^\perp \) is a \( T \times (T - q) \) full column rank matrix such that \((w^*)^T \times w^\perp = 0_{q \times (T - q)}\) and \((w^\perp)^T \times w^\perp = I_{(T - q) \times (T - q)}\). Therefore

\[
\gamma = \begin{pmatrix} \gamma^* \\ \gamma^\perp \end{pmatrix} = \begin{pmatrix} w^* & w^\perp \end{pmatrix}^{-1} z \quad \text{with} \quad \begin{pmatrix} w^* & w^\perp \end{pmatrix}^{-1} = \begin{pmatrix} ((w^*)^T w^*)^{-1}(w^*)^T \\ (w^\perp)^T \end{pmatrix} \tag{2–10} 
\]

Thus, \( z = \begin{pmatrix} w^* & w^\perp \end{pmatrix} \begin{pmatrix} \gamma^* \\ \gamma^\perp \end{pmatrix} \), is a 1–1 transformation from \( z \) to \( \gamma \).

We define \( E(z) = \mu \) and \( \text{Var}(z) = \sigma^2 \) based on the multivariate normal prior on \( z \). Under the Beta(1,1) prior on \( \pi \), \( \mu = 0 \) and \( \sigma^2 = \pi/12 \); under the Beta(2,1) prior (triangular prior), \( \mu = \frac{1}{2} \) and \( \sigma^2 = 0.5722 \). The corresponding prior for \( \gamma \) is also multivariate normal with mean and variance given below,

\[
E(\gamma) = \begin{pmatrix} w^* & w^\perp \end{pmatrix}^{-1} E(z(\pi)) = \begin{pmatrix} ((w^*)^T w^*)^{-1}(w^*)^T \\ (w^\perp)^T \end{pmatrix} \mu 1_{T \times 1} 
\]

\[
= \mu \begin{pmatrix} ((w^*)^T w^*)^{-1}(w^*)^T \\ (w^\perp)^T \end{pmatrix} 1_{T \times 1} 
\]

and

\[
\text{Var}(\gamma) = \begin{pmatrix} w^* & w^\perp \end{pmatrix}^{-1} \text{Var}(z(\pi))\left(\begin{pmatrix} w^* & w^\perp \end{pmatrix}^{-1}\right)^T = \begin{pmatrix} ((w^*)^T w^*)^{-1}(w^*)^T \\ (w^\perp)^T \end{pmatrix} \sigma^2 I_{T \times T} \begin{pmatrix} ((w^*)^T w^*)^{-1}(w^*)^T w^\perp \\ ((w^*)^T w^*)^{-1}(w^*)^T \end{pmatrix} = \sigma^2 \begin{pmatrix} ((w^*)^T w^*)^{-1}(w^*)^T \\ (w^\perp)^T \end{pmatrix} \begin{pmatrix} ((w^*)^T w^*)^{-1}(w^*)^T w^\perp \\ ((w^*)^T w^*)^{-1}(w^*)^T \end{pmatrix} = \sigma^2 \begin{pmatrix} 0 \\ 0 \end{pmatrix}. 
\]
The resulting prior for $\gamma^*$ is also multivariate normal with expectation and variance,

$$E(\gamma^*) = \mu (w^*^T w^*)^{-1} w^*^T 1_{T \times 1} \text{ and } \text{Var}(\gamma^*) = \sigma^2 (w^*^T w^*)^{-1}.$$  

The dimension reduction from $z$ to $\gamma^*$ results in the prior variance being too small. To see this note that the variance of $i$th component of $z$ in $z = w \gamma$ is

$$\text{var}(z_i(\gamma)) = \text{var}(w_{(i,:)\gamma})$$
$$= \text{var}(w^*_{(i,:)\gamma^*}) + \text{var}(w^\perp_{(i,:)\gamma^\perp})$$
$$= \sigma^2(w^*_{(i,:)\gamma^*})^T (w^*_{(i,:)\gamma^*})^T + w^\perp_{(i,:)\gamma^\perp})^T.$$  

The $i$th component of $z = w^* \gamma^*$ has variance

$$\text{var}(z_i(\gamma^*)) = \text{var}(w^*_{(i,:)\gamma^*})$$
$$= \sigma^2(w^*_{(i,:)\gamma^*})^T (w^*_{(i,:)\gamma^*})^T.$$  

Clearly, $\text{var}(z_i(\gamma)) > \text{var}(z_i(\gamma^*))$. It is easy to adjust for this by noting that the average variance of $z(\pi_{jk}) = w^*_{jk} \gamma^*$, $\text{var}(z)$ is

$$\overline{\text{var}(z)} = \frac{1}{T} \sum_{i=1}^{T} \text{var}(z_i)$$
$$= \sigma^2 \frac{1}{T} \text{tr}\{w^* (w^*)^T w^* (w^*)^T\}$$
$$= \sigma^2 \frac{1}{T} \text{tr}\{((w^*)^T w^*)^{-1}(w^*)^T w^*\}$$
$$= \sigma^2 \frac{1}{T} \text{tr}\{I_{q \times q}\}$$
$$= \frac{q}{T} \sigma^2.$$  

where $\sigma^2$ is the desired variance. Hence we can inflate $\text{var}(\gamma^*)$ by a factor of $\frac{T}{q}$. The resulting prior for $\gamma^*$ is $\gamma^* \sim N(\mu ((w^*)^T w^*)^{-1}(w^*)^T 1_{T \times 1}, \frac{T}{q} \sigma^2 ((w^*)^T w^*)^{-1}).$
2.3.4 Extension to Unit-level Covariates

We can easily extend this prior to subject specific covariates. Suppose, for \( i = 1, \ldots, n \),

\[
z(\pi_{i,jk}) = w^*_{i,jk} \gamma.
\]  

(2–11)

Let \( w_i^* \) be a \( p \times q \) matrix such that \( z_i = w_i^* \gamma \). We first stack \( z_1, \ldots, z_n \) and \( w_1^*, \ldots, w_n^* \) together,

i.e., \( z = (z_1^T, z_2^T, \ldots, z_n^T)^T \), and \( w^* = (w_1^*T, w_2^*T, \ldots, w_n^*T)^T \). So, we have

\[
z = w^* \gamma
\]

and \( w^* \) is \( nT \times q \) full column rank matrix. Similar to the previous case,

\[
\gamma \sim N(\mu(\mu(\mu(\mu(w^*T w^*)^{-1}w^*T 1_{nT \times 1}) \frac{nT}{q} \sigma^2 (w^*T w^*)^{-1})).
\]

We can rewrite this as

\[
\gamma \sim N(\mu(\mu(\mu(\mu(\mu(\sum_{i=1}^n w_i^*T w_i^*)^{-1}w^*T 1_{nT \times 1}) \frac{nT}{q} \sigma^2 (\sum_{i=1}^n w_i^*T w_i^*)^{-1})),
\]  

(2–12)

which is our recommended prior in the general case.

2.3.5 Connection to g-priors

Our priors on \( \gamma \) have similar form to the \( g \)-priors introduced by Zellner (1986). However, our derivation begins with a prior on an unconstrained parameter space as opposed to Zellner’s construction of a prior based on the posterior distribution of imaginary data \( y_0, y_0 = x^T \beta + \epsilon \) where \( \epsilon \sim N(0, \sigma_0^2 I_n) \) (with independent priors on \( \beta \propto 1 \) and \( \sigma_0 \propto \frac{1}{\sigma_0} \)). The Zellner prior for \( \beta | \sigma_0 \) has the form \( \beta \sim N(\hat{\beta}_0, \frac{\sigma_0^2}{g} (xx^T)^{-1}) \), where \( \hat{\beta}_0 \) is the least squares estimate based on the imaginary data and \( g \) is a penalty parameter; in practice, the mean is typically set to zero so no imaginary data is actually required. Our prior has a similar form but it is based on the projection of \( z(\pi_i) \) on \( w_i^* \) with weights based on the original prior for \( \pi \) on the unconstrained space (here a hypercube). The ‘weights’ based on the prior in Eq. 2–12 come in through \( \mu \) in the prior mean,
\[(\sum_{i=1}^{n} w_i^* w_i^*)^{-1} w^T \mu 1_{n \times 1} \text{ and } \sigma^2 \text{ in the prior variance, } \sigma^2 (w^T w)^{-1}. \] As a result, with these priors, we do not have to deal with the issue of the choice of \( g \) (for some discussion see George and Foster (2000); Clyde and George (2000)).

### 2.3.6 Prior for \( \eta \)

The most commonly used prior on the marginal variances \( \sigma \) is the inverse gamma prior, which facilitates computations due to conditional conjugacy. Daniels (2006) used a uniform prior on the transformed innovation (IV) parameters with or without structure similar to the models in Section 2.2 for the marginal standard deviations. Barnard et al. (2000) discussed independent normal priors on logarithmic transformed \( \sigma \). In particular, they proposed the following prior

\[
\log(\sigma_i) \sim N(\xi, \Lambda),
\]

with \( \Lambda \) diagonal. We will derive a prior for \( \eta \) similar to that for \( \gamma \) based on Barnard et al.’s prior for the marginal standard deviations. The resulting prior is

\[
\eta \sim N(\lambda(\sum_{i=1}^{n} A_i^T A_i)^{-1} A^T 1_{np \times 1}, \tau^2 \frac{np}{q_0}(\sum_{i=1}^{n} A_i^T A_i)^{-1})
\]

(2–14)

Note in the derivation, we have assumed \( \xi = \lambda 1_{p \times 1} \) and \( \Lambda = \tau^2 I_{p \times p} \) in Eq. 2–13, where \( \lambda \) and \( \tau \) are fixed.

**Posterior distribution and computations:** The full data likelihood, \( L(\eta, \beta, \gamma | y) \) is proportional to

\[
\sum_{i=1}^{n} |D_i(\eta)R_i(\gamma)D_i(\eta)|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \text{tr}[\sum_{i=1}^{n} R_i(\gamma)^{-1}D_i(\eta)^{-1}(Y_i-x_i\beta)(Y_i-x_i\beta)^T D_i(\eta)^{-1}]\}
\]

We specify the following priors for \( \beta, \eta, \) and \( \gamma \),

\[
\beta \propto 1
\]

(2–15)

\[
\eta \sim N(\lambda_0(\sum_{i=1}^{n} A_i^T A_i)^{-1} A^T 1_{np \times 1}, \tau^2 \frac{np}{q_0}(\sum_{i=1}^{n} A_i^T A_i)^{-1})
\]

(2–16)
\[ \gamma \sim N(\mu \left( \sum_{i=1}^{n} w_i^* T \right)^{-1} w_i^* T 1_{nT \times 1}, \frac{nT}{q} \sigma^2 (\sum_{i=1}^{n} w_i^* T w_i^*)^{-1}) \]  

(2–17)

where \( \sum_{i=1}^{n} A_i^T A_i \) and \( \sum_{i=1}^{n} w_i^* T w_i^* \) are non-singular. In the setting of incomplete longitudinal responses, under an assumption of ignorable missingness, we only need to specify the full data response model and the likelihood of interest is the observed data likelihood, \( L(\beta, \gamma, \eta | y_{obs}, x) \), where the observed data response is \( y_{obs} \) (Daniels and Hogan (2008)); the form of the observed data likelihood is given in the supplementary materials.

Since we specify an improper prior on \( \beta \), we need to prove the posterior distribution of \( (\beta, \gamma, \eta) \) is proper. In the next section, we provide a theorem which gives simple sufficient conditions under which the posterior is proper. The supplementary materials contain details on the MCMC algorithm to sample from the posterior distribution.

### 2.3.7 Posterior Propriety

In the following theorem, we state conditions that are sufficient for the posterior to be proper. First, we need to introduce some notation. Suppose full-data \( Y_i : i = 1, \ldots, n \) are independently distributed random variables with distribution \( Y_i \sim N_p(x_i^T \beta, \Sigma_i) \), where \( x_i \) is a \( p \times \beta \) covariate matrix, \( \beta \) is a \( \beta \times 1 \) (mean) regression parameter vector, \( \Sigma_i = D_i^{1/2} \gamma R_i(\gamma) D_i^{1/2} \gamma^T \), and \( D_i(\eta) = \text{diag}(\sigma^2_{i1}, \sigma^2_{i2}, \ldots, \sigma^2_{ip}) \) specified by Eq. 2–6 and \( R_i(\gamma) \) specified by Eq. 2–1 and Eq. 2–5; define \( \mathfrak{H}_\beta, \mathfrak{H}_\gamma, \mathfrak{H}_\eta \) to be sample spaces of \( \beta, \gamma, \eta \), respectively. Let \( (Q_{i1}, \ldots, Q_{ip})^T \) be a vector of missing data indicators and let \( Y_i^{k_{i1}} = \{ Y_{ij}, j = 1, \ldots, k, \text{where } k_i : Q_{ik_i} = 1, Q_{ik_{i+1}} = 0 \} \), \( S_k = \{ i, Q_{ik_i} = 1 \} \), and \( Q_{ik_{i+1}} = 0, \text{and } k_i = k \), where \( 1 \leq k \leq p - 1; i = 1, \ldots, n \).

**Theorem 2.1.** We assume the observed data distribution for the \( i \)th subject \( (i = 1, \ldots, n) \) is \( Y_i^{k_{i1}} \sim N(k_i x_{i1}^T \beta, \Sigma_i^{k_i}) \), where \( x_{i1}^k = x_{i[1:k_i]} \) is a \( p \times k_i \) submatrix of \( x_i \) and \( \Sigma_i^{k_i} = \Sigma_{i[1:k_i]} \) is a \( k_i \times k_i \) principal submatrix of \( \Sigma_i \). We also assume the priors on the parameters are given by Eq. 2–15-Eq. 2–17 and missingness is ignorable. Then the posterior of \( (\beta, \gamma, \eta) \) will be proper under the following three (easy to check) conditions:
1. $\sum_{i \in S_k} x_i^k x_i^{kT}$ is non-singular for all $k \in \{1, 2, \ldots, p - 1\}$.

2. $\sum A_i A_i^T$ is non-singular.

3. $\sum w_i^* w_i^{*T}$ is non-singular.

The proof is given in the supplementary materials. Note that the three conditions are conditions for the three design matrices in our model (for the mean, the variance, and the correlations, respectively); the latter two guarantee that the priors Eq. 2–16 and Eq. 2–17 are proper.

### 2.4 Simulations

#### 2.4.1 Models

To assess the importance of the correlation structure on estimating (mean) longitudinal trajectories in incomplete data, we conducted a simulation. The true model was Eq. 2–4 with $p = 6$. For each individual the rows of the mean design matrix were specified as an (orthogonal) quadratic trajectory, specified as in Eq. 2–19. We set $\beta = (27, -2.3, 0.50)^T$. We considered three sample sizes (30, 100, and 400). For each scenario we simulated 200 data sets.

The true models for marginal variances and partial autocorrelation coefficients were given by

$$z(\pi_{jk}) = \gamma_1 I(|k - j| = 1 \cap j = 1) + \gamma_2 I(|k - j| = 2) + \gamma_3 I(|k - j| = 1 \cap j > 1) \quad (2–18)$$

and

$$\log \sigma_{jj} = \eta_1 I(j = 1) + \eta_2 I(j > 1),$$

with $\gamma = (0.65, 0.21, 0.85)^T$, and $\eta = (150, 200)^T$. The structure on $\pi$ represents a second order model with the lag one partial autocorrelations constant except for time 1, the lag two partial autocorrelations constant over time, and higher lag partial autocorrelations equal to zero. The structure on the variances corresponds to a constant variance over time after time one.
After simulating the complete data, we induce ignorable missingness via the following missing data mechanism,

\[
\text{logit} P(Q_{ik} = 1|Q_{i,k-1} = 1, y_{\text{obs}}) = \alpha_1 + \alpha_2 y_{k-1},
\]

where \(Q_{jk} = I\{Y_{jk} \text{ is observed}\}\), \(y_{\text{obs}}\) denotes realization of observed data, and \(\alpha = (3.86324, -0.05)\).

We fit four models to the simulated data. For each model, we use the same true mean and marginal variance models, but different partial autocorrelation models. Our objective is to evaluate the impact of mis-specifying the partial autocorrelation model on inference on the marginal mean regression coefficients, \(\beta\). Specifically, the models we compare are:

1) True model for \(\pi\) given in Eq. 2–18
2) Independence model, \(\pi = 0\)
3) AR(1) model: \(z(\pi_{jk}) = \gamma_1 I\{|k - j| = 1\}\)
4) Unstructured model (no structure on \(\pi\))

For each of the 200 simulated datasets for each sample size, we ran 20,000 iterations for each of the four models. To compare inference on the mean under all four models, we computed the following two quantities: 1) Total MSE, sum of mean squared error of the components of \(\beta\) and 2) Change from Baseline, change of estimated mean responses from time one to time six. We also compare the mean trajectories graphically.

2.4.2 Results

The simulation results are given in Tables 2-1, 2-2 and Figures (2-2A, 2-2B, and 2-2C). As the sample size increases, the estimates for \(\beta\) quickly approach the true value for the true correlation model, more slowly for the unstructured correlation model and to the wrong values for the AR(1) and independence correlation models (with the latter with considerable bias) (Table 2-1). Table 2-2 presents a similar story with bias from the incorrect models and larger MSE’s for the estimates of the \(\beta\)’s and the change.
from baseline. Graphically, the fitted trajectories can be seen in Figures (2-2A, 2-2B, and 2-2C) and illustrate the bias in the fitted trajectory when the correlation structure is incorrect. Clearly, the unstructured model is consistent, however, it is quite unstable and variable for the smaller sample sizes. Wishart distribution is popular prior as we do Bayesian inference on variance-covariance of normal random vector. To illustration, we do model selection among Independent Model, AR(1) model, True Model, and Wishart prior. The DIC values are recorded in Table (2-5). From this table, it is easy to see that True model do best among all model we consider here and True model does much better job than wishart prior as sample size is large.

2.5 Data Example: Schizophrenia Trial

The data were collected as part of a randomized, double-blind clinical trial for a "new" pharmacologic treatment of schizophrenia (Lapierre et al. (1990)). The trial compared three doses of the "new" treatment (low, medium, high) to the standard dose of haloperidol, an effective antipsychotic that had known side effects. At the time of the study, the trial was designed to find the appropriate dosing level since the experimental therapy was thought to have similar antipsychotic effectiveness with fewer side effects. Two hundred forty-five patients were enrolled and randomized to one of the four treatment arms. The intended length of follow-up was 6 weeks, with measures taken weekly except for week 5. Schizophrenia severity was assessed using the Brief Psychiatric Rating Scale (BPRS) a sum of scores of 18 items that reflect behaviors, mode, and feelings. The scores ranged from 0 to 108 with higher scores indicating higher severity. To enter the study, the BPRS score had to be no less than 20. We will illustrate our approach using only the medium dose arm. The main inferential interest is the change in BPRS from the beginning to the end of the study.

The dropout rate on the medium dose arm was high, with only 40 out of 61 (about 66%) participants having a measurement at week 7 (the sixth measurement time). Reasons for dropout included adverse events (e.g., side effects), lack of treatment effect, and
withdrawal for unspecified reasons. The trajectories of completers vs. non-completers is shown in Figure 2-3A. Clearly those dropping out were doing worse prior to dropping out (higher BPRS).

Let the longitudinal vector of outcomes for subject \(i\) be \(Y_i = (Y_{i1}, \ldots, Y_{i6})^T\), measured at weeks \(t = (t_1, \ldots, t_6) = (1, 2, 3, 4, 5, 7)\). We assume \(Y_i\) follows \((2–4)\) with mean.

\[
E(Y_i) = \beta_0 + \beta_1 x_1^j + \beta_2 x_2^j, \tag{2–19}
\]

where \(x_1^j = (t_j - \bar{t})\) and \(x_2^j = (t_j - \bar{t})^2 - \frac{\sum_{k=1}^{5}(t_k - \bar{t})^3}{\sum_{k=1}^{5}(t_k - \bar{t})^2}(t_j - \bar{t}) - \frac{\sum_{k=1}^{5}(t_k - \bar{t})^2}{6}(t_j - \bar{t})\), i.e., an orthogonal quadratic polynomial. We assume missingness is ignorable.

We fit the five partial autocorrelation models given below:

**Independence Model:** \(z(\pi_{jj+k}) = 0, \log(\sigma_j) = I(j = 1)\eta_1 + I(j > 1)\eta_2\).

**AR(1) Model:** \(z(\pi_{jj+k}) = I(k = 1)\gamma_1, \log(\sigma_j) = I(j = 1)\eta_1 + I(j > 1)\eta_2\).

**Unstructured Covariance Model:** \(z(\pi_{jj+k}) = \gamma_{jj+k}\) (with \(\gamma = (\gamma_{12}, \gamma_{13}, \ldots, \gamma_{p-1,p})\)), \(\log(\sigma_j) = \eta_j\).

After examining the unstructured covariance matrix in Table 2-4, we consider two structured models;

**Structured Model 1:**

\[
z(\pi_{jj+k}) = I(k = 1 \cap j < 2)\gamma_1 + I(k = 1 \cap j > 1)\gamma_2 + I(k = 2)\gamma_3, \\
\log(\sigma_j) = I(j = 1)\eta_1 + I(j > 1)\eta_2.
\]

This is the same model as the one considered in the simulation.

**Structured Model 2:**

\[
z(\pi_{jj+k}) = I(k = 1 \cap j < 2)\gamma_1 + I(k = 1 \cap j > 1)\gamma_2 + I(k = 2 \cap j < 3)\gamma_3 \\
+ I(k = 2 \cap j > 2)\gamma_4 + I(k = 3)\gamma_5 + I(k = 4)\gamma_6 + I(k = 5)\gamma_7, \\
\log(\sigma_j) = I(j = 1)\eta_1 + I(j > 1)\eta_2.
\]
This model is more flexible for the partial autocorrelations allowing nonstationary lag one and lag two autocorrelations and stationary lag three, four and five (with no structural zeros). The structure on the variances is the same as Structured Model 1.

We use priors specified in Eq. 2–15, Eq. 2–16, and Eq. 2–17 for $\beta$, $\eta$, $\gamma$, respectively. For all models, we ran 200,000 iterations with no burn-in since they converged after a few iterations. The plot of all fitted mean trajectories is given in Figure 2-3B.

The mean BPRS initially decreased but started to go back up by week 5. This is related to those dropping out doing more poorly than those staying in the study. Table 2-5 contains the posterior mean of $\beta$, the change from baseline to week 7, their 95% credible intervals, and the DIC based on the observed data likelihood. The changes from baseline in all models were negative with 95% credible interval excluding 0, showing that Medium-dose reduced the BPRS score significantly, which agrees with earlier analysis done in Daniels and Hogan (2008). The changes from baseline varied from $-14$ to $-11$ based on the covariance model chosen. Based on the DIC, Structured Model 2 provided the best fit. The change from baseline in Structured Model 2 was almost a full point different from the unstructured model.

### 2.6 Discussions

In this paper, we first extended the priors in Daniels and Pourahmadi (2009) for partial autocorrelations for the unstructured case by introducing a set of triangular priors which favor positive marginal correlations. Using Fisher’s z-transformation on the partial autocorrelations $\pi$, we introduced a GLM framework for regression models to induce structure and/or unit-specific covariates in the correlation matrix. Based on priors proposed for the partial autocorrelations in the non-regression setting, we introduced a prior for the coefficients in the partial autocorrelation regressions (and for the coefficients of the marginal variance regressions). We conducted simulations that illustrated the importance of correct specification of the correlation structure in the
setting of ignorable missingness in longitudinal data and fit the models to data from a longitudinal schizophrenia clinical trial.

There are a variety of extensions to the modeling proposed here. Clearly, it can be difficult to find a good parametric model that imposes structure on the correlation matrix. Thus extending approaches developed under different parameterizations (Smith and Kohn (2002); Wong et al. (2003)) to our setting is an important extension. Correlation matrices (instead of covariance matrices) arise commonly in models for longitudinal data modeled using Gaussian copulas (e.g., the multivariate probit model) (Nelsen (1999)); efficient computations using the partial autocorrelation in these settings will be a challenging problem due to the lack of conjugacy. Finally, to offer some robustness to a selected model for the correlation structure, an alternative would be to shrink the partial autocorrelations to the structure using independent Beta priors as has been done previously using normal priors on other parameterizations of a covariance matrix (Daniels and Kass (2001); Daniels and Pourahmadi (2002))
Table 2-1. Posterior means of $\beta$

<table>
<thead>
<tr>
<th></th>
<th>Sample Size 30</th>
<th></th>
<th>Sample Size 100</th>
<th></th>
<th>Sample Size 400</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\text{Unstr}$</td>
<td>26.5</td>
<td>26.9</td>
<td>26.7</td>
<td>25.3</td>
<td>26.9</td>
<td>26.7</td>
</tr>
<tr>
<td>$\text{AR(1)}$</td>
<td>0.52</td>
<td>0.51</td>
<td>0.50</td>
<td>0.58</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>$\text{Indep.}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2-1. Marginal priors
Figure 2-2. Trajectories of sample size 30, 100, 400
Figure 2-3. Observed trajectories and fitted trajectories
Table 2-2. Summary measures for the simulation

<table>
<thead>
<tr>
<th></th>
<th>Total MSE</th>
<th></th>
<th>Change from Baseline</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstr</td>
<td>True</td>
<td>AR(1)</td>
<td>Indep.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.9</td>
<td>6.5</td>
<td>6.6</td>
<td>10.1</td>
<td>12.6</td>
</tr>
<tr>
<td>1.8</td>
<td>1.7</td>
<td>2.0</td>
<td>4.8</td>
<td>12.2</td>
</tr>
<tr>
<td>0.42</td>
<td>0.41</td>
<td>0.52</td>
<td>3.4</td>
<td>12.1</td>
</tr>
</tbody>
</table>
Table 2-3. Summary count from the simulations

<table>
<thead>
<tr>
<th></th>
<th>Sample Size 30</th>
<th></th>
<th>Sample Size 100</th>
<th></th>
<th>Sample Size 400</th>
<th></th>
<th>Sample Size 1000</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR(1)</td>
<td>Ind</td>
<td>True</td>
<td>Wish</td>
<td>AR(1)</td>
<td>Ind</td>
<td>True</td>
<td>Wish</td>
</tr>
<tr>
<td></td>
<td>True</td>
<td>Wish</td>
<td>AR(1)</td>
<td>Ind</td>
<td>True</td>
<td>Wish</td>
<td>AR(1)</td>
<td>Ind</td>
</tr>
<tr>
<td>69 (0.345)</td>
<td>0 (0)</td>
<td>131 (0.655)</td>
<td>0 (0)</td>
<td>2 (0.10)</td>
<td>0 (0)</td>
<td>198 (0.990)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>131 (0.655)</td>
<td>0 (0)</td>
<td>69 (0.345)</td>
<td>0 (0)</td>
<td>146 (0.730)</td>
<td>0 (0)</td>
<td>2 (0.010)</td>
<td>52 (0.260)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>0 (0)</td>
<td>30 (0.150)</td>
<td>0 (0)</td>
<td>170 (0.850)</td>
<td>52 (0.260)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>148 (0.740)</td>
<td>200 (1)</td>
</tr>
<tr>
<td>0 (0)</td>
<td>170 (0.850)</td>
<td>0 (0)</td>
<td>30 (0.150)</td>
<td>0 (0)</td>
<td>200 (1)</td>
<td>0 (0)</td>
<td>0 (0)</td>
<td>200 (1)</td>
</tr>
</tbody>
</table>
Table 2-4. Descriptive summaries of schizophrenia trial

<table>
<thead>
<tr>
<th>Variances (Diagonal) and Correlations (Off Diagonal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>126.25 0.6578 -0.0738 0.0804 -0.0253 -0.5230</td>
</tr>
<tr>
<td>0.7889 210.35 0.8543 -0.0593 -0.3328 0.0292</td>
</tr>
<tr>
<td>-0.0740 1.2718 224.42 0.8559 0.2648 0.4375</td>
</tr>
<tr>
<td>0.0806 -0.0594 1.2779 240.84 0.8961 0.3506</td>
</tr>
<tr>
<td>-0.0253 -0.3460 0.2713 1.4522 221.98 0.8433</td>
</tr>
<tr>
<td>-0.5805 0.0292 0.4692 0.3661 1.2325 243.08</td>
</tr>
</tbody>
</table>

Table 2-5. Posterior summaries of schizophrenia

<table>
<thead>
<tr>
<th>(β₀, β₁, β₂)</th>
<th>Changes from Baseline (95% CI)</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent</td>
<td>(25.6, -2.35, 0.68)</td>
<td>-14.1 (-18.7, -9.4)</td>
</tr>
<tr>
<td>AR(1)</td>
<td>(27.5, -1.97, 0.69)</td>
<td>-11.8 (-15.8, -7.9)</td>
</tr>
<tr>
<td>Unstructured</td>
<td>(26.9, -2.03, 0.62)</td>
<td>-12.2 (-16.3, -7.9)</td>
</tr>
<tr>
<td>Structured Model 1</td>
<td>(27.9, -1.81, 0.58)</td>
<td>-10.8 (-14.6, -7.0)</td>
</tr>
<tr>
<td>Structured Model 2</td>
<td>(27.9, -1.89, 0.54)</td>
<td>-11.3 (-15.3, -7.3)</td>
</tr>
</tbody>
</table>
CHAPTER 3
ESTIMATING LARGE CORRELATION MATRICES BY BANDING THE PARTIAL AUTOCORRELATION MATRIX

In this chapter, we propose a computationally efficient approach to estimate (large) p-dimensional correlation matrices of ordered data based on an independent sample of size \( n \). To do this, we construct the estimator based on a k-band partial autocorrelation matrix with the number of bands chosen using an exact multiple hypothesis testing procedure. This approach is considerably faster than many existing methods and only requires inversion of \( k \) dimensional covariance matrices. In addition, the resulting estimator is guaranteed to be positive definite as long as \( k \leq n - 2 \) (even when \( n < p \)). We evaluate our estimator via extensive simulations and compare it to the Ledoit-Wolf estimator. We also illustrate the approach using high-dimensional sonar data.

3.1 Background

Estimating a covariance matrix is essential in multivariate data analysis. Although the sample covariance matrix is an unbiased estimator of the covariance matrix of a Gaussian random vector, it has poor properties if the dimension (\( p \)) is large. In addition, covariance matrices are often sparse for large \( p \). Regularizing large sample covariance matrices has been proposed by using ridge regression (Warton (2008); Witten and Tibshirani (2009)) and discriminant analysis (Friedman (1989)). Recently, this area has seen an upsurge in practical and theoretical approaches due to a plethora of high dimensional data. Furrer and Bengtsson (2007) consider 'tapering' the sample covariance matrix by gradually shrinking the off-diagonal elements toward zeros. Johnstone and Lu (2007) consider a regularization of PCA, using a sparse basis and thresholding. Fan and Lv (2008) impose sparsity on the covariance matrix via a factor model.

Correlation matrices have special features under ordered (longitudinal) data, such as \( Y_j \) and \( Y_k \) being closer to independence or conditional independence as \( |j - k| \) increases. Wu and Pourahmadi (2003) exploit this feature by using the Cholesky
decomposition of the covariance matrix to perform what they called 'banding the inverse covariance matrix'. Bickel and Levina (2008b) achieve sparsity by banding the sample covariance matrix while Rothman et al. (2010) band the Cholesky factor of the covariance matrix. The former uses 'tapering' to achieve positive-definite (Furrer and Bengtsson (2007)). In addition, these methods use cross-validation to find the number of bands, which can be computationally intensive.

In this chapter, we present a banding method based on a partial autocorrelation matrix, which has several favorable properties including the estimator guaranteed to be positive-definiteness (w/o any adjustment), even for \( n < p \), and being based on exact, small sample results (not asymptotic). Computationally the estimator finds the number of bands sequentially, so if the estimate has \( k \) bands, only \( k \)-dimensional matrices need to be inverted (not \( p \)-dimensional, where \( p \) is often large).

This chapter is arranged as follows. In Section 3.2, we briefly review the partial autocorrelation matrix. In Section 3.3, we discuss a computational algorithm to estimate the non-zero elements in a banded partial autocorrelation matrix. After proving some necessary theorems, we propose a simple algorithm to estimate the number of bands, which relays on a multivariate hypothesis testing set-up, and exact small sample results. Sections 3.4 and 3.5 investigate the operating characteristics of our procedure via risk simulations and apply it to two real data examples, respectively.

### 3.2 Review of Partial Autocorrelations

We first review reparameterizing the correlation matrix \( R = (\rho_{j,k}) \) using the elements in the partial autocorrelation matrix \( \Pi = (\pi_{j,k}) \). In the partial autocorrelation matrix, \( \pi_{j,j} = 1 \) and for \( 1 \leq j < k \leq p \), \( \pi_{j,k} \) is the partial autocorrelation between \( Y_j \) and \( Y_k \) adjusted for the intervening variables, \( \{Y_{j+1}, \ldots, Y_{k-1}\} \).
There is a simple relationship between the elements of $R$ and $\Pi$. The partial autocorrelation $\pi_{j+1} = \rho_{j+1}$ and $\pi_{j+k}$, $(2 \leq k \leq p - 1)$ is

\[
\pi_{j+k} = \frac{\rho_{j+k} - \nu^T(j, k)R_2^{-1}(j, k)r(j, k)}{[1 - \nu^T(j, k)R_2^{-1}(j, k)\nu(j, k)][1 - r^T(j, k)R_2^{-1}(j, k)r(j, k)]^{1/2}},
\]

with $\nu(j, k) = (\rho_{j+1}, ..., \rho_{j+k-1})^T$ and $r(j, k) = (\rho_{j+k+1}, ..., \rho_{j+k+k-1})^T$. $R_2(j, k)$ is a matrix that contains the middle $k - 1$ rows and columns of $R[j : j + k]$, i.e.

\[
R[j : j + k] = \begin{pmatrix}
1 & \nu^T(j, k) & \rho_{j+k} \\
\nu(j, k) & R_2(j, k) & r(j, k) \\
\rho_{j+k} & r^T(j, k) & 1
\end{pmatrix}
\]

Correspondingly, $\rho_{j+k}$ can be written as a function of $\pi_{j+k}$,

\[
\rho_{j+k} = r_{jk} + \pi_{j+k}A_{jk},
\]

where

\[
r_{jk} = \nu^T(j, k)R_2^{-1}(j, k)r(j, k),
\]

and

\[
A_{jk}^2 = [1 - \nu^T(j, k)R_2^{-1}(j, k)\nu(j, k)][1 - r^T(j, k)R_2^{-1}(j, k)r(j, k)].
\]

The partial autocorrelation matrix will often be sparse (Dempster (1972); Friedman et al. (2008)) since the correlations in it are conditional (on the intervening variables) correlations. For example, an AR(1) correlation matrices corresponds to a partial autocorrelation matrix with only one non-zero band. More generally the partial autocorrelation will have $k$ non-zero bands under a k-th order ante-dependence model (Zimmerman and Nunez-Anton (2010); Gabriel (1962)).
3.3 Banding the Partial Autocorrelation Matrix

For a \( p \times p \) matrix \( M = [m_{ij}] \) and any \( k \in \{0, 1, 2, \ldots, p - 1\} \), we define a \( k \)-band matrix \( (B_k(M)) \) of \( M \) (Bickel and Levina (2008b)) as follows,

\[
B_k(M) = [m_{ij}I(|i - j| \leq k)].
\]

A diagonal matrix and full matrix are special cases of \( k \)-band matrices with \( k=0 \) and \( p-1 \), respectively. Here we will band the partial autocorrelation matrix, \( \Pi = (\pi_{jk}) \).

The computational attractiveness of our approach will rely on estimating the partial autocorrelations one band at a time and then for each band, doing a simple hypothesis test of whether to add another band. Our multiple hypothesis testing procedure is based on exact small sample results for \( n \geq p + 1 \), not asymptotic ones, which will be especially important when \( p \) is large relative to \( n \); for \( n \leq p \), the corresponding tests will be approximate. For our procedure, the largest matrix we will need to manipulate will be a \((k + 1)\)-dimensional matrix where \( k \) is the number of bands. In the following two subsections, we will provide details on first estimating the partial autocorrelations for each band and second, the procedure for hypothesis testing.

3.3.1 Estimating the Partial Autocorrelations for each Band

As stated in the previous section, we first estimate the partial autocorrelations for each band, then test whether to keep the band or stop. We will need several results to easily and efficiently estimate the partial autocorrelations in each band which we provide next. However, first, we introduce some notation.

In the following, we assume the data, \( \{Y_i : i = 1, \ldots, n\} \) are independent and identically normally distributed \( p \)-vector with mean 0 and covariance matrix \( \Sigma \). The MLE of \( \Sigma \) is

\[
S_\Sigma = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})(Y_i - \bar{Y})^T = \left( \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)^T \right).
\]

and the corresponding mle of the correlation matrix is
\[
\hat{R} = \left( \frac{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2}} \right)_{p \times p}.
\]

The \( jk^{th} \) sample correlation coefficient

\[
\hat{\rho}_{j,k} = \frac{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2}},
\]

is only related to \( j^{th} \) and \( k^{th} \) components of those normal random vectors. This implies that we can estimate correlation coefficients and variances one at a time.

Partial autocorrelation coefficients have one to one mapping to the correlation coefficients (Section 2). Furthermore, Eq. 3–1 shows \( \hat{\rho}_{j,k} \) is only related to the \( j^{th} \) through \( k^{th} \) rows and columns in \( \hat{R} \).

3.3.1.1 Statement of needed results

We briefly review and prove some useful results that will be needed in our approach.

**Result 1**: (Anderson (1984)) Let \( \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \) be a square matrix, where \( \Sigma_{11}, \Sigma_{22} \) are square submatrices with non-zero determinants. We denote \( \Sigma_{11\bullet 2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \) and \( \Sigma_{22\bullet 1} = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \).

Then

a)

\[
det(\Sigma) = det(\Sigma_{11})det(\Sigma_{22\bullet 1}).
\]

b)

\[
\Sigma^{-1} = \begin{pmatrix}
\Sigma_{11\bullet 2}^{-1} & -\Sigma_{11\bullet 2}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \\
-\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11\bullet 2}^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11\bullet 2}^{-1} \Sigma_{12} \Sigma_{22}^{-1}
\end{pmatrix},
\]

with similar results for \( \Sigma_{11\bullet 2} \) and \( \Sigma_{22} \).
**Result 2**: (Joe (2006)) For $-1 < \pi_{j,k} < 1, j, k \in \{1, 2, \ldots, p\}$ and $j \neq k$, the corresponding correlation matrix $R = (\rho_{j,k})$ is positive-definite.

**Result 3**: (Daniels and Pourahmadi (2009)) For $-1 < \pi_{j,k} < 1$, for $j, k \in \{1, 2, \ldots, p\}$, $j \neq k$, and $R$ the corresponding correlation matrix constructed from the elements of $\Pi$, we have

$$
\det(R) = \prod_{t=2}^{p} \prod_{j=1}^{t-1} (1 - \pi_{j,t}^2).
$$

**Result 4**: Inverting the correlation matrix constructed from a k-band partial autocorrelation matrix only requires inversion of $(k + 1)$-dimensional matrices, and its precision matrix is also a k-band matrix.

*Proof* in the appendix.

**Result 5**: For a k-band partial autocorrelation matrix of a multivariate normal random vector, the exponential part of the multivariate normal density is only a function of partial autocorrelations of lags $\{1, \ldots, k\}$.

*Proof* in the appendix.

### 3.3.1.2 Results for computing the mle for each band

The key idea is that we can compute the mle of a k-band partial autocorrelation matrix one band at a time. This will be much more efficient than having to manipulate the entire p-dimensional matrix. We can do this since the mle of the partial autocorrelations in band $j$ only depends on the mle of the partial autocorrelations in the bands $< j$, not those $> j$.

The next result presents an expression for the determinant of the correlation matrix as a function of the partitioned matrix in Eq. 3–2.

**Lemma 3.0.1.** Let $R_{p \times p} = (\rho_{j,k})$ be a correlation matrix, and $\Pi_{p \times p} = (\pi_{j,k})$ be the corresponding partial autocorrelation matrix of $R_{p \times p}$. Then,

$$
det(R_{p \times p}) = \left[ det(R_2)(1 - \nu_2^T R_2^{-1} \nu_2)(1 - r_2^T R_2^{-1} r_2) \right] \cdot (1 - \pi_{1,p}^2).
$$
Proof in appendix.

We introduce a sequential procedure to estimate the partial autocorrelations in each band and show that is equivalent to the mle \((\hat{\Pi})\) of full multivariate normal likelihood \(L(\Pi|y_1, \ldots, y_n)\). Let \(f(y_{ij}, \ldots, y_{ij+l}|\Pi_{j,j+l})\) be the marginal pdf of the random sub-vector \(Y_{j,j+l} = (Y_{ij}, \ldots, Y_{ij+l})\) of a multivariate normal random vector \(Y_i\), and let \(G(\pi_{j,j+l}) = L(\Pi_{j,j+l}|y_{ij}, \ldots, y_{ij+n})\) be the corresponding likelihood function.
Define \(e_{\pi} = (e_{\pi_1}, e_{\pi_2}, \ldots, e_{\pi_{p-1}})\) as the maximizer of the following objective function,
\[
G^{*}(\pi_l) = \prod_{j=1}^{p-l} G(\pi_{j,j+l}), \text{ and } \pi_l = (\pi_{1,1+l}, \ldots, \pi_{p-l,p}), \text{ for } l = 1, \ldots, p-1 \quad (3-6)
\]
Since \(\pi_{j,k}, j, k \in \{1, 2, \ldots, p\} \text{ and } k \neq j\), independently varys in the interval \((-1, 1)\), we can estimate \(\pi_l\) sequentially by maximizing the objective functions in (Eq. 3–6) for \(l = 1, \ldots, p-1\), i.e.
\[
\tilde{\pi}_l = \max_{\pi_l} G^{*}(\pi_l), \text{ for } l = 1, 2, \ldots, p-1 \quad (3-7)
\]
In the following, we prove the maximizer of the objective function for each \(\pi_l\) is equivalent to the mle of \(\Pi\) based on the multivariate normal likelihood.

**Theorem 3.1.** The MLE \(\hat{\pi}\) of the partial autocorrelation coefficients
\[
\{\hat{\pi}_{1,2}, \hat{\pi}_{2,3}, \ldots, \hat{\pi}_{p-1,p}, \hat{\pi}_{1,3}, \ldots, \hat{\pi}_{p-2,p}, \ldots, \hat{\pi}_{1,p}\}
\]
based on the multivariate normal likelihood function \(L(\Pi|y_1, y_2, \ldots, y_n)\) is same as \(\tilde{\pi}\),
\[
\{\tilde{\pi}_{1,2}, \tilde{\pi}_{2,3}, \ldots, \tilde{\pi}_{p-1,p}, \tilde{\pi}_{1,3}, \ldots, \tilde{\pi}_{p-2,p}, \ldots, \tilde{\pi}_{1,p}\} \text{ obtained from Eq. 3–7.}
\]
Proof in appendix.

Combining with Eq. 3–1, this means the estimated lag k partial autocorrelation coefficients are invariant to the estimated partial autocorrelation coefficients with lag greater than k. We have following corollary immediately.

**Corollary 3.1.1.** Let \(\hat{\Pi}\) be the MLE of partial autocorrelation based on the multivariate normal likelihood function \(L(\Sigma|y_1, y_2, \ldots, y_n)\) and \(\tilde{\Pi}\) be as in Eq. 3–7. Then the mle \((\hat{\Pi})\) of
a $k$-band partial autocorrelation matrix is equivalent to the corresponding $\tilde{\Pi}$, based on a $k$-band matrix.

3.3.2 Choosing the Band

3.3.2.1 Theorems

To estimate the number of bands ($k$) of a partial correlation matrix $\Pi = (\pi_{j,k})$, our strategy will be to sequentially test the null hypothesis that each band is zero starting from the first band. Implicitly, if the $j$th band is zero, the subsequent bands, $j + 1, \ldots, p - 1$ are zero as well. In general, for $\pi_k = (\pi_{1,1+k}, \ldots, \pi_{p-k,k})$, we construct multiple tests under the following hypotheses:

$$H_0 : \pi_l = 0, \ l \geq k \ \text{vs} \ H_\alpha : \pi_k \neq 0.$$  \hfill (3–8)

We choose the band as the first $k$ for which $H_0$ can not be rejected. Note we just need to test the partial autocorrelations in the $k$th band under the assumption of a true band $k-1$ matrix. For computational reasons, we do not explicitly include in the test for band $k$ the partial autocorrelations in bands $> k$; we assess the impact of this on operating characteristics of the test in Section 4. The lemma below provides the key result for the theorem that gives the result needed to efficiently conduct these hypothesis tests.

**Lemma 3.1.1.** Suppose $Y_1, Y_2, \ldots, Y_n \ (n > p)$ are iid $N(0, DRD)$, where $D$ is a diagonal matrix of marginal standard deviations. Let $\hat{\Pi} = (\hat{\pi}_{j,t})_{p \times p}$ be the mle of the partial autocorrelation matrix. For a band $k_0$ partial autocorrelation matrix, the mle’s of the partial autocorrelations with lags greater than $k_0$ are independent with marginal distributions given by

$$f(\hat{\pi}_{j,j+k}) \propto (1 - \hat{\pi}_{j,j+k})^\alpha (1 + \hat{\pi}_{j,j+k})^\beta$$

where

$$\alpha = \beta = \begin{cases} 
  \frac{n-k-2}{4} & \text{for } k_0 < k \leq p - 2 \\
  \frac{n-p-1}{2} & \text{for } k_0 < k = p - 1 
\end{cases}.$$
Proof in the appendix.

The above lemma is the key result for the proof of the following theorem.

**Theorem 3.2.** Suppose $Y_1, Y_2, \ldots, Y_n$ are iid $N(0, DRD)$ and $\Pi = (\pi_{j,t})_{p \times p}$ is the partial correlation matrix of R. Then, for $n \geq p + 1$, under hypotheses in Eq. 3–8, the mle's of $\pi_{j,j+l}$, denoted as $\hat{\pi}_{j,j+l}$, $l = 1, \ldots, p - 1$ follow independent transformed Beta distributions on $(-1, 1)$ with parameters

$$\alpha = \beta = \begin{cases} \frac{n-l-2}{4} & \text{if } 1 \leq k \leq l \leq p - 2, \\ \frac{n-p-1}{2} & \text{if } k \leq l = p - 1. \end{cases}$$

**Proof.** It is a direct result of Lemma 3.1.1. \qed

For $n \leq p$, the lag $k$ ($k > k_0$) sample partial autocorrelations are correlated even if the true band is $k_0$, but the correlations among them are very small (from empirical checks). Therefore, we still approximate the distribution of the $k$-band estimated partial autocorrelations with the independent Beta distributions given in Theorem 3.2.

Since the partial autocorrelations are independent Beta distributions, we adjust for multiple testing using a Bonferroni correction. We explore this in the simulations in Section 3.4 for both $n > p$ and $n \leq p$. Before that though, we will outline the exact procedure in the next subsection.

### 3.3.2.2 Overall procedure

Here are the steps we use to estimate the partial autocorrelation matrix.

1. Starting with band $j$,
   1. Compute the mle of the $\pi$'s in band $j$ using the results in Section 3.1.2.
   2. Test the null hypothesis that all the partial autocorrelations in the $j$th band are zero using the multivariate hypotheses testing procedure outlined above.
   3. If reject, repeat the first two steps for band $j + 1$; otherwise, stop and the number of bands is $j - 1$. 

61
Practically, we do hypothesis testing starting from band 1 and stop at the first $j$ ($j \leq \min(n - 2, p - 1)$) for which $H_0$ can not be rejected. Therefore this procedure only relates to $(j+1)$-dimensional principal submatrices of $R$ and only requires a sample size $n \geq j + 2$.

3.4 Simulations

3.4.1 Models

To evaluate our banding method in practice, we conduct several simulations using Bonferroni criteria at different levels of $\alpha$.

We consider several true matrices which we describe next:

- Two sets of matrices based on banding the estimated partial autocorrelation matrices of the Metal and Rock data (available at http://www.ics.uci.edu/~melearn/MLRepository.html); for more details on the data, see Section 3.5. In particular, we consider 4, 9, and 14 bands.

- $AR(1)$ correlation matrices for several values of the lag 1 correlation.

- Four band partial autocorrelation matrix: $\pi_{j,j+l} = 0.4I(|l| = 1) + 0.2I(2 \leq |l| \leq 3) + 0.1I(|l| = 4)$.

For each true matrix, we estimate the band $k$ by averaging the estimated bands over 100 replicated data sets generated from a multivariate normal distribution with mean zero and variance-covariance matrix $\Sigma$ be the corresponding correlation matrix for each case above. We estimate the partial autocorrelation matrix by averaging these 100 estimated banded partial autocorrelation matrices corresponding to each replicated data set and then constructing the estimated correlation matrix, which, in general, is not banded. Recall, for a $k$-band partial autocorrelation matrix, we only need to invert at most $k$-dimensional matrices to construct the corresponding "inverse correlation matrix" instead.

3.4.2 Results

We compare our estimator to the sample correlation matrix and the Ledoit-Wolf shrinkage estimator (Ledoit and Wolf (2003)) of the covariance matrix (shrinkage target is a diagonal matrix). The Ledoit-Wolf (L-W) estimator is computationally tractable like
our estimator but does not exploit the decaying correlation that we do with our banded estimator. And to make an appropriate comparison, we compute the L-W estimator of the covariance matrix and then use as a comparison, the resulting correlation matrix.

We use the procedure in Section 3.2.1 with the Bonferroni (Bonf) criteria at different levels of $\alpha \in \{0.05, 0.2, 0.3, 0.5\}$. To compare our estimator to the Ledoit-Wolf and the sample correlation matrix, we use squared error loss of the estimated correlation matrix.

Tables 3-1 to 3-3 contain the simulation results. The risk results indicate our approach does much better than the sample correlation matrix (which is not surprising). The top part of Table 3-1 contains simulation results based on banding the estimated partial autocorrelation matrix of the metal data with 4, 9, and 14 bands as the true matrices. We first focus on $n > p$. The choice of $\alpha = 0.05$ does very well for the larger sample sizes, $n > 100$, but severely underestimates for the smaller sample sizes. Choices of $\alpha = 0.2, 0.3$ do considerably better for the smaller sample sizes with less than 1 band of overestimation for the larger samples. The bottom part of Table 3-1 shows corresponding results for the rock data with the same behavior as the metal data.

Table 3-2 contains results based on an AR(1) correlation matrix for different choices of the lag 1 correlation (0, 0.2, 0.5, 0.9). We see similar behavior as the metal and rock data in that we have increased power for the smallest sample sizes for $\alpha = 0.2, 0.3$ over $\alpha = 0.05$ and less overestimation than $\alpha = 0.5$. Finally Table 3-3 contains the results of a true four band partial autocorrelation matrix, that was used in Rothman et al. (2010) for covariance matrix. By $n = 60$, the estimator does very well for $\alpha = 0.2, 0.3$.

For $n \leq p$, we continue to see substantial risk improvements, but a consistent underestimation of the bands. This is due to low power in small sample sizes. The best performance appears to be based on using a Bonferroni correction with $\alpha \geq 0.30$.

We also compare our estimators to Ledoit and Wolf (2003) shrinkage estimator, which shrinks the sample covariance matrix toward a diagonal matrix. Risks of our estimators are much smaller than those of Ledoit-Wolf (L-W) estimator. The
performance versus the L-W estimator is not surprising since the L-W estimator is not designed for our setting where the true covariance matrices here are sparse.

3.4.3 Choice of $\alpha$ for Bonferroni Correction

Simulation results demonstrated that different choices of $\alpha$ give quite different power for the smaller sample sizes with minimal overestimation of the bands for the larger sample sizes (at most overestimate by 1 or 2 bands on average) as seen in Tables 3-1 to 3-3. To optimize the power and minimize the overestimation for larger sample sizes, we recommend $\alpha$ in the range of 0.2 to 0.3 for $n > p$. However, we recommend the more conservative $\alpha$ in the range of 0.3 to 0.5 for the case of $n \leq p$ when the sample partial autocorrelations in each band are no longer independent.

3.5 Applications to Sonar Data

We illustrate our approach on two data sets, the Metal and Rock data of the sonar data, which is available at http://www.ics.uci.edu/~mlearn/MLRepository.html. This data set contains 111 signals from a Metal cylinder and 97 signals from a Rock, where each signal has 60 frequency energy measurements ranging from 0.0 to 1.0. These signals were measured at different angles for the same objects. We assume the signals are iid normal random vectors. Images of absolute sample correlation matrices of the Metal and Rock data in Figure 3-1 (hot to cool corresponds to 1.0 to 0.0) show a general pattern of decaying correlations with increasing lag, which motivates the banded estimator here. Since the sample sizes are not large relative to the dimension of matrices ($p=60$), based on our simulations we use estimators based on $\alpha = 0.2$ or 0.3. The resulting estimator for the rock data had four bands and the metal data had eleven bands (Table 3-4). The images of estimated correlation matrices based on the banded partial autocorrelation matrix estimators are shown in Figure 3-1. The top part of Figure 3-1 for the metal data shows that most of the marginal correlations up to lag 37 are captured quite well and the bottom part of Figure 3-1 for the Rock data show similar results up to lag 17. This (to some extent) agrees with previous analysis of the
data in Rothman et al. (2010) (they found bands of 37 and 17 for metal and rock data, respectively, based on banding the sample covariance matrix and the Cholesky factor). However, our estimator has considerably fewer parameters to estimate (4 vs. 17 bands for the rock data and 11 vs. 37 for the metal data) and is computationally much quicker, with less than one second for each $\alpha$ with Matlab R2006a.

3.6 Discussions

We have proposed a k-banded estimator for a correlation matrix that exists when $n \leq p$ and only requires inversion of at most $k$-dimensional matrices. The algorithm for the estimator relies on exact distributional results under the null hypothesis for $n > p$. The estimator can be computed very quickly. We recommend a Bonferroni correction with $\alpha \in [0.2, 0.3]$ for $n > p$ and $\alpha \in [0.3, 0.5]$ for $n \leq p$. Related banded estimators based on the sample covariance matrices need adjustments to ensure positive definiteness and banded estimators for the sample covariance and its Cholesky factor do not provide exact distributional results and require computationally intensive cross-validation procedures to obtain the number of bands (Bickel and Levina (2008b); Rothman et al. (2010)). In addition, these estimators require $n > p$.

Current extensions of this work include determining the rate of convergence of our estimator and the addition of smoothing the partial autocorrelations within bands (similar to previous work by Wu and Pourahmadi (2003) using the GARPs of the modified Cholesky decomposition of a covariance matrix) to better estimate the partial autocorrelations in each non-zero band.
Table 3-1. Risk simulations based on metal and rock data

<table>
<thead>
<tr>
<th>n</th>
<th>Band</th>
<th>Risk</th>
<th>$R_L$</th>
<th>W</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.2$</th>
<th>$\alpha = 0.3$</th>
<th>$\alpha = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>359</td>
<td>126</td>
<td>1.4</td>
<td>63.9</td>
<td>2.2</td>
<td>61.4</td>
<td>2.6</td>
<td>64.1</td>
</tr>
<tr>
<td>9</td>
<td>354</td>
<td>157</td>
<td>1.6</td>
<td>132</td>
<td>2.1</td>
<td>117</td>
<td>2.5</td>
<td>117</td>
</tr>
<tr>
<td>14</td>
<td>355</td>
<td>167</td>
<td>1.5</td>
<td>166</td>
<td>2.1</td>
<td>152</td>
<td>2.6</td>
<td>149</td>
</tr>
<tr>
<td>20</td>
<td>173</td>
<td>91.1</td>
<td>2.7</td>
<td>32.3</td>
<td>3.3</td>
<td>33.0</td>
<td>3.5</td>
<td>33.6</td>
</tr>
<tr>
<td>9</td>
<td>164</td>
<td>105</td>
<td>2.6</td>
<td>87.0</td>
<td>3.4</td>
<td>80.6</td>
<td>3.9</td>
<td>76.6</td>
</tr>
<tr>
<td>14</td>
<td>162</td>
<td>105</td>
<td>2.7</td>
<td>116</td>
<td>3.4</td>
<td>109</td>
<td>3.9</td>
<td>106</td>
</tr>
<tr>
<td>30</td>
<td>116</td>
<td>80.1</td>
<td>3.5</td>
<td>69.7</td>
<td>5.2</td>
<td>58.7</td>
<td>6.2</td>
<td>54.9</td>
</tr>
<tr>
<td>14</td>
<td>105</td>
<td>82.6</td>
<td>3.7</td>
<td>95.3</td>
<td>5.9</td>
<td>78.5</td>
<td>6.7</td>
<td>75.3</td>
</tr>
<tr>
<td>40</td>
<td>84.2</td>
<td>58.0</td>
<td>3.6</td>
<td>16.3</td>
<td>4.0</td>
<td>16.4</td>
<td>4.1</td>
<td>16.6</td>
</tr>
<tr>
<td>9</td>
<td>81.3</td>
<td>64.2</td>
<td>4.6</td>
<td>54.3</td>
<td>6.8</td>
<td>43.7</td>
<td>7.8</td>
<td>40.3</td>
</tr>
<tr>
<td>14</td>
<td>80.6</td>
<td>64.8</td>
<td>5.0</td>
<td>80.4</td>
<td>7.7</td>
<td>61.5</td>
<td>9.4</td>
<td>57.3</td>
</tr>
<tr>
<td>50</td>
<td>68.0</td>
<td>50.2</td>
<td>3.8</td>
<td>12.6</td>
<td>4.2</td>
<td>12.7</td>
<td>4.2</td>
<td>12.9</td>
</tr>
<tr>
<td>9</td>
<td>64.5</td>
<td>52.4</td>
<td>5.3</td>
<td>45.4</td>
<td>7.9</td>
<td>32.6</td>
<td>8.6</td>
<td>30.3</td>
</tr>
<tr>
<td>14</td>
<td>63.2</td>
<td>53.8</td>
<td>5.8</td>
<td>69.1</td>
<td>9.2</td>
<td>52.9</td>
<td>10.9</td>
<td>48.2</td>
</tr>
<tr>
<td>60</td>
<td>53.7</td>
<td>42.7</td>
<td>4.0</td>
<td>10.2</td>
<td>4.2</td>
<td>10.4</td>
<td>4.3</td>
<td>10.6</td>
</tr>
<tr>
<td>9</td>
<td>53.2</td>
<td>44.8</td>
<td>6.7</td>
<td>33.1</td>
<td>8.8</td>
<td>25.3</td>
<td>9.1</td>
<td>25.0</td>
</tr>
<tr>
<td>14</td>
<td>48.3</td>
<td>40.3</td>
<td>7.5</td>
<td>52.9</td>
<td>11.0</td>
<td>39.3</td>
<td>12.4</td>
<td>35.8</td>
</tr>
<tr>
<td>100</td>
<td>28.0</td>
<td>25.1</td>
<td>8.7</td>
<td>14.4</td>
<td>9.2</td>
<td>14.0</td>
<td>9.4</td>
<td>14.0</td>
</tr>
<tr>
<td>9</td>
<td>28.0</td>
<td>25.1</td>
<td>8.7</td>
<td>14.4</td>
<td>9.2</td>
<td>14.0</td>
<td>9.4</td>
<td>14.0</td>
</tr>
<tr>
<td>14</td>
<td>31.1</td>
<td>29.2</td>
<td>12.2</td>
<td>21.8</td>
<td>13.9</td>
<td>20.2</td>
<td>14.2</td>
<td>20.0</td>
</tr>
<tr>
<td>300</td>
<td>28.0</td>
<td>25.1</td>
<td>8.7</td>
<td>14.4</td>
<td>9.2</td>
<td>14.0</td>
<td>9.4</td>
<td>14.0</td>
</tr>
<tr>
<td>500</td>
<td>6.93</td>
<td>6.75</td>
<td>4.1</td>
<td>1.20</td>
<td>4.2</td>
<td>1.24</td>
<td>4.3</td>
<td>1.26</td>
</tr>
<tr>
<td>9</td>
<td>6.62</td>
<td>6.36</td>
<td>9.0</td>
<td>2.76</td>
<td>9.3</td>
<td>2.83</td>
<td>9.4</td>
<td>2.85</td>
</tr>
<tr>
<td>14</td>
<td>6.10</td>
<td>6.04</td>
<td>14.0</td>
<td>3.68</td>
<td>14.0</td>
<td>3.71</td>
<td>14.3</td>
<td>3.73</td>
</tr>
<tr>
<td>1000</td>
<td>3.55</td>
<td>3.47</td>
<td>4.0</td>
<td>0.58</td>
<td>4.2</td>
<td>0.60</td>
<td>4.3</td>
<td>0.61</td>
</tr>
<tr>
<td>9</td>
<td>2.94</td>
<td>2.89</td>
<td>9.0</td>
<td>1.33</td>
<td>9.2</td>
<td>1.31</td>
<td>9.3</td>
<td>1.32</td>
</tr>
<tr>
<td>14</td>
<td>2.99</td>
<td>2.95</td>
<td>14.0</td>
<td>1.87</td>
<td>14.2</td>
<td>1.88</td>
<td>14.4</td>
<td>1.89</td>
</tr>
</tbody>
</table>
Table 3-1. Continued

<table>
<thead>
<tr>
<th>n</th>
<th>Band</th>
<th>Risk</th>
<th>( R_L )</th>
<th>( W )</th>
<th>( \alpha = 0.05 )</th>
<th>( \alpha = 0.2 )</th>
<th>( \alpha = 0.3 )</th>
<th>( \alpha = 0.5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>364</td>
<td>139</td>
<td>1.1</td>
<td>46.9</td>
<td>1.8</td>
<td>58.3</td>
<td>2.0</td>
<td>61.8</td>
</tr>
<tr>
<td>10</td>
<td>354</td>
<td>145</td>
<td>1.2</td>
<td>74.6</td>
<td>1.9</td>
<td>78.9</td>
<td>2.3</td>
<td>84.4</td>
</tr>
<tr>
<td>14</td>
<td>349</td>
<td>161</td>
<td>1.2</td>
<td>116</td>
<td>1.8</td>
<td>118</td>
<td>2.2</td>
<td>117</td>
</tr>
<tr>
<td>4</td>
<td>171</td>
<td>93.7</td>
<td>2.3</td>
<td>33.1</td>
<td>3.2</td>
<td>32.3</td>
<td>3.5</td>
<td>32.5</td>
</tr>
<tr>
<td>20</td>
<td>166</td>
<td>98.7</td>
<td>2.1</td>
<td>60.2</td>
<td>3.1</td>
<td>54.2</td>
<td>3.5</td>
<td>52.2</td>
</tr>
<tr>
<td>14</td>
<td>167</td>
<td>109</td>
<td>2.1</td>
<td>99.8</td>
<td>3.3</td>
<td>90.2</td>
<td>3.8</td>
<td>87.0</td>
</tr>
<tr>
<td>4</td>
<td>114</td>
<td>72.6</td>
<td>3.2</td>
<td>22.7</td>
<td>3.7</td>
<td>21.5</td>
<td>4.0</td>
<td>22.2</td>
</tr>
<tr>
<td>30</td>
<td>109</td>
<td>80.9</td>
<td>3.2</td>
<td>44.4</td>
<td>4.2</td>
<td>38.3</td>
<td>4.7</td>
<td>36.9</td>
</tr>
<tr>
<td>14</td>
<td>109</td>
<td>81.9</td>
<td>3.1</td>
<td>81.6</td>
<td>3.9</td>
<td>73.8</td>
<td>4.5</td>
<td>69.5</td>
</tr>
<tr>
<td>4</td>
<td>82.0</td>
<td>59.1</td>
<td>3.6</td>
<td>15.2</td>
<td>3.9</td>
<td>15.1</td>
<td>4.2</td>
<td>15.9</td>
</tr>
<tr>
<td>40</td>
<td>82.4</td>
<td>61.1</td>
<td>3.8</td>
<td>34.9</td>
<td>5.3</td>
<td>31.9</td>
<td>5.9</td>
<td>32.1</td>
</tr>
<tr>
<td>14</td>
<td>77.3</td>
<td>61.9</td>
<td>3.6</td>
<td>70.7</td>
<td>4.8</td>
<td>61.8</td>
<td>5.8</td>
<td>57.9</td>
</tr>
<tr>
<td>4</td>
<td>68.2</td>
<td>51.1</td>
<td>3.7</td>
<td>13.0</td>
<td>4.1</td>
<td>13.3</td>
<td>4.3</td>
<td>14.1</td>
</tr>
<tr>
<td>50</td>
<td>65.4</td>
<td>53.1</td>
<td>3.8</td>
<td>33.4</td>
<td>5.7</td>
<td>28.8</td>
<td>6.8</td>
<td>28.0</td>
</tr>
<tr>
<td>14</td>
<td>63.6</td>
<td>54.2</td>
<td>4.1</td>
<td>65.5</td>
<td>6.3</td>
<td>53.6</td>
<td>7.8</td>
<td>49.8</td>
</tr>
<tr>
<td>60</td>
<td>55.4</td>
<td>42.6</td>
<td>3.7</td>
<td>10.5</td>
<td>4.2</td>
<td>10.1</td>
<td>4.4</td>
<td>10.5</td>
</tr>
<tr>
<td>9</td>
<td>49.9</td>
<td>41.8</td>
<td>4.6</td>
<td>26.9</td>
<td>6.5</td>
<td>23.4</td>
<td>7.9</td>
<td>22.8</td>
</tr>
<tr>
<td>14</td>
<td>52.0</td>
<td>43.1</td>
<td>4.7</td>
<td>58.3</td>
<td>8.3</td>
<td>43.9</td>
<td>10.4</td>
<td>39.0</td>
</tr>
<tr>
<td>4</td>
<td>34.9</td>
<td>29.7</td>
<td>4.1</td>
<td>5.90</td>
<td>4.3</td>
<td>6.04</td>
<td>4.3</td>
<td>6.09</td>
</tr>
<tr>
<td>100</td>
<td>33.4</td>
<td>31.4</td>
<td>4.8</td>
<td>16.4</td>
<td>8.8</td>
<td>13.8</td>
<td>9.1</td>
<td>13.6</td>
</tr>
<tr>
<td>14</td>
<td>30.3</td>
<td>27.5</td>
<td>8.5</td>
<td>35.5</td>
<td>13.0</td>
<td>21.7</td>
<td>13.6</td>
<td>20.4</td>
</tr>
<tr>
<td>4</td>
<td>10.8</td>
<td>10.2</td>
<td>4.1</td>
<td>19.2</td>
<td>4.2</td>
<td>19.9</td>
<td>4.3</td>
<td>20.2</td>
</tr>
<tr>
<td>14</td>
<td>10.5</td>
<td>9.67</td>
<td>14.0</td>
<td>6.09</td>
<td>14.2</td>
<td>6.17</td>
<td>14.3</td>
<td>6.18</td>
</tr>
<tr>
<td>4</td>
<td>6.17</td>
<td>6.03</td>
<td>4.1</td>
<td>1.17</td>
<td>4.2</td>
<td>1.19</td>
<td>4.4</td>
<td>1.22</td>
</tr>
<tr>
<td>500</td>
<td>6.89</td>
<td>6.87</td>
<td>9.1</td>
<td>2.37</td>
<td>9.2</td>
<td>2.41</td>
<td>9.3</td>
<td>2.43</td>
</tr>
<tr>
<td>14</td>
<td>5.78</td>
<td>5.60</td>
<td>14.1</td>
<td>3.53</td>
<td>14.2</td>
<td>3.54</td>
<td>14.3</td>
<td>3.56</td>
</tr>
<tr>
<td>4</td>
<td>3.43</td>
<td>3.33</td>
<td>4.1</td>
<td>0.59</td>
<td>4.3</td>
<td>0.61</td>
<td>4.4</td>
<td>0.62</td>
</tr>
<tr>
<td>1000</td>
<td>2.94</td>
<td>2.97</td>
<td>9.1</td>
<td>1.16</td>
<td>9.2</td>
<td>1.17</td>
<td>9.3</td>
<td>1.19</td>
</tr>
<tr>
<td>14</td>
<td>3.33</td>
<td>3.32</td>
<td>14.1</td>
<td>1.77</td>
<td>14.2</td>
<td>1.78</td>
<td>14.4</td>
<td>1.79</td>
</tr>
</tbody>
</table>
Figure 3-1. Images of sample correlation and estimated correlation matrices
Table 3-2. Risk simulation based on AR(1) correlation matrices

<table>
<thead>
<tr>
<th>n</th>
<th>ρ</th>
<th>R risk</th>
<th>L-W</th>
<th>α = 0.05 EB risk</th>
<th>α = 0.2 EB risk</th>
<th>α = 0.3 EB risk</th>
<th>α = 0.5 EB risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>394</td>
<td>16.6</td>
<td>0.0 0.35 0.1 1.11</td>
<td>0.1 1.76 0.3 4.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.2</td>
<td>390</td>
<td>19.9</td>
<td>0.1 5.44 0.2 7.63</td>
<td>0.4 9.36 0.6 12.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>385</td>
<td>49.3</td>
<td>0.4 29.8 0.9 19.8</td>
<td>1.0 18.7 1.2 19.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>326</td>
<td>207</td>
<td>1.0 34.9 1.1 35.5</td>
<td>1.1 35.5 1.3 39.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>186</td>
<td>2.10</td>
<td>0.1 0.53 0.2 1.30</td>
<td>0.3 1.87 0.5 3.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>185</td>
<td>6.57</td>
<td>0.2 5.42 0.7 6.68</td>
<td>0.8 7.06 1.1 8.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>184</td>
<td>34.5</td>
<td>1.0 7.54 1.1 8.38</td>
<td>1.2 9.22 1.5 10.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>152</td>
<td>122</td>
<td>1.0 15.5 1.1 15.7</td>
<td>1.2 16.2 1.4 17.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>122</td>
<td>0.60</td>
<td>0.0 0.20 0.2 0.82</td>
<td>0.2 1.17 0.5 2.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>122</td>
<td>5.37</td>
<td>0.3 4.74 0.7 4.88</td>
<td>1.0 5.58 1.4 6.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>120</td>
<td>30.3</td>
<td>1.1 4.81 1.2 5.34</td>
<td>1.2 5.63 1.4 6.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>98.3</td>
<td>86.5</td>
<td>1.0 10.1 1.1 10.1</td>
<td>1.2 10.6 1.4 11.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0</td>
<td>90.7</td>
<td>0.31</td>
<td>0.0 0.14 0.2 0.67</td>
<td>0.4 1.16 0.6 1.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>90.7</td>
<td>5.01</td>
<td>0.6 4.01 1.1 3.80</td>
<td>1.3 3.99 1.5 4.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>89.4</td>
<td>27.7</td>
<td>1.1 3.52 1.2 4.03</td>
<td>1.3 4.43 1.5 5.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>71.9</td>
<td>65.3</td>
<td>1.1 7.64 1.3 8.09</td>
<td>1.4 8.29 1.7 9.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0</td>
<td>72.1</td>
<td>0.18</td>
<td>0 0 0.1 0.381 0.3 0.659 0.5 1.30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>72.0</td>
<td>4.77</td>
<td>0.8 3.08 1.2 3.18</td>
<td>1.3 3.44 1.7 4.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>72.1</td>
<td>25.7</td>
<td>1.1 2.82 1.3 3.51</td>
<td>1.5 3.96 1.8 4.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>58.0</td>
<td>53.5</td>
<td>1.0 5.76 1.2 6.08</td>
<td>1.3 6.33 1.7 7.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>0</td>
<td>59.8</td>
<td>0.12</td>
<td>0.1 0.18 0.3 0.63</td>
<td>0.4 0.80 0.7 1.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>60.1</td>
<td>4.67</td>
<td>0.8 2.69 1.2 2.54</td>
<td>1.4 2.87 1.8 3.58</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>58.6</td>
<td>23.7</td>
<td>1.1 2.37 1.3 2.85</td>
<td>1.3 3.03 1.5 3.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>49.1</td>
<td>46.2</td>
<td>1.1 5.21 1.2 5.58</td>
<td>1.3 6.03 1.6 6.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>35.8</td>
<td>0.051</td>
<td>0.1 0.10 0.3 0.36</td>
<td>0.3 0.46 0.7 0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>35.7</td>
<td>4.38</td>
<td>1.0 1.32 1.2 1.52</td>
<td>1.4 1.72 1.7 2.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>35.0</td>
<td>18.6</td>
<td>1.1 1.35 1.2 1.56</td>
<td>1.3 1.65 1.5 1.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>27.9</td>
<td>27.4</td>
<td>1.0 3.10 1.3 3.37</td>
<td>1.4 3.51 1.6 3.92</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3-2. Continued

<table>
<thead>
<tr>
<th>n</th>
<th>( \rho )</th>
<th>R risk</th>
<th>L-W risk</th>
<th>( \alpha = 0.05 ) EB risk</th>
<th>( \alpha = 0.2 ) EB risk</th>
<th>( \alpha = 0.3 ) EB risk</th>
<th>( \alpha = 0.5 ) EB risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>17.7</td>
<td>0.01</td>
<td>0.1</td>
<td>0.03</td>
<td>0.13</td>
<td>0.26</td>
<td>0.7</td>
</tr>
<tr>
<td>200</td>
<td>0.2</td>
<td>17.8</td>
<td>3.87</td>
<td>1.0</td>
<td>0.62</td>
<td>1.2</td>
<td>0.71</td>
</tr>
<tr>
<td>0.5</td>
<td>17.5</td>
<td>12.1</td>
<td>1.0</td>
<td>0.63</td>
<td>1.1</td>
<td>0.70</td>
<td>1.3</td>
</tr>
<tr>
<td>0.9</td>
<td>14.4</td>
<td>13.9</td>
<td>1.1</td>
<td>1.57</td>
<td>1.3</td>
<td>1.66</td>
<td>1.4</td>
</tr>
<tr>
<td>500</td>
<td>0.2</td>
<td>7.13</td>
<td>0.01</td>
<td>0.1</td>
<td>0.02</td>
<td>0.4</td>
<td>0.10</td>
</tr>
<tr>
<td>0.5</td>
<td>6.99</td>
<td>5.97</td>
<td>1.1</td>
<td>0.28</td>
<td>1.2</td>
<td>0.32</td>
<td>1.3</td>
</tr>
<tr>
<td>0.9</td>
<td>5.61</td>
<td>5.58</td>
<td>1.1</td>
<td>0.62</td>
<td>1.3</td>
<td>0.65</td>
<td>1.4</td>
</tr>
<tr>
<td>1000</td>
<td>0.2</td>
<td>3.53</td>
<td>0.00</td>
<td>0.1</td>
<td>0.01</td>
<td>0.2</td>
<td>0.03</td>
</tr>
<tr>
<td>0.5</td>
<td>3.47</td>
<td>3.19</td>
<td>1.0</td>
<td>0.13</td>
<td>1.2</td>
<td>0.15</td>
<td>1.3</td>
</tr>
<tr>
<td>0.9</td>
<td>2.84</td>
<td>2.83</td>
<td>1.1</td>
<td>0.28</td>
<td>1.2</td>
<td>0.29</td>
<td>1.3</td>
</tr>
</tbody>
</table>

Table 3-3. Risk simulations of 4 band partial autocorrelation matrices

<table>
<thead>
<tr>
<th>n</th>
<th>R risk</th>
<th>L-W risk</th>
<th>( \alpha = 0.05 ) EB risk</th>
<th>( \alpha = 0.2 ) EB risk</th>
<th>( \alpha = 0.3 ) EB risk</th>
<th>( \alpha = 0.5 ) EB risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>378</td>
<td>72.7</td>
<td>0.2</td>
<td>67.4</td>
<td>0.7</td>
<td>60.4</td>
</tr>
<tr>
<td>20</td>
<td>180</td>
<td>53.8</td>
<td>1.0</td>
<td>44.8</td>
<td>1.7</td>
<td>36.8</td>
</tr>
<tr>
<td>30</td>
<td>118</td>
<td>46.3</td>
<td>1.5</td>
<td>35.0</td>
<td>2.4</td>
<td>26.0</td>
</tr>
<tr>
<td>40</td>
<td>87.2</td>
<td>40.2</td>
<td>1.9</td>
<td>26.2</td>
<td>3.0</td>
<td>16.8</td>
</tr>
<tr>
<td>50</td>
<td>69.8</td>
<td>36.6</td>
<td>2.1</td>
<td>24.1</td>
<td>3.3</td>
<td>14.5</td>
</tr>
<tr>
<td>60</td>
<td>58.1</td>
<td>32.4</td>
<td>2.7</td>
<td>16.4</td>
<td>3.6</td>
<td>9.54</td>
</tr>
<tr>
<td>100</td>
<td>34.1</td>
<td>23.7</td>
<td>3.3</td>
<td>6.09</td>
<td>3.8</td>
<td>5.57</td>
</tr>
<tr>
<td>300</td>
<td>11.2</td>
<td>9.88</td>
<td>4.0</td>
<td>1.77</td>
<td>4.1</td>
<td>1.74</td>
</tr>
<tr>
<td>500</td>
<td>6.85</td>
<td>6.27</td>
<td>4.0</td>
<td>1.05</td>
<td>4.2</td>
<td>1.09</td>
</tr>
<tr>
<td>1000</td>
<td>3.45</td>
<td>3.29</td>
<td>4.0</td>
<td>0.49</td>
<td>4.1</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Table 3-4. Estimated number of bands for rock and metal of sonar data

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>0.05</th>
<th>0.2</th>
<th>0.3</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rock</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Metal</td>
<td>6</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

70
CHAPTER 4
NONPARAMETRIC ESTIMATION OF LARGE CORRELATION MATRICES BY SMOOTHING BANDS IN THE PARTIAL AUTOCORRELATION

In this paper, we improve the estimator of a correlation matrix by using nonparametric regression method on the banded estimator of partial autocorrelation matrix obtained from Wang and Daniels (2012). In particular, we smooth the partial autocorrelations within bands. An asymptotic optimal bandwidth is derived based on Toeplitz condition on a banded partial autocorrelation matrix, and a plug-in bandwidth selector is suggested. The favorable properties of this estimator is demonstrated by simulation. The estimator is illustrated on two high dimensional data sets. The simulations suggests that the Toeplitz condition on partial autocorrelation matrices within lag is not restrictive.

4.1 Introduction

Estimating a large covariance matrix is very important in many applications and has been an active research area in the last ten years. The sample covariance matrix is a simple nonparametric estimator of covariance matrix, but it behaves poorly when the data are severely unbalanced or the dimension of covariance matrix is large compared to the number of observations (Stein (1975)). We review some of the most relevant literature here. Rice and Silverman (1991) suggested a nonparametric method to estimate a covariance matrix by smoothing the first few eigenfunctions. Hall et al. (1994) used kernel methods for estimating covariance functions for a stationary stochastic process, and employed the Fourier transformation to ensure the positive semidefiniteness condition for the estimated covariance functions. Diggle and Verbyla (1998) provided a kernel-weighted local linear regression estimator for estimating the nonstationary variogram for longitudinal studies, but can not guarantee the estimated covariance matrix is positive definite. Wu and Pourahmadi (2003) examined a nonparametric method based on Cholesky decomposition of a covariance matrix with a local regression polynomial smoother. Huang et al. (2007) consider a method following Cholesky decomposing a covariance matrix, then approximating the subdiagonals...
of Cholesky decomposition by splines. Li et al. (2007) proposed a nonparametric estimator of correlation functions and proved asymptotic normality for its sampling distribution under certain assumptions, but they need use the procedure in Hall et al. (1994) to ensure positive definiteness. Fan et al. (2007) and Fan and Wu (2008) propose a viable semiparametric model for covariance matrix by estimating variance functions nonparametrically, while estimating the correlation function parametrically according to information from irregular and sparse data points on each subject. Yin et al. (2010) suggested a nonparametric model to estimate conditional covariance matrices. Li (2011) proposed a method to estimate covariance matrices for longitudinal data and showed that kernel covariance estimation provides uniformly consistent estimators for the within-subject covariance matrices.

Boundary bias is a problem associated with kernel smoothing for estimation near the boundaries of the range. Jones (1993) outlined boundary bias correction methods for kernel density estimation and pointed out important connections with the regression situation. Specifically, a linear combination of two kernels provides a simple and effective bias correction Hastie and Loader (1993).

Selecting the smoothing parameter is another issue linked to nonparametric estimation. Cross-validation and variations provide simple methods to choose the smoothing parameter with the given answer asymptotically optimal under very weak conditions (Stone (1984)). However, it performs poorly for kernel density estimation Park and Marron (1990b) and can be computationally expensive.

Ruppert et al. (1995) proposed a method to estimate an asymptotically optimal bandwidth for independent data by plugging in estimators based on a blocked locally quartic polynomial fit with the number of block, N, chosen by Mallows’ $C_p$. They adjust for boundary bias by truncating boundary regions. Plug-in methods assuming independent data underestimate (overestimate) the appropriate bandwidth as data is positively (negatively) correlated. Chiu (1989) provided a bandwidth selection method
for correlated data by modifying Mallows’ Criterion, while Herrmann et al. (1992) modified the independent plug-in bandwidth with a correction factor for correlated data.

Generally, covariance matrices are sparse as the dimension m is large. Bickel and Levina (2008b) achieve sparsity by banding the sample covariance matrix while Rothman et al. (2010) band the Cholesky factor of the covariance matrix, but the former has to use ‘tapering’ to achieve positive-definite (Furrer and Bengtsson (2007)). Wang and Daniels (2012) investigate sparsity by banding the partial autocorrelation matrix without concern about the positive definiteness condition on the estimated correlation matrix. In this follow up paper of Wang and Daniels (2012), we present a method to smooth the bands of the partial autocorrelation matrix, using kernel smoothing as in Altman (1990). We use the approach in Fun and Gijbels (1996) to accommodate the boundary bias problem in kernel estimation. The procedure is computationally quick even for high dimensional covariance matrices via the use of a plug-in bandwidth (no cross-validation).

This chapter is arranged as follows. In Section 4.2, we briefly review the partial autocorrelation matrix. In Section 4.3, we discuss some relevant properties of a banded partial autocorrelation matrix. In Section 4.4, we discuss some theoretical results related to our smoothing estimator and propose a simple algorithm to compute the estimator. Section 4.5 and 4.6 investigate the operating characteristics of our procedure via risk simulations and apply it to two real data examples, respectively.

Note, in this chapter, m denotes the dimension of partial autocorrelation matrix, and n denotes the number of observations.

4.2 Review of Partial Autocorrelations

We first review reparameterizing the correlation matrix $R = (\rho_{jk})$ using the elements in the partial autocorrelation matrix $\Pi = (\pi_{jk})$, where $\pi_{jj} = 1$ and for $1 \leq j < k \leq m$, $\pi_{jk}$ is the correlation between $Y_j$ and $Y_k$ adjusted for the intervening variables,
\{Y_{j+1}, \ldots, Y_{k-1}\}. Unlike \( R \), the off-diagonal elements of \( \Pi \) can independently vary in \((-1, 1)\) with the corresponding correlation matrix, \( R \), remaining positive-definite.

There is a simple relationship between the elements of \( R \) and \( \Pi \). The partial autocorrelation \( \pi_{jj+1} = \rho_{jj+1} \) and \( \pi_{j,j+k}, \,(2 \leq k \leq m - 1) \) is

\[
\pi_{j,j+k} = \frac{\rho_{j,j+k} - r_1(j, k)R_2^{-1}(j, k)r_3^T(j, k)}{[1 - r_1(j, k)R_2^{-1}(j, k)]^{1/2}[1 - r_3(j, k)R_2^{-1}(j, k)r_3^T(j, k)]^{1/2}}.
\] (4–1)

with \( r_1(j, k) = (\rho_{j, j+1}, \ldots, \rho_{j, j+k-1}) \), \( r_3(j, k) = (\rho_{j+k, j+1}, \ldots, \rho_{j+k, j+k-1}) \), and \( R_2(j, k) \) contains the middle \( k - 1 \) rows and columns of \( R[j : j + k] \), i.e.

\[
R[j : j + k] = \begin{pmatrix}
1 & r_1(j, k) & \rho_{j,j+k} \\
 r_1^T(j, k) & R_2(j, k) & r_3^T(j, k) \\
\rho_{j+k,j} & r_3(j, k) & 1
\end{pmatrix}
\] (4–2)

Correspondingly, \( \rho_{j,j+k} \) can be written as a function of \( \pi_{j,j+k} \),

\[
\rho_{j,j+k} = r_{jk} + \pi_{j,j+k}D_{jk},
\] (4–3)

where

\[
r_{jk} = r_1(j, k)R_2^{-1}(j, k)r_3^T(j, k)
\]

\[
D_{jk}^2 = [1 - r_1(j, k)R_2^{-1}(j, k)r_3^T(j, k)][1 - r_3(j, k)R_2^{-1}(j, k)r_3^T(j, k)].
\]

A partial autocorrelation matrix will often be sparse (Dempster (1972); Friedman et al. (2008)) since the correlations in it are conditional (on the intervening variables) correlations. For example, an AR(1) correlation matrix corresponds to a partial autocorrelation matrix with only one non-zero band. More generally, a partial autocorrelation matrix will have \( k \) non-zero bands under a \( k \)-th order ante-dependence model (Gabriel (1962); Zimmerman and Nunez-Anton (2010)).
4.3 Some Properties of $k_0$ Band Partial Autocorrelation Matrices

The strategy in this paper will be to first band the partial autocorrelation matrix and then smooth the non-zero bands. We first provide some useful properties of the partial autocorrelation matrix.

Assume $\Pi$ is a partial autocorrelation matrix and $R$ is its corresponding correlation matrix. To investigate properties of $k_0$ band partial autocorrelation matrices, we first state some useful facts here.

Fact 1: (Joe (2006)) If $\Pi$ is a toeplitz matrix, so is $R$.

Fact 2: (Wang and Daniels (2012)) If $\Pi$ is a $k_0$-band partial autocorrelation matrix, then the inverse of the corresponding correlation matrix, $R^{-1}$ is a $k_0$-band matrix.

Fact 3: (Demko et al. (1984)) Let $A$ be a position definite, $k_0$-band, bounded and invertible matrix with an Euclidean metric, and $[a, b]$ be the smallest interval containing the spectrum, $\sigma(A)$, of the matrix $A$. Set $r = \frac{b}{a}$, $C_0 = \frac{(1+\sqrt{r})^2}{2ar}$ and $\lambda = \left(\frac{\sqrt{r}-1}{\sqrt{r}+1}\right)^{\frac{k_0}{2}}$. Then, the $(i, j)^{th}$ element of $A^{-1}$ satisfies

$$|A_{(i,j)}^{-1}| \leq C \lambda^{|i-j|}$$  \hspace{1cm} (4–4)

where $C = \max\{a^{-1}, C_0\}$.

This result shows that entries of inverse matrix of a banded matrix exponentially decay with lag. Moreover the decay rate is only related to number of bands ($k_0$) and the range of its eigenvalues, regardless of the dimension of the matrix. In the case of a banded partial autocorrelation matrix, the entries of corresponding correlation matrix are exponentially decaying as lag increases. This is clearly seen in an AR(1) correlation matrix which induces a 1-band partial autocorrelation matrix.

In the following, let $m$ denote the dimension of the random vector $Y_i : i = 1, 2, \ldots, n$. We have following proposition immediately.

Proposition 4.1. Let $R_m = (\rho_{m(j+k)})$ be the corresponding correlation matrix of a $k_0$-band partial autocorrelation matrix $\Pi_m$, and let $[a_m, b_m]$ be the smallest interval
containing the spectrum $\sigma(R_m^{-1})$ of the matrix $R_m^{-1}$. Set $r_m = \frac{b_m}{a_m}$, $C_m^* = \frac{(1+\sqrt{r_m})^2}{2a_mr_m}$, and 
\[ \lambda = \left( \frac{\sqrt{r_m}-1}{\sqrt{r_m}+1} \right)^2. \]

Then, 
\[ |\rho_{m(j,j+k)}| \leq C_m\lambda^{[k]}, \]
where $C_m = \max\{a_m^{-1}, C_m^*\}$.

Furthermore, for $m = 2, 3, ...$ assume $a_0 = \inf_m\{a_m\} > 0$, $b_0 = \sup_m\{b_m\} < +\infty$, and $C_0 = \sup_m\{C_m\} < +\infty$. Set $r_0 = \frac{b_0}{a_0}$, and $\lambda_0 = \left( \frac{\sqrt{r_0}-1}{\sqrt{r_0}+1} \right)^2$.

Then, 
\[ |\rho_{m(j,j+k)}| \leq C_0\lambda_0^{[k]}, \text{ for all } m \geq 2. \]

Now, assume $Y_1, Y_2, ..., Y_n$ are iid m-dimensional multivariate normal random vectors with mean zero and variance-covariance matrix $\Sigma_m = D_m R_m D_m$, where $D_m$ is a diagonal matrix with standard deviations on its main diagonal and $R_m$ is a correlation matrix which corresponds to partial autocorrelation matrix $\Pi_m$. Let $\hat{\Pi}_m$ be the corresponding sample partial autocorrelation matrix of sample correlation matrix $\hat{R}_m$. $\hat{\Pi}_m$ is the MLE of $\Pi_m$ (Wang and Daniels (2012)). Now, we define $vh(\hat{\Pi}_m)$ as
\[ vh(\hat{\Pi}_m) = (\hat{\pi}_{21}, \hat{\pi}_{31}, ..., \hat{\pi}_{m1}, \hat{\pi}_{32}, ..., \hat{\pi}_{m2}, ..., \hat{\pi}_{mm-1})^T, \]
a vector with all the elements in the lower triangular part of $\hat{\Pi}_m$ (without the main diagonal). A property of the asymptotic covariance matrix of $vh(\hat{\Pi}_m)$ is given below.

**Theorem 4.1.** Let $\hat{\Pi}_m = (\hat{\pi}_{m(j,j+k)})$ be the sample partial autocorrelation matrix of an m-dimensional multivariate normal distribution based on an iid sample, $Y_i \sim N(\mu_m, D_m R_m D_m), i = 1, 2, ..., n$. $R_m$ is the correlation matrix which corresponds to a $k_0$-band partial autocorrelation matrix. Let $[a_m, b_m]$ be the smallest interval containing the spectrum $\sigma(R_m^{-1})$ of the matrix $R_m^{-1}$. Set $r_m = \frac{b_m}{a_m}$, $C_m^* = \frac{(1+\sqrt{r_m})^2}{2a_mr_m}$ and $\lambda = \left( \frac{\sqrt{r_m}-1}{\sqrt{r_m}+1} \right)^2$.

Assume
a. \( a_0 = \inf_m \{a_m\} > 0 \) and \( b_0 = \sup_m \{b_m\} < +\infty \), \( C_0 = \sup_m \{C_m\} < +\infty \). Set \( r_0 = \frac{b_0}{a_0} \)

and \( \lambda_0 = \left( \frac{\sqrt{\pi_0 - 1}}{\sqrt{r_0} + 1} \right)^{\frac{2}{r_0}} \).

b. \(-1 < a_0 < \pi_m(j, j + k) < a_1^* < 1\) for some \( a_0^* \) and \( a_1^* \), and all \( m \in \{2, 3, \ldots\} \).

Let \( \Omega_{\pi, m, k} = (\omega_{\pi, m, k(j, j_1)}) \) denote the asymptotic covariance matrix of all the lag \( k \) elements in \( \hat{\Pi}_m \), where \( \omega_{\pi, m, k(j, j_1)} \) represents the covariance between \( \hat{\pi}_{m(j, j + k)} \) and \( \hat{\pi}_{m(j_1, j_1 + k)} \). Then,

\[
\omega_{\pi, m, k(j, j_1)} = \sum_{s=1}^{k} \sum_{t=0}^{s-1} \frac{\partial \pi_{ij+k}}{\partial \rho_{i+sj+t+1}} \omega_{ij+sj+t+1, ij+sj+t+1} \frac{\partial \pi_{ij+k}}{\partial \rho_{ij+sj+t+1}} = o\left( \lambda_0^{\frac{2}{|j|+j_1}} \right),
\]

(4–5)

where \( \omega_{ij+sj+t+1, ij+sj+t+1} \) is the covariance between \( \rho_{ij+sj+t+1} \) and \( \rho_{ij+sj+t+1} \).

**Proof:** See the appendix.

This theorem indicates that under conditions a) and b), the covariances of the estimated partial autocorrelations within a band decay exponentially. For example, entries of an AR(1) correlation matrix, which corresponds to a one band partial autocorrelation matrix, are exponentially decaying. At the same time, the covariances of estimated partial autocorrelations within a lag are also exponentially decaying with a faster rate. Similar behavior is seen with \( m^{th}\)-order antedependence models (Gabriel (1962)). We also have the following proposition.

**Proposition 4.2.** Assume \( \Pi_m \) is a Toeplitz matrix with \( k_0 \) bands and satisfies the conditions in Theorem 1. Then \( \Omega_{\pi, m, k} = (\omega_{\pi, m, k(j, j_1)}) \), the asymptotic covariance matrix of all lag \( k \) elements in \( \hat{\Pi}_m \), is a Toeplitz matrix with common element, \( \sigma^2_{\pi, k} \), in each entry on the main diagonal. Thus we can denote \( \Omega_{\pi, m, k} = (\omega_{\pi, m, k}(t)) \) where \( t = |j - j_1| \). Then we have

- \( \sum_{t=1}^{\tilde{k}} |\omega_{\pi, m, k}(t)| \) converges as \( m \) and \( \tilde{k} \rightarrow \infty \),
- \( \sum_{t=1}^{\tilde{k}} t |\omega_{\pi, m, k}(t)| = o(\tilde{k}) \) as \( m \) and \( \tilde{k} \rightarrow \infty \).

For subsequent development, we define

\[
S_{\pi, k} = \sum_{t=1}^{\tilde{k}} \omega_{\pi, m, k}(t) \text{ and } S_{\pi, k} = \lim_{\tilde{k} \to \infty} S_{\pi, k}.
\]

(4–6)
4.4 Nonparametric Estimation of Partial Autocorrelation Coefficients within Bands

We use the results in Section 3 to smooth the partial autocorrelations in each band. Let \( \pi^{**}_{k,m} \) denote the vector \( (\pi_{1k+1}, \pi_{2k+2}, \ldots, \pi_{m-k,m}) \), \( k = 1, 2, \ldots, m-1 \), the vector of entries on \( k \)th subdiagonals of \( \Pi_m \). We assume that components of \( \pi_{k,m} \), are realizations of a smooth function \( \pi^*_{k,m}(\cdot) \) on \([0, 1]\), i.e., \( \pi_{k,m}(t) = \pi^*_{k,m}(\frac{t}{m-k}) \).

For convenience, we denote \( x = \frac{t}{m-k} \) and \( x_j = \frac{j}{m-k}, j=1, 2, \ldots, m-k \). Let \( \hat{\pi}_{k,m}(\frac{t}{m-k}) := \hat{\pi}^*_{k,m}(x) \).

Assume

\[
\hat{\pi}^*_{k,m}(x) = \pi^*_{k,m}(x) + \epsilon_k(x)
\]

where \( \pi^*_{k,m}(x) \) is a smooth mean function and \( \epsilon_k \) is an error process with mean zero.

Now we choose a kernel \( K(\cdot) \) satisfying condition A, B, C in Altman (1990),

A. \( K \) is symmetric about 0.

B. \( K \) has support only on the interval \((-\frac{1}{2}, \frac{1}{2})\).

C. \( K \) is Lipschitz continuous of order \( \alpha^* > 0 \).

We also assume that \( K \) is a \( p \)th order kernel (Gasser and Muller (1979)), which satisfies,

\[
s_K = \int x^i K(x) dx = \begin{cases} 
0 & \text{if } i < p \\
\neq 0 & \text{if } i = p 
\end{cases}
\]

Most commonly used kernels have \( p=2 \), e.g., the Epanechnikov kernel, \( K(x) = \frac{3}{4}(1 - x^2)I(|x| \leq 1) \), which we will use here.

Let the estimator as a function of the bandwidth \( h \) be

\[
\hat{\pi}^*_{k,m}(x, h, K) = \sum_{t=0}^{m-k} w_{h,m-k}(x, t) \hat{\pi}^*_{k,m}(t),
\]

with weights

\[
w_{h,m-k}(x, t) = K\left(\frac{x-h \cdot t}{h}\right) / (m-k)h.
\]
where h is the bandwidth, m is the dimension of the partial autocorrelation matrix, and k is lag.

4.4.1 Theoretic Results

Without loss of generality, we drop subscripts k and m of \( \hat{\pi}^*_{k,m}(x) \) and \( \hat{\pi}^*_{k,m}(x, h, K) \), and then using results in Altman (1990) we arrive at the following theorem.

**Theorem 4.2.** Suppose that the mean function \( \pi^*(\cdot) \) in Eq. 4–7 has p derivatives and the kernel function K is of order p and satisfies conditions A-C. Suppose that the points are equally spaced on support of function interval of estimation and partial autocorrelation matrix \( \Pi_m \) satisfies the conditions in Theorem 1.

Then for \( \frac{\varsigma}{2} \leq x \leq 1 - \frac{\varsigma}{2} \), where \( \varsigma \) is a small positive number.

\[
MSE(\hat{\pi}(x, h, K)) = E(\hat{\pi}^*_{(x,h,K)} - \pi^*(x))^2
\]

\[
= \left( \frac{h^p \pi^*(p)(x)s_K}{p!} \right)^2 + \frac{N_K(\sigma_{\pi,k}^2 + 2S_{\pi,k})}{(m-k)h} + o(1/(m-k)h) + o(h^{2p})
\]

where \( s_K = \int x^p K(x)dx, N_K = \int K^2(x)dx, \) and \( S_{\pi,k} \) is defined in Eq. 4–6.

Typically, we want to estimate \( \pi^*(x) \) globally over the support of \( \pi^*(x) \) and an appropriate error criteria is conditional mean integrated squared error (Ruppert and Wand (1994)).

\[
MISE(\hat{\pi}(x, h, K))
\]

\[
= E \int [\hat{\pi}^*_{(x,h,K)} - \pi^*(x)]^2 v(x)dx
\]

\[
= \left( \frac{h^p s_K}{p!} \right)^2 \int (\pi^*(p)(x))^2 v(x)dx + \frac{N_K(\sigma_{\pi,k}^2 + 2S_{\pi,k}) \int v(x)dx}{(m-k)h} + o(1/(m-k)h) + o(h^{2p})
\]

where \( v(x) \) is a weight function. By minimizing MISE as a function of \( h \), we obtain asymptotic optimal bandwidth given in the following corollary.

**Corollary 4.2.1.** Under the conditions of Theorem 2, the asymptotically optimal bandwidth is

\[
h_{m,k} = \left\{ \frac{(p!)^2 N_K(\sigma_{\pi,k}^2 + 2S_{\pi,k}) \int v(x)dx}{2ps_K \int (\pi^*(p)(x))^2 v(x)dx} \right\}^{\frac{1}{2p+1}} (m-k)^{-\frac{1}{2p+1}}.
\]

(4–8)
The asymptotic optimal bandwidth given above is a generalization of asymptotic optimal bandwidth in Hart (1991) and Herrmann et al. (1992) for \( p=2 \).

**Corollary 4.2.2.** Under the conditions of Theorem 2, the kernel estimator is consistent, as \( h \to 0 \) and \( (m-k)h \to +\infty \).

**Proofs:** Trivial extensions of results on Altman (1990).

Boundary bias, which happens near the boundaries of the range. To balance between bias and variance, a local linear regression smoother, which is a linear combination of \( K(x) \) and \( xK(x) \) (Hastie and Loader (1993)) is chosen for our situation due to that our primary goal is estimate \( \pi^*(x) \) instead some degree of derivative of \( \pi^*(x) \). It also provides desirable properties at boundaries.

**4.4.2 Estimation**

Our procedure for estimating a correlation matrix proceeds in two steps,

- Estimate, \( \hat{k}_0 \), the number of bands in the sample partial autocorrelation matrix.
- Smooth the partial autocorrelation coefficients within each band \( k (k \leq \hat{k}_0) \).

**4.4.2.1 Choose number of bands**

Wang and Daniels (2012) developed a procedure to estimate the number of bands in a partial autocorrelation matrix based on a sequential multiple hypothesis testing procedure with modified type I error rate. The advantages of this procedure include: 1) Hypothesis testing is based on exact distributions, 2) Inversion of high dimensional matrices. 3) No concerns about the positive-definite condition on the estimated correlation matrix. We estimate the number of bands, \( \hat{k}_0 \), with this procedure.

**4.4.2.2 Smooth estimates within each band**

Once we estimate the number of bands, we smooth their entries. Corollary 4.2.2 guarantees our smoothing estimator is consistent. The main remaining concern is choosing the smoothing parameter \( h \). To estimate the bandwidth in Eq. 4–8, we need to estimate \( \sigma^2_{\pi,k}, S_{\pi,k} \), and \( \theta_{r,p} = \int (\pi^*(p)(x))^2 v(x) dx \). Ruppert et al. (1995) suggested a way to estimate \( \sigma^2_{\pi,k}, \theta_{r,s} = \int \pi^*(r)(x)\pi^*(s)(x) v(x) dx \). Details are given in the appendix.
4.5 Simulation Study

To evaluate the operating characteristics of our estimator, we conduct several simulations with true models for partial autocorrelations, $\pi_{j,j+k}$, below.

- **Model 1:** AR(1) structure
  $\pi_{j,j+k} = \rho \cdot I(|k| = 1)$, for $\rho = 0, 0.2, 0.5, 0.9$,
  which corresponds to all higher lag partial autocorrelations equal to zero except lag 1. $\rho = 0$ is a special case of within subject independence, and $\rho = 0.2, 0.5, 0.9$ represent to weak, moderate, and strong within subject correlation.

- **Model 2:** 4 band partial autocorrelation
  $\pi_{j,j+k} = 0.4I(|k| = 1) + 0.2I(2 \leq |k| \leq 3) + 0.1I(|k| = 4),$
  which corresponds to the partial autocorrelation matrix satisfying Toeplitz condition and reflects the partial autocorrelation matrix having moderate partial autocorrelations on lag 1 and small partial autocorrelations on lag 2, and lag 3, very small partial autocorrelations on lag 4.

- **Model 3:** 5 band partial autocorrelation (I)
  $\pi_{j,j+k} = I(|k| < 6) \cdot \max\{0, \exp(-(|k| + 2)/4) - (1 + j)^{-1.5}\}$.
  The partial autocorrelation coefficients are decreasing with time ($j$) within each lag, and are exponentially decaying as lag ($k$) increases for a given time.

- **Model 4:** 9 band partial autocorrelations (II)
  $\pi_{j,j+k} = \frac{1}{2} \cdot \left(1 + \sin\left(\frac{60-j}{6}\right)\right) \cdot \left(\frac{1}{2} + k\right)^{-1.02}I(k < 10)$.
  The partial autocorrelations are periodic functions of time.

For each true matrix, we simulate 100 data sets with dimension $m=60$ and different sample sizes from a multivariate normal distribution with mean zero and variance-covariance matrix $\Sigma$ equal to corresponding correlation matrix for each case above.

We estimate the smoothed partial autocorrelation matrix by averaging the smoothed estimators of partial autocorrelation matrices over 100 replicated data sets and then constructing the estimated correlation matrix. We compare the smoothed correlation estimator to sample correlation, Ledoit-Wolf estimator, the banded estimator and smoothed estimator with smoothing parameter chosen by Ruppert et al. (1995). We calculate a mean sum of square error loss (MSSR) of the estimated correlation matrix by first calculating squared $L_2$ norm between the corresponding correlation matrix of
estimated partial autocorrelation and true correlation matrix, then average this squared error loss over 100 replicates. The results are recorded in Tables 4-1 to 4-4.

Table 4-1 provides the mean sum of square error for the five estimators mentioned for model 1. We can see that smoothed partial autocorrelation coefficients within each band significantly reduce the mean sum of square error (MSSR) compared to banded partial autocorrelation estimator and sample correlation estimator. Furthermore, our smoothing estimator reduces this quantity around 5% compared to the method suggested by Ruppert et al. (1995). Table 4-2 provides results for model 2 and has a similar story to model 1. Table 4-3 records the results for model 3, which does not satisfy Toeplitz condition. We can see that Wang and Daniels (2012) estimate the band very well, even if a sample size is small (comparing to dimension p=60), and the smoothed estimators reduce their MSSR more than 50% for small sample size (comparing to MLE and W-L estimator). Although model 4 is more complicated due to periodically varying pattern, MSSR shows similar improvements (Table 4-4).

4.6 Applications

We illustrate our approach on two data sets, the Metal and Rock data of the sonar data, which is available at http://www.ics.uci.edu/~mlearn/MLRepository.html. This data set contains 111 signals from a Metal cylinder and 97 signals from a Rock, where each signal has 60 frequency energy measurements ranging from 0.0 to 1.0. These signals were measured at different angles for the same objects. As in previous analysis of this data we assume the signals are iid normal random vectors (Rothman et al. (2010)).

Figure 4-2 and 4-3 display estimated partial autocorrelations and the estimated smoothed partial autocorrelations for each band for metal and rock data, respectively. For the metal data, the fit at the boundaries is very good, except on the right boundary of lag 2 fitting. The fit here is driven by the estimated partial autocorrelation at the final frequency, which may be an outlier. For lag 5, there is a big change around frequency
30, which is confirmed by the image plot of correlation matrix of metal data in Figure 4-4A. The image of the correlation matrix which corresponds to smoothed partial autocorrelation within estimated bands, appears to capture the main features of the sample correlation matrix. Compared to metal data, the correlation structure of rock data appears to be much simpler. Figure 4-3 shows clearly how the smoothed curves capture the characteristic of estimated partial autocorrelation within each band. Note that the smoothed curves violate the Toeplitz condition used in our theorems. However, the curve fitting still works well which we observed from the decrease in sum of mean squared error in the simulation study in section 4.5.

4.7 Discussion

We have proposed a nonparametric smoothing method to improve the estimation of a banded partial autocorrelation matrix, even when the dimension m is less than sample size n. This estimator is consistent as long as partial autocorrelations satisfy some regularity conditions for the kernel and a Toeplitz condition on the partial autocorrelation matrix. The main condition, convergence of covariances between the partial autocorrelations does not require the Toeplitz condition, and in any case, the Toeplitz condition does not appear to be restrictive from our simulation results. We are doing further work to see if the Toeplitz condition is really necessary. Another condition in our theorem is equally spaced design points, a common assumption in the nonparametric literature (Altman (1990); Hart (1991), and Ruppert et al. (1995)). However, Rice and Wu (2001) discuss a plug-in bandwidth estimator for unequally spaced design points. We plan to explore more on this topic as well.
Table 4-1. Simulation records for AR(1) structure

\[
\text{Type I Error} = 0.2I(n > m) + 0.3I(n \leq m)
\]

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>n</th>
<th>MSSE(MLE)</th>
<th>MSSE(L-W)</th>
<th>MSSE(B-E)</th>
<th>MSSE(S-R)</th>
<th>MSSE(S-D)</th>
<th>EB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10</td>
<td>394.42</td>
<td>16.562</td>
<td>9.8683</td>
<td>0.7755</td>
<td>0.7380</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>186.29</td>
<td>2.1049</td>
<td>6.5160</td>
<td>0.4697</td>
<td>0.4508</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>122.00</td>
<td>0.5982</td>
<td>4.7379</td>
<td>0.3338</td>
<td>0.3193</td>
<td>1.0</td>
</tr>
<tr>
<td>0.2</td>
<td>40</td>
<td>90.746</td>
<td>0.3062</td>
<td>4.4952</td>
<td>0.3502</td>
<td>0.3319</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>72.118</td>
<td>0.1821</td>
<td>2.4859</td>
<td>0.1868</td>
<td>0.1827</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>59.848</td>
<td>0.1216</td>
<td>2.5685</td>
<td>0.2203</td>
<td>0.2120</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>35.810</td>
<td>0.0514</td>
<td>0.5966</td>
<td>0.0423</td>
<td>0.0401</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>17.699</td>
<td>0.0125</td>
<td>0.3665</td>
<td>0.0328</td>
<td>0.0313</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>7.1338</td>
<td>0.0059</td>
<td>0.1587</td>
<td>0.0118</td>
<td>0.0111</td>
<td>0.6</td>
</tr>
<tr>
<td>0.5</td>
<td>10</td>
<td>390.44</td>
<td>19.893</td>
<td>17.176</td>
<td>3.2490</td>
<td>3.1923</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>184.76</td>
<td>6.5728</td>
<td>9.6111</td>
<td>1.1581</td>
<td>1.1156</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>121.92</td>
<td>5.3733</td>
<td>8.6966</td>
<td>0.9157</td>
<td>0.8848</td>
<td>1.9</td>
</tr>
<tr>
<td>0.1</td>
<td>40</td>
<td>90.668</td>
<td>5.0077</td>
<td>6.2899</td>
<td>0.4947</td>
<td>0.4678</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>72.004</td>
<td>4.7718</td>
<td>5.9839</td>
<td>0.5233</td>
<td>0.4999</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>60.102</td>
<td>4.6738</td>
<td>5.1496</td>
<td>0.4366</td>
<td>0.4195</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>35.675</td>
<td>4.3781</td>
<td>1.8353</td>
<td>0.1598</td>
<td>0.1517</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>17.756</td>
<td>3.8690</td>
<td>0.8891</td>
<td>0.0785</td>
<td>0.0760</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>7.0793</td>
<td>2.9074</td>
<td>0.3643</td>
<td>0.0324</td>
<td>0.0313</td>
<td>1.5</td>
</tr>
<tr>
<td>0.1</td>
<td>10</td>
<td>385.32</td>
<td>49.318</td>
<td>21.559</td>
<td>2.9131</td>
<td>2.8288</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>183.58</td>
<td>34.549</td>
<td>14.213</td>
<td>2.0915</td>
<td>2.0309</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>119.93</td>
<td>30.303</td>
<td>9.0232</td>
<td>1.1918</td>
<td>1.1386</td>
<td>1.9</td>
</tr>
<tr>
<td>0.5</td>
<td>40</td>
<td>89.388</td>
<td>27.660</td>
<td>7.0238</td>
<td>0.9396</td>
<td>0.9051</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>72.073</td>
<td>25.738</td>
<td>6.7978</td>
<td>0.9736</td>
<td>0.9511</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>60</td>
<td>58.620</td>
<td>23.725</td>
<td>3.9621</td>
<td>0.5906</td>
<td>0.5700</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>35.015</td>
<td>18.588</td>
<td>1.8269</td>
<td>0.2619</td>
<td>0.2538</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>17.486</td>
<td>12.106</td>
<td>0.8639</td>
<td>0.1299</td>
<td>0.1260</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>6.9915</td>
<td>5.9671</td>
<td>0.3755</td>
<td>0.0553</td>
<td>0.0531</td>
<td>1.4</td>
</tr>
</tbody>
</table>
Table 4-1. Continued

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$n$</th>
<th>$\text{MSSE(MLE)}$</th>
<th>$\text{MSSE(L-W)}$</th>
<th>$\text{MSSE(B-E)}$</th>
<th>$\text{MSSE(S-R)}$</th>
<th>$\text{MSSE(S-D)}$</th>
<th>$EB$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>326.50</td>
<td>206.62</td>
<td>42.099</td>
<td>24.830</td>
<td>24.488</td>
<td>1.5</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>152.37</td>
<td>122.30</td>
<td>20.853</td>
<td>13.248</td>
<td>13.028</td>
<td>1.9</td>
</tr>
<tr>
<td>30</td>
<td>30</td>
<td>98.265</td>
<td>86.534</td>
<td>13.608</td>
<td>8.3355</td>
<td>8.2051</td>
<td>1.8</td>
</tr>
<tr>
<td>0.9</td>
<td>40</td>
<td>71.973</td>
<td>65.318</td>
<td>10.157</td>
<td>6.2784</td>
<td>6.1875</td>
<td>2.2</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>58.033</td>
<td>53.484</td>
<td>8.9294</td>
<td>5.5626</td>
<td>5.4666</td>
<td>2.4</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>49.068</td>
<td>46.224</td>
<td>7.2929</td>
<td>4.5916</td>
<td>4.5422</td>
<td>2.2</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>27.892</td>
<td>27.390</td>
<td>3.7617</td>
<td>2.5758</td>
<td>2.5555</td>
<td>1.5</td>
</tr>
<tr>
<td>200</td>
<td>200</td>
<td>14.378</td>
<td>13.946</td>
<td>1.8139</td>
<td>1.2421</td>
<td>1.2288</td>
<td>1.6</td>
</tr>
<tr>
<td>500</td>
<td>500</td>
<td>5.6128</td>
<td>5.5801</td>
<td>0.7474</td>
<td>0.5007</td>
<td>0.4957</td>
<td>1.6</td>
</tr>
</tbody>
</table>

Table 4-2. Simulation records for model 2

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\text{MSSE(MLE)}$</th>
<th>$\text{MSSE(L-W)}$</th>
<th>$\text{MSSE(B-E)}$</th>
<th>$\text{MSSE(S-R)}$</th>
<th>$\text{MSSE(S-D)}$</th>
<th>$EB$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>378.05</td>
<td>72.727</td>
<td>54.018</td>
<td>25.008</td>
<td>24.884</td>
<td>2.1</td>
</tr>
<tr>
<td>20</td>
<td>180.92</td>
<td>53.819</td>
<td>33.838</td>
<td>11.430</td>
<td>11.230</td>
<td>3.7</td>
</tr>
<tr>
<td>30</td>
<td>117.82</td>
<td>46.258</td>
<td>22.205</td>
<td>5.8647</td>
<td>5.7664</td>
<td>4.3</td>
</tr>
<tr>
<td>40</td>
<td>87.246</td>
<td>40.171</td>
<td>16.551</td>
<td>3.8079</td>
<td>3.7195</td>
<td>4.5</td>
</tr>
<tr>
<td>50</td>
<td>69.836</td>
<td>36.566</td>
<td>13.503</td>
<td>2.8064</td>
<td>2.7274</td>
<td>4.9</td>
</tr>
<tr>
<td>60</td>
<td>58.135</td>
<td>32.433</td>
<td>11.135</td>
<td>2.5623</td>
<td>2.5045</td>
<td>4.8</td>
</tr>
<tr>
<td>100</td>
<td>34.142</td>
<td>23.698</td>
<td>5.5270</td>
<td>1.3442</td>
<td>1.3154</td>
<td>4.1</td>
</tr>
<tr>
<td>300</td>
<td>11.237</td>
<td>9.8833</td>
<td>1.8131</td>
<td>0.3677</td>
<td>0.3576</td>
<td>4.4</td>
</tr>
<tr>
<td>500</td>
<td>6.8513</td>
<td>6.6292</td>
<td>1.1541</td>
<td>0.2453</td>
<td>0.2383</td>
<td>4.6</td>
</tr>
</tbody>
</table>

Table 4-3. Simulation records for model 3

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\text{MSSE(MLE)}$</th>
<th>$\text{MSSE(L-W)}$</th>
<th>$\text{MSSE(B-E)}$</th>
<th>$\text{MSSE(S-R)}$</th>
<th>$\text{MSSE(S-D)}$</th>
<th>$EB$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>346.60</td>
<td>164.76</td>
<td>135.11</td>
<td>113.69</td>
<td>113.75</td>
<td>2.5</td>
</tr>
<tr>
<td>20</td>
<td>158.13</td>
<td>107.41</td>
<td>53.902</td>
<td>29.456</td>
<td>29.430</td>
<td>4.9</td>
</tr>
<tr>
<td>30</td>
<td>104.46</td>
<td>83.950</td>
<td>34.439</td>
<td>15.433</td>
<td>15.424</td>
<td>5.8</td>
</tr>
<tr>
<td>40</td>
<td>76.972</td>
<td>62.546</td>
<td>25.854</td>
<td>10.659</td>
<td>10.681</td>
<td>6.1</td>
</tr>
<tr>
<td>50</td>
<td>63.148</td>
<td>53.993</td>
<td>19.944</td>
<td>8.7859</td>
<td>8.8376</td>
<td>6.1</td>
</tr>
<tr>
<td>60</td>
<td>52.322</td>
<td>46.120</td>
<td>16.510</td>
<td>7.1734</td>
<td>7.1864</td>
<td>6.2</td>
</tr>
<tr>
<td>100</td>
<td>31.005</td>
<td>30.334</td>
<td>8.5979</td>
<td>4.3624</td>
<td>4.4437</td>
<td>5.5</td>
</tr>
<tr>
<td>300</td>
<td>10.259</td>
<td>9.9234</td>
<td>2.9361</td>
<td>1.8543</td>
<td>1.8816</td>
<td>5.4</td>
</tr>
<tr>
<td>500</td>
<td>6.1594</td>
<td>6.0401</td>
<td>1.7328</td>
<td>1.2049</td>
<td>1.1795</td>
<td>5.4</td>
</tr>
</tbody>
</table>
Figure 4-1. True partial autocorrelation curves for models
Table 4-4. Simulation records for model 4

<table>
<thead>
<tr>
<th>n</th>
<th>MSSE(MLE)</th>
<th>MSSE(L-W)</th>
<th>MSSE(B-E)</th>
<th>MSSE(S-R)</th>
<th>MSSE(S-D)</th>
<th>EB</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>365.56</td>
<td>119.18</td>
<td>86.525</td>
<td>72.510</td>
<td>72.210</td>
<td>2.0</td>
</tr>
<tr>
<td>20</td>
<td>169.23</td>
<td>83.046</td>
<td>46.664</td>
<td>28.213</td>
<td>27.971</td>
<td>4.3</td>
</tr>
<tr>
<td>30</td>
<td>113.29</td>
<td>67.569</td>
<td>35.801</td>
<td>17.893</td>
<td>17.667</td>
<td>5.8</td>
</tr>
<tr>
<td>40</td>
<td>84.891</td>
<td>56.258</td>
<td>25.361</td>
<td>12.150</td>
<td>11.981</td>
<td>6.2</td>
</tr>
<tr>
<td>50</td>
<td>66.769</td>
<td>47.538</td>
<td>21.171</td>
<td>8.1035</td>
<td>7.9282</td>
<td>7.2</td>
</tr>
<tr>
<td>60</td>
<td>55.232</td>
<td>41.339</td>
<td>18.541</td>
<td>6.2615</td>
<td>6.1104</td>
<td>8.1</td>
</tr>
<tr>
<td>100</td>
<td>33.123</td>
<td>27.829</td>
<td>11.283</td>
<td>4.6177</td>
<td>4.4801</td>
<td>8.1</td>
</tr>
<tr>
<td>300</td>
<td>10.822</td>
<td>10.079</td>
<td>3.9460</td>
<td>1.4429</td>
<td>1.3732</td>
<td>9.7</td>
</tr>
<tr>
<td>500</td>
<td>6.5410</td>
<td>6.2943</td>
<td>2.3379</td>
<td>0.9378</td>
<td>0.8585</td>
<td>9.7</td>
</tr>
</tbody>
</table>

Type I Error = 0.3I(n > m) + 0.4I(n ≤ m)
Figure 4-2. Sample partial autocorrelations and smoothed curves within each band of the metal data
Figure 4-3. Sample partial autocorrelations and smoothed curves within each band of the rock data.
Figure 4-4. Intensity plots for the metal and rock data
In this dissertation, we developed new methods to model the dependence structure in multivariate data based on re-parameterizing the correlation matrix using partial autocorrelations.

In Chapter 2, we proposed several priors for partial autocorrelation coefficients. These priors capture both decaying correlation as lag increases and place more weight on positive correlations, which is appreciate for longitudinal data. We transformed the partial autocorrelations using Fisher’s Z-transform, and introduced a generalized linear model framework to model partial autocorrelation (correlation) structure. Modeling structured partial autocorrelations greatly reduces the number of parameters that need to be estimated.

In Chapter 3, we investigated estimating correlation (covariance) matrices in high dimensions. Our approach was based on banding a partial autocorrelation matrix, and discovered that all lag $k_0$ partial autocorrelation coefficients are independent transformed Beta distributions on (-1, 1) with equal shape and scale parameters. This finding provided a foundation for doing sequential multiple hypothesis testing to estimate band using a Bonferroni correction. We also explored the relationship of type I error with sample size and dimension via simulations. Our estimation process is much faster than many existing methods and only requires inversion of matrices with dimension equal to the number of bands plus 1. We are working on proving consistency properties of our estimator of the correlation matrix, and the theory underlying our type I error correction.

In Chapter 4, we improved estimators of our banded partial autocorrelation matrix estimators discussed in Chapter 3 by nonparametric smoothing the partial autocorrelation functions within each band. We derive an asymptotically optimal bandwidth for kernel estimation of a partial autocorrelation function that provides a plug-in bandwidth estimator. Our theoretical derivation for the error is based on
equally spaced design points and a Toeplitz condition for the true partial autocorrelation matrices. Our simulation shows that the Toeplitz condition is not overly restrictive, which may due to the correlations between the partial autocorrelations within a band exponentially decaying to zero. We are currently exploring ways to weaken the Toeplitz condition for deriving the optimal bandwidth.
APPENDIX A
SUPPORTING MATERIAL FOR CHAPTER 2

A.1 Sampling Algorithm

We sample \((\beta, \eta, \gamma)\) using a (block) Gibbs sampler along with data augmentation for the missing responses (straightforward given the full data response is multivariate normal).

We sample the full conditional for the parameters as follows. At iteration \(k_0\),

1. Sample \(Y_{\text{imiss}}|(\beta^{(k_0-1)}, \gamma^{(k_0-1)}, \eta^{(k_0-1)}, Y_{\text{obs}})\) by data augmentation.
   \[
   Y_{\text{imiss}} \sim N(x_{\text{imiss}}^T \beta^{(k_0-1)} + \sum_{i=2}^{(k_0-1)} (\sum_{i=2}^{(k_0-1)}-1)(Y_{\text{obs}} - x_{\text{obs}}^T \beta^{(k_0-1)}),
   \]
   \[
   \sum_{i=2}^{(k_0-1)} - \sum_{i=2}^{(k_0-1)} (\sum_{i=2}^{(k_0-1)}-1) (\sum_{i=2}^{(k_0-1)}-1)
   \]
   where \(x_i^T = (x_{\text{obs}}^T, x_{\text{imiss}}^T)^T\), and \(\sum_{i=2}^{(k_0-1)} = 
   \left[
   \begin{array}{c}
   \sum_{i=2}^{(k_0-1)} \\
   \sum_{i=2}^{(k_0-1)}
   \end{array}
   \right].
   \]

2. Sample \(\beta|(\eta^{(k_0-1)}, \gamma^{(k_0-1)}, Y^{(k_0)})\) from a normal distribution with mean
   \[
   \mu^{(k_0-1)} = \sum_{i=1}^{(k_0-1)} \sum_{i=1}^{n} x_i D^{-1}_{i(\eta^{(k_0-1)})} R^{-1}_{i(\gamma^{(k_0-1)})} D^{-1}_{i(\gamma^{(k_0-1)})} x_i^T.
   \]
   and variance,
   \[
   \Sigma^{(k_0-1)} = \left(\sum_{i=1}^{n} x_i D^{-1}_{i(\eta^{(k_0-1)})} R^{-1}_{i(\gamma^{(k_0-1)})} D^{-1}_{i(\gamma^{(k_0-1)})} x_i^T\right)^{-1}.
   \]

3. Sample \(\eta|(\beta^{(k_0)}, \gamma^{(k_0-1)}, Y^{(k_0)})\) using a random walk Metropolis-Hastings algorithm. The full conditional is proportional to
   \[
   \pi(\eta|\beta, \gamma, Y, X) \propto \{(\prod_{i=1}^{n} |D_{i(\eta)}|^{-1}) \sum_{i=1}^{n} x_i D^{-1}_{i(\eta)} R^{-1}_{i(\gamma^{(k_0-1)})} D^{-1}_{i(\gamma^{(k_0-1)})} x_i^T \}^{-\frac{1}{2}}
   \]
   \[
   \exp\{-\frac{1}{2} \sum_{i=1}^{n} (y_i^{(k_0)} - x_i \beta^{(k_0)} D^{-1}_{i(\eta)} R^{-1}_{i(\gamma^{(k_0-1)})} D^{-1}_{i(\gamma^{(k_0-1)})} (y_i^{(k_0)} - x_i \beta^{(k_0)}))\}
   \]
   \[
   \exp(-\frac{1}{2} \eta^T \Sigma \eta).
   \]

4. Sample \(\gamma|(\beta^{(k_0)}, \eta^{(k_0)}, Y^{(k_0)})\) using a Quasi-Newton Metropolis-Hastings Algorithm (details in the web appendix). The full conditional is proportional to
\[ \pi(\gamma|\beta, \eta, Y, X) \]
\[ \propto \left( \prod_{i=1}^{\alpha} |R_{(i)}(\gamma)|^{-\frac{1}{2}} \right) \left| \sum_{i=1}^{\alpha} x_i^T D_{(i)(k)}^{-1} R_{(i)(k)}^{-1} D_{(i)(k)}^{-1} x_i \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{\alpha} (y_i - x_i^T \beta^{(k)}) \right\} \exp\left\{ -\frac{1}{2} (\gamma - \mu_\gamma)^T \Sigma^{-1} (\gamma - \mu_\gamma) \right\}. \]

**A.2 Simulating from the Full Conditional for \( \gamma \)**

For convenience, we denote log full conditional \( \pi(\gamma|\beta, \eta, Y, X) \) by \( \ell^*(\gamma) \). Let \( \gamma \) denote the mode of \( \ell^*(\gamma) \) and assume \( \gamma \) is in a neighborhood of \( \gamma_0 \). Then,

\[ \frac{\partial \ell^*(\gamma)}{\partial \gamma} = \frac{\partial \ell^*(\gamma)}{\partial \gamma} |_{\gamma=\gamma_0} + \frac{\partial^2 \ell^*(\gamma)}{\partial \gamma \partial \gamma^T} |_{\gamma=\gamma_0} (\gamma - \gamma_0) + o(|\gamma - \gamma_0|^2) 1_{q \times 1}. \]

Therefore,

\[ \gamma \approx \gamma_0 - \left( \frac{\partial^2 \ell^*(\gamma)}{\partial \gamma \partial \gamma^T} \right)^{-1} \frac{\partial \ell^*(\gamma)}{\partial \gamma} |_{\gamma=\gamma_0}. \]

We can approximate the observed information matrix \( -\frac{\partial^2 \ell^*(\gamma)}{\partial \gamma \partial \gamma^T} |_{\gamma=\gamma_0} \) by the expected fisher information matrix at \( \gamma_0 \). Using a quasi-Newton method to update \( \gamma \), \( \gamma = \gamma_0 + \alpha I(\gamma_0)^{-1} \nabla \ell^*(\gamma) |_{\gamma=\gamma_0} \), we choose \( \alpha \) to satisfy Wolf’s condition (Wolfe (1969)). The form of the expected information, \( I(\gamma) \) is given in the following section.

At iteration \( k_0 \), for Step 4 in our algorithm, we first approximate the model of \( \ell^*(\cdot) \), \( (\gamma^{(k_0)}) \) by \( \hat{\gamma}^{(k_0)} = \gamma^{(k_0-1)} + \alpha I(\gamma^{(k_0-1)})^{-1} \nabla \ell^*(\gamma) |_{\gamma=\gamma^{(k_0-1)}} \). Then we sample

\[ \gamma^{(k_0)} \sim N(\hat{\gamma}^{(k_0)}, I^{-1}(\hat{\gamma}^{(k_0)})), \tag{A-1} \]

and accept \( \gamma^{(k_0)} = \gamma^{(k_0)} \) with probability

\[ p = \min\{1, \frac{\pi(\gamma^{(k_0)})|\beta^{(k_0)}, \eta^{(k_0)}, Y^{(k_0)}, X)}{\pi(\gamma^{(k_0-1)})|\beta^{(k_0)}, \eta^{(k_0)}, Y^{(k_0)}, X)} \times \frac{h(\gamma^{(k_0-1)})|\beta^{(k_0)}, \eta^{(k_0)}, Y^{(k_0)}, X)}{h(\gamma^{(k_0)})|\beta^{(k_0)}, \eta^{(k_0)}, Y^{(k_0)}, X} \}. \tag{A-2} \]

where \( h(\cdot) \) is the pdf of Eq. A-1.

**A.3 Deriving the Expected Information Matrix for \( \gamma \)**

We derive the expected information matrix for \( \gamma \) (and for \( (\gamma, \sigma_0) \)), where \( \sigma_0 = (\sigma_{11}, ..., \sigma_{pp})^T \). To do this, we first define some needed quantities. Let \( A \) be a \( n \times n \)
symmetric matrix with elements \{a_{ij}\}, \(B\) be a \(m \times n\) matrix with elements \{b_{ij}\}, and \(C\) be a \(s \times s\) matrix with elements \{c_{ij}\} matrices. We define
\[
\text{vec}(A) = \left( a_{11} \ a_{21} \ \cdots \ a_{n1} \ a_{12} \ \cdots \ a_{n2} \ \cdots \ a_{nn} \right)^T
\]
to be a vector including all the elements in the matrix sorted by column. We define
\[
\nu(A) = \left( a_{11} \ a_{21} \ \cdots \ a_{n1} \ a_{22} \ \cdots \ a_{n2} \ \cdots \ a_{nn} \right)^T
\]
to be a vector including all the elements in the lower triangular part of a square matrix (sorted by column). Finally, we define
\[
\nu h(A) = \left( a_{21} \ a_{31} \ \cdots \ a_{n1} \ a_{32} \ \cdots \ a_{n2} \ \cdots \ a_{nn-1} \right)^T,
\]
to be a vector with all the elements in the lower triangular part of a square matrix (without the main diagonal). Define the Kronecker product of two square matrices as
\[
B \otimes C = (b_{ij} C).
\]
Now, define a matrix \(D_n\) such that \(D_n \nu(A) = \text{vec}(A)\). So, \(\nu(A) = D_n^\dagger \text{vec}(A)\), where \(D_n^\dagger\) is the Moore-Penrose inverse of \(D_n\), and

Let \(Y_1, Y_2, ..., Y_n\) be independently and identically distributed \(p \times 1\) random vector such that

\[
Y_i \sim N_p(\mu, \Sigma)
\]
where \(i = 1, 2, ..., n, \Sigma = (\sigma_{jk})\) is positive definite, and let \(n \geq p + 1\). The expected information matrix for \(\nu(\Sigma)\) is

\[
F_n = \frac{n}{2} D_p^T (\Sigma^{-1} \otimes \Sigma^{-1}) D_p
\]
(Magnus and Neudecker (1984))

To derive the expected Fisher information of \((\gamma, \sigma_0)\), we specify the following series of transformations,
\[
\sigma \rightarrow \sigma^* \rightarrow \rho \rightarrow \rho^* \rightarrow \pi \rightarrow z \rightarrow (\gamma, \gamma^*, \sigma_0)
\]

The transformations are defined as \(\sigma = (\nu(\Sigma))\), \(\sigma^* = (\nu h(\Sigma)^T, \sigma_0^T)^T\), \(\rho = (\nu h(R)^T, \sigma_0^T)^T\), \(\rho^* = (\rho_{12}, \rho_{23}, \rho_{13}, ..., \rho_{p-1p}, ..., \rho_{1p}, \sigma_0^T)^T\), \(\pi = (\pi_{12}, \pi_{23}, \pi_{13}, ..., \pi_{p-1p}, ..., \pi_{1p}, \sigma_0^T)^T\), and \(z = (z_{12}, z_{23}, z_{13}, ..., z_{p-1p}, ..., z_{1p}, \sigma_0)^T\).

Details on the Jacobian of each transformation follow.

1. \(g_1(\sigma; \sigma^*) : \sigma \rightarrow \sigma^*\) separates the diagonal and off-diagonal elements. The Jacobian matrix \(J^\rho\) of \(g_1(\sigma; \sigma^*)\) is obtained by re-ordering the \(\frac{p(p+1)}{2} \times \frac{p(p+1)}{2}\) dimensional identity matrix corresponding to this re-ordering transformation.

2. \(g_2(\sigma^*; \rho) : \sigma^* \rightarrow \rho\) is a 1-1 transformation from the covariance parameters to variance/correlation parameters. The Jacobian, \(J^\rho\) is

\[
J^\rho = \begin{pmatrix}
J_{11}^\rho & J_{12}^\rho \\
0 & I_{p \times p}
\end{pmatrix}
\]

where \(J_{11}^\rho = \text{diag}(\nu h(\sigma_0 \otimes \sigma_0^T)^{1/2})\) and

\[
J_{12}^\rho = \begin{pmatrix}
\nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & ... & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & 0 & ... & 0 & 0 \\
\nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & ... & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & 0 & ... & 0 & 0 \\
\nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & ... & \nu_{12} \sqrt{\nu_{11}^2 + \nu_{12}^2} & 0 & ... & 0 & 0 \\
... & ... & ... & ... & ... & ... & ... & ... & ... \\
0 & 0 & 0 & ... & \nu_{2p-1p-1} \nu_{2p-1p} & \nu_{2p-1p-1} \nu_{2p-1p} & \nu_{2p-1p-1} \nu_{2p-1p} & ... & \nu_{2p-1p-1} \nu_{2p-1p}
\end{pmatrix}
\]

3. \(g_3(\rho; \rho^*) : \rho \rightarrow \rho^*\) is a 1-1 transformation which changes the order of the parameters \(\rho\) to \(\rho^*\) with Jacobian,

\[
J^{\rho^*} = \begin{pmatrix}
J_{11}^{\rho^*} & 0 \\
0 & I_{p \times p}
\end{pmatrix},
\]

where \(J_{11}^{\rho^*}\) is a matrix obtained by reordering identity matrix \(I_{\frac{p(p-1)}{2}} \times \frac{p(p-1)}{2}\) corresponding to the reordering from \((\rho_{12}, ..., \rho_{1p}, \rho_{23}, ..., \rho_{2p}, ..., \rho_{p-1p})\), to \((\rho_{12}, \rho_{23}, \rho_{13}, ..., \rho_{p-1p}, ..., \rho_{1p})\).

4. \(g_4(\rho^*; \pi) : \rho^* \rightarrow \pi\) is a 1-1 transformation defined in (1) with \(\sigma_{ji}\) unchanged. The Jacobian is

\[
J^{\pi} = \begin{pmatrix}
J_{11}^{\pi} & 0 \\
0 & I_{p \times p}
\end{pmatrix}
\]
where \( J_{11}^\pi \) is Jacobian matrix of transformation from \((\rho_{12}, \rho_{23}, \rho_{13}, ..., \rho_{p-1p}, ..., \rho_{1p})\) to 
\((\pi_{12}, \pi_{23}, \pi_{13}, ..., \pi_{p-1p}, ..., \pi_{1p})\), which is a lower triangular matrix with elements 
\( \left\{ \frac{\partial \rho_{jk}}{\partial \pi_{lm}} \right\} \) in position \( \left( \frac{(l-1)(2p-j)}{2} + k, \frac{(l-1)(2p-l)}{2} + m \right) \).

5. \( g_5(\pi; z) : \pi \rightarrow z \) is a 1-1 transformation which transforms the \( \pi \) to \( z(\pi) \). The Jacobian is
\[
J^z = \begin{pmatrix} J_{11}^z & 0 \\ 0 & I_{p \times p} \end{pmatrix}
\]
where
\[
J_{11}^z = \begin{pmatrix}
1 - \pi_{12}^2 & 0 & 0 & \cdots & 0 \\
0 & 1 - \pi_{23}^2 & 0 & \cdots & 0 \\
0 & 0 & 1 - \pi_{13}^2 & \cdots & 0 \\
& & & \ddots & \\
0 & 0 & 0 & \cdots & 1 - \pi_{1p}^2 \\
\end{pmatrix}
\]

6. \( g_6(z; (\gamma, \gamma^\perp, \sigma_0)) : z \rightarrow (\gamma, \gamma^\perp, \sigma_0) \) is defined in Eq. 2–5 with \( \sigma_{ji} \) unchanged, and the Jacobian
\[
J(\gamma, \gamma^\perp, \sigma_0) = \begin{pmatrix} J_{11}^{(\gamma, \gamma^\perp)} & 0 \\ 0 & I_{p \times p} \end{pmatrix},
\]
where
\[
J_{11}^{(\gamma, \gamma^\perp)} = \begin{pmatrix} w^* & w^\perp \end{pmatrix}.
\]

Now, let \( I(\sigma^*) \) denote the expected information of \( \sigma^* \). Since the information matrix is invariant under transformation,
\[
I(\sigma^*) = J^{*T} \left\{ \frac{n}{2} [D_T^T (\Sigma^{-1} \bigotimes \Sigma^{-1}) D_T] \right\} J^* = \begin{pmatrix} I_{11}^{*} & I_{12}^{*} \\ I_{21}^{*} & I_{22}^{*} \end{pmatrix}.
\]
(\(A-3\))

Then,
\[
I(\rho) = J^{\rho T} I(\sigma^*) J^\rho
\]
\[
= \begin{pmatrix} J_{11}^{\rho} & J_{12}^{\rho} \\ 0 & I_{p \times p} \end{pmatrix}^T \begin{pmatrix} I_{11}^{*} & I_{12}^{*} \\ I_{21}^{*} & I_{22}^{*} \end{pmatrix} \begin{pmatrix} J_{11}^{\rho} & J_{12}^{\rho} \\ 0 & I_{p \times p} \end{pmatrix}
\]
\[
= \begin{pmatrix} J_{11}^{\rho T} I_{11}^{*} J_{11}^{\rho} & J_{11}^{\rho T} I_{12}^{*} J_{12}^{\rho} + J_{11}^{\rho T} I_{12}^{*} + I_{12}^{\rho T} I_{12}^{*} I_{12}^{\rho} \\ J_{12}^{\rho T} I_{11}^{*} J_{11}^{\rho} + I_{12}^{\rho T} I_{11}^{*} I_{12}^{\rho} & J_{12}^{\rho T} I_{12}^{*} J_{12}^{\rho} + (I_{21}^{*} J_{12}^{\rho} + J_{12}^{\rho T} I_{21}^{*}) + I_{22}^{*} \end{pmatrix}
\]

97
\[ l(\rho_0) = \begin{pmatrix} l(\rho_0) & l(\rho_0 \sigma_0) \\ l(\rho_0 \sigma_0) & l(\sigma_0) \end{pmatrix} \]

where \( \rho_0 = (\rho_{12}, ..., \rho_{1p}, \rho_{23}, ..., \rho_{p-1p}) \).

It then follows that

\[
I(\gamma, \gamma^+, \sigma_0) = (J(\gamma, \gamma^+, \sigma_0)^T J \gamma^T J^T J \gamma \sigma_0^T) l(\rho) (J \gamma^T J \gamma \sigma_0 T l(\rho) )
\]

\[
= \begin{pmatrix} (J_{11}(\gamma^+)^T J_{11} T J_{11} \gamma^T J_{11} \gamma T J_{11} \gamma^T l(\rho_0) )^T \\ (J_{11}(\gamma^+)^T J_{11} T J_{11} \gamma^T J_{11} \gamma^T l(\rho_0 \sigma_0) )^T \\ J_{11} \gamma^T J_{11} \gamma^T l(\rho_0 \sigma_0) \end{pmatrix} l(\sigma_0) \]

So the three blocks of the information matrix for \((\gamma, \sigma_0)\) are

\[
I(\gamma) = (w^T J_{11} T J_{11} \gamma^T J_{11} \gamma T J_{11} \gamma^T l(\rho_0) ) (J_{11} \gamma^T J_{11} \gamma T l(\sigma_0) ) \]

\[
I(\gamma, \sigma_0) = w^T J_{11} T J_{11} \gamma^T J_{11} \gamma^T l(\rho_0 \sigma_0) \]

\[
I(\sigma_0) = J_{12} \gamma^T J_{12} \gamma + (J_{12} \gamma^T J_{12} \gamma + J_{12} \gamma^T l(\sigma_0) + l(\sigma_0)^T \]

A.4 Proof of Theorem 1

Proof. Let \( Y_i^{k_i} = \{ Y_j, j = 1, ..., k_i \} \) where \( k_i : Q_{i;k_i} = 1, Q_{i;k_i+1} = 0 \), \( S_k = \{ i, Q_{i;k_i} = 1 \) and \( Q_{i;k_i+1} = 0 \), and \( k_i = k \), where \( 1 \leq k \leq p - 1 \); \( i = 1, ..., n \) and \( S_\beta, S_\gamma, S_\eta \) be sample spaces of \( \beta, \gamma, \eta \), respectively. Thus the observed data distribution of \( ith \) subject is \( Y_i^{k_i} \sim N_k(x_i^{k_i T} \beta, \Sigma_i^{k_i}) \), where \( x_i^{k_i} = x_i[i;1:k_i] \) is a \( p_\beta \times k_i \) submatrix of \( x_i \) and \( \Sigma_i^{k_i} = \Sigma_i[i;1:k_i;1:k_i] \) is a \( k_i \times k_i \) principal submatrix of \( \Sigma_i \). Define the observed data, \( Y_{obs} = (Y_1^{k_1}, ..., Y_n^{k_n}) \).

Therefore,

\[
f(y_i^{k_i} | \beta, \gamma, \eta, x_i^{k_i}) \propto |\Sigma_i^{k_i}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(y_i^{k_i} - x_i^{k_i T} \beta)^T (\Sigma_i^{k_i})^{-1}(y_i^{k_i} - x_i^{k_i T} \beta)],
\]

and

\[
f(y_{obs} | \gamma, \beta, \eta, x) \propto \prod_{i=1}^{n} |\Sigma_i^{k_i}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\sum_{i=1}^{n}(y_i^{k_i} - x_i^{k_i T} \beta)^T (\Sigma_i^{k_i})^{-1}(y_i^{k_i} - x_i^{k_i T} \beta))]
\]
\[
\begin{align*}
&= \prod_{k=1}^{p-1} \prod_{i \in S_k} |(\Sigma_i^k)|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \sum_{k=1}^{p-1} \sum_{i \in S_k} (y_i - x_i^k \beta)^T (\Sigma_i^k)^{-1} (y_i - x_i^k \beta)\} \\
&= \prod_{k=1}^{p-1} \prod_{i \in S_k} |(\Sigma_i^k)|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \sum_{k=1}^{p-1} \sum_{i \in S_k} (y_i^k T \beta)^T (\Sigma_i^k)^{-1} y_i^k \} \\
&\quad - 2 \sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T \beta + \beta^T \sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T \beta} \} \pi(\gamma) \pi(\eta)
\end{align*}
\]

Therefore, posterior distribution of \((\beta, \gamma, \eta)\) is

\[
\pi(\beta, \gamma, \eta|y_{\text{obs}}, x) = \frac{m(\beta, \gamma, \eta|y_{\text{obs}}, x)}{\int_{\beta \in \mathcal{B}} \int_{\gamma \in \mathcal{G}} \int_{\eta \in \mathcal{H}} m(\beta, \gamma, \eta|y_{\text{obs}}, x) d\beta d\gamma d\eta},
\]

where

\[
m(\beta, \gamma, \eta|y_{\text{obs}}, x) = \prod_{k=1}^{p-1} \prod_{i \in S_k} |(\Sigma_i^k)|^{-\frac{1}{2}} \exp\{-\frac{1}{2} \sum_{k=1}^{p-1} \sum_{i \in S_k} (y_i^k T \beta)^T (\Sigma_i^k)^{-1} y_i^k \} - 2 \sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T \beta + \beta^T \sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T \beta} \} \pi(\gamma) \pi(\eta).
\]

Since \(\sum x_i^k x_i^k T\) is full rank, \(\sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T\) is a positive-definite matrix.

Therefore, its smallest eigenvalue \(\lambda_p\) is larger than zero and

\[
|\sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T|^{-\frac{1}{2}} < \lambda_p^{-\frac{\theta}{2}}.
\]

Define \(\hat{\beta} = (\sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T)^{-1} (\sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} y_i^k)\). We obtain,

\[
\int_{\mathcal{B}} m(\beta, \gamma, \sigma|y_{\text{obs}}, x) d\beta \\
\propto |\sum_{k=1}^{p-1} \sum_{i \in S_k} x_i^k (\Sigma_i^k)^{-1} x_i^k T|^{-\frac{1}{2}} (\prod_{k=1}^{p-1} \prod_{i \in S_k} |(\Sigma_i^k)|^{-\frac{1}{2}}) \\
\exp\{-\frac{1}{2} \sum_{k=1}^{p-1} \sum_{i \in S_k} (y_i^k - \hat{\beta} x_i^k T \beta)^T (\Sigma_i^k)^{-1} (y_i^k - \hat{\beta} x_i^k T \beta)\} \pi(\gamma) \pi(\eta) \\
\leq \lambda_p^{-\frac{\theta}{2}} (\prod_{k=1}^{p-1} \prod_{i \in S_k} |(\Sigma_i^k)|^{-\frac{1}{2}}) \exp\{-\frac{1}{2} \sum_{k=1}^{p-1} \sum_{i \in S_k} (y_i^k - \hat{\beta} x_i^k T \beta)^T (\Sigma_i^k)^{-1} (y_i^k - \hat{\beta} x_i^k T \beta)\} \pi(\gamma) \pi(\eta).
\]
Now, define $M_i^{k_i}(\beta) = (y_i^{k_i} - x_i^{k_i} T \beta)(y_i^{k_i} - x_i^{k_i} T \beta)^T$. The only positive eigenvalue of $M_i^{k_i}(\beta)$ is $\lambda_1(M_i^{k_i}(\beta)) = (y_i^{k_i} - x_i^{k_i} T \beta)(y_i^{k_i} - x_i^{k_i} T \beta) > 0$ (Marshall and Olkin (1979)).

Similar to Daniels (2006), we remove the dependence of $M_i^{k_i}(\beta)$ on $\Sigma_i^{k_i}$ by bounding the exponential term. Re-write the exponential term in the above expression as:

$$(y_i^{k_i} - x_i^{k_i} T \beta)^T (\Sigma_i^{k_i})^{-1}(y_i^{k_i} - x_i^{k_i} T \beta)$$

$$= trace[(\Sigma_i^{k_i})^{-1} M_i^{k_i}(\beta)]$$

$$\geq \sum_{t=1}^{k} \lambda_t(\Sigma_i^{k_i}) \lambda_{k-t+1}(M_i^{k_i}(\beta))$$

$$= \lambda_k(\Sigma_i^{k_i}) \lambda_1(M_i^{k_i}(\beta))$$

where $\lambda_t(\bullet)$, defined as $\lambda_1(A)$, $\lambda_2(A)$, ..., $\lambda_p(A)$ are the ordered eigenvalues of a $p \times p$ matrix $A$. The first inequality is from dimciteMarshallOlkin1979.

Let $\lambda_{\text{min},k} = \min_{i \in S_k} \{\lambda_k(\Sigma_i^{k_i})\} > 0$ and $\lambda^{(M)}_{\text{min},k} = \min_{i \in S_k} \{\lambda_1(M_i^{k_i})\} > 0$. Then,

$$\sum_{i \in S_k} (y_i^{k_i} - x_i^{k_i} T \beta)^T (\Sigma_i^{k_i})^{-1}(y_i^{k_i} - x_i^{k_i} T \beta)$$

$$\geq \sum_{i \in S_k} \lambda_k(\Sigma_i^{k_i}) \lambda_1(M_i^{k_i})$$

$$\geq s_k \lambda_{\text{min},k} \lambda^{(M)}_{\text{min},k}$$

$$= \text{trace}\left\{ s_k \lambda_{\text{min},k} l_k \right\}$$

where $s_k$ denotes cardinality of set $S_k$ and $l_k$ is a $k \times k$ identity matrix. Finally, for each $i$, simulate a 'new' set of data, $y_i^{k_i}$ from a multivariate normal distribution such that

$\sum_{i \in S_k} y_i^{k_i} y_i^{k_i} T = \lambda^{(M)}_{\text{min},k} l_k$.

We then obtain

$$\left(\prod_{k=1}^{p} \prod_{i \in S_k} |(\Sigma_i^{k_i})|^{\frac{1}{2}} \right) \exp\left\{-\frac{1}{2} \sum_{k=1}^{p} \sum_{i \in S_k} (y_i^{k_i} - x_i^{k_i} T \beta)^T (\Sigma_i^{k_i})^{-1}(y_i^{k_i} - x_i^{k_i} T \beta)\right\} \pi(\gamma) \pi(\eta)$$

100
\[
\begin{align*}
&\leq (\prod_{k=1}^{p-1} \prod_{i \in S_k} [\prod_{t=1}^{\pi_k} \lambda_t (\sum_{i}^{-k})]^{\frac{1}{2}}) \exp\left\{-\frac{1}{2} \sum_{k=1}^{p-1} \sum_{i \in S_k} y_i^{*k_i} T \frac{S_k \lambda_{\text{min}, k}}{k} l_k y_i^{*k_i}\right\} \pi(\gamma) \pi(\eta) \\
&= (\prod_{k=1}^{p-1} \prod_{i \in S_k} [\lambda_t (\sum_{i}^{-k})]^{\frac{1}{2}} [\frac{S_k \lambda_{\text{min}, k}}{k}]^{\frac{1}{2}}) \exp\left\{-\frac{1}{2} \sum_{k=1}^{p-1} \sum_{i \in S_k} y_i^{*k_i} T \frac{S_k \lambda_{\text{min}, k}}{k} l_k y_i^{*k_i}\right\} \pi(\gamma) \pi(\eta)
\end{align*}
\]
\[
\leq M_0
\]

where \(M_0\) is a finite constant since all three terms above are bounded. Therefore,

\[
\int_{\mathcal{A}_\beta} m(\beta, \gamma, \sigma | y_{\text{obs}}, x) d\beta \text{ is finite.}
\]

Since the priors on \(\gamma, \eta\) are proper under the assumption that \(\sum A_i A_i^T\) and \(\sum w_i w_i^T\) are full rank, we have

\[
\int_{\mathcal{A}_\beta} \int_{\mathcal{A}_\gamma} \int_{\mathcal{A}_\eta} m(\beta, \gamma, \eta | y_{\text{obs}}, x) d\beta d\gamma d\eta \leq \int_{\mathcal{A}_\gamma} \int_{\mathcal{A}_\eta} M_0 \pi(\gamma) \pi(\eta) d\gamma d\eta < \infty.
\]

Hence, the posterior of \((\beta, \gamma, \eta)\) is proper. \(\square\)
**APPENDIX B**

**SUPPORTING MATERIAL FOR CHAPTER 3**

**B.1 Proof of Lemma 3.0.1**

First, we partition $R_{p\times p}$ as follows,

$$R_{p\times p} = \begin{pmatrix} R_{p-1\times p-1} & r_1 \\ r_1^T & \rho_{p,p} \end{pmatrix},$$

where $R_{p-1\times p-1} = R_{[1:p-1\times p-1]}$ and $r_1 = R_{[1,1:p-1]} = (\rho_{1,p}, r_2^T)^T$.

By Fact 1,

$$\det(R_{p\times p}) = \det(R_{p-1\times p-1})(1 - r_1^T R_{p-1\times p-1}^-1 r_1).$$

Furthermore, we can partition $R_{p-1\times p-1}$ in the following way,

$$R_{(p-1)\times(p-1)} = \begin{pmatrix} \rho_{1,1} & \nu_2^T \\ \nu_2 & R_2 \end{pmatrix},$$

where $\nu_2 = R_{p\times p}^T[1, 2 : p - 1]$ and $R_2 = R_{p\times p}[2 : p-2, 2 : p-2]$.

So, $\det(R_{(p-1)\times(p-1)}) = (1 - \nu_2^T R_2^{-1} \nu_2) \det(R_2)$. Now, let $A_{11\cdot 2} = 1 - \nu_2^T R_2^{-1} \nu_2$.

Then,

$$R_{(p-1)\times(p-1)}^{-1} = \begin{pmatrix} A_{11\cdot 2}^{-1} & -A_{11\cdot 2}^{-1} \nu_2^T R_2^{-1} \\ -R_2^{-1} \nu_2 A_{11\cdot 2}^{-1} & R_2^{-1} \nu_2 A_{11\cdot 2}^{-1} \nu_2^T R_2^{-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & R_2^{-1} \end{pmatrix} = A_{11\cdot 2}^{-1} \begin{pmatrix} 1 & -\nu_2^T R_2^{-1} \\ -R_2^{-1} \nu_2 & R_2^{-1} \nu_2^T R_2^{-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & R_2^{-1} \end{pmatrix}. \quad (B-1)$$

We now show

$$r_1^T R_{(p-1)\times(p-1)} r_1$$

$$= \begin{pmatrix} \rho_{1,p} & r_2^T \end{pmatrix} \begin{pmatrix} A_{11\cdot 2}^{-1} \begin{pmatrix} 1 & -\nu_2^T R_2^{-1} \\ -R_2^{-1} \nu_2 & R_2^{-1} \nu_2^T R_2^{-1} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & R_2^{-1} \end{pmatrix} \end{pmatrix} \begin{pmatrix} \rho_{1,p} \\ r_2 \end{pmatrix}$$

$$= A_{11\cdot 2}^{-1} \cdot \left[ \rho_{1,p}^2 - 2 \rho_{1,p} \cdot r_2^T R_2^{-1} \nu_2 + (r_2^T R_2^{-1} \nu_2)^2 \right] + r_2 R_2^{-1} r_2^T.$$
\[ = A_{11}^{-1} \cdot (\rho_{1,p}^2 - \nu_2^T R_2^{-1} r_2)^2 + r_2^T R_2^{-1} r_2. \]

Therefore,

\[ \det (R_{p \times p}) \]
\[ = (1 - \nu_2^T R_2^{-1} \nu_2) \det (R_2) \cdot \{1 - r_2^T R_2^{-1} r_2 - A_{11}^{-1} \cdot (\rho_{1,p}^2 - \nu_2^T R_2^{-1} r_2)^2\} \]
\[ = (1 - \nu_2^T R_2^{-1} \nu_2)(1 - r_2^T R_2^{-1} r_2) \det (R_2) \cdot \left[1 - \left(\frac{\rho_{1,p}^2 - \nu_2^T R_2^{-1} r_2}{\sqrt{(1 - \nu_2^T R_2^{-1} \nu_2)(1 - r_2^T R_2^{-1} r_2)}}\right)^2\right] \]
\[ = (1 - \nu_2^T R_2^{-1} \nu_2)(1 - r_2^T R_2^{-1} r_2) \det (R_2) \cdot (1 - \pi_{1,p}). \]

B.2 Preliminaries and Notation for Proof of Theorem 3.1

Before proving Theorem 3.1, we introduce some notation and derive some preliminary results. Let \( \{Y_i : i = 1, ..., n\} \) be \( p \times 1 \) vectors of independent, normally distributed random variables with mean 0 and covariance matrix \( \Sigma \). The mle of \( \Sigma \) is

\[ S_{\Sigma} = \frac{1}{n} \sum_{i=1}^n Y_i Y_i'. \]

Applying the transformation \( X_i = T Y_i \), where \( T = \text{diag}(s_{j,j}^{-\frac{1}{2}}) \), we define

\[ S = \frac{1}{n} \sum_{i=1}^n X_i X_i' = (s_{j,k}). \]

The likelihood for \( \pi_{jk} \), \( G(\pi_{jk}) \) is proportional to

\[ G(\pi_{jk}) \propto (1 - \pi_{jk}^2)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2} \text{tr}(R^{-1}[j : k] S[j : k])\right\}. \]

The corresponding log likelihood can be re-written as

\[ \log G(\pi_{jk}) = g(\pi_{jk}) \propto -\frac{n}{2} \log(1 - \pi_{jk}^2) - \frac{n}{2 \det (R[j : k])} \text{tr}(A[j : k] S[j : k]) \]

where \( A[j : k] \) is the adjoint matrix of \( R[j : k] \) and \( R^{-1}[j : k] = \frac{1}{\det(R[j : k])} A[j : k] \). So \( A[j : k] \)

is a quadratic function of \( \rho_{jk} \), i.e. \( A[j : k] = A_0 + A_1 \rho_{jk} + A_2 \rho_{jk}^2 \), where \( A_0, A_1, A_2 \) are \( (k - j + 1) \times (k - j + 1) \) matrices.

The first derivative of the log likelihood for \( \pi_{jk} \) is

\[ \frac{\partial g(\pi_{jk})}{\partial \pi_{jk}} \]
\[ \begin{align*}
&= \frac{n}{2} \left\{ D_{jk}(1 - \pi_{j,k}^2) \left[ \text{tr}(A_1 S[j : k]) + 2 \rho_{j,k} \text{tr}(A_2 S[j : k]) \right] \\
&\quad + 2 \pi_{j,k} \text{tr}(A[j : k] S[j : k]) \right\} + \left\{ \frac{n}{2} \left\{ D_{jk}(1 - \pi_{j,k}^2) \left[ \text{tr}(A_1 S[j : k]) + 2 \rho_{j,k} \text{tr}(A_2 S[j : k]) \right] \\
&\quad + 2 \pi_{j,k} \text{tr}(A_0 S[j : k]) + \rho_{j,k} \text{tr}(A_1 S[j : k]) + \rho_{j,k}^2 \text{tr}(A_2 S[j : k]) \right\} \right\}.
\end{align*} \]

where \( a = \det(R[j : k])/(1 - \pi_{j,k}^2) \), which is not a function of \( \pi_{j,k} \) (cf: Result 3).

Based on the following partition for \( R_{p \times p} \),

\[
R_{p \times p} = \begin{pmatrix}
1 & \nu_2^T & \rho_{1,p} \\
\nu_2 & R_2 & r_2 \\
\rho_{p,1} & r_2^T & 1
\end{pmatrix},
\]

where \( R_2 = R_{p \times p}[2 : p - 1, 2 : p - 1] \), \( \nu_2 = R_{p \times p}^T[1, 2 : p - 1] \), \( r_2 = R_{p \times p}[2 : p - 1, p] \), we obtain

\[
A_1 = \begin{pmatrix}
0 & r_2^T R_2^{-1} \cdot \det(R_2) & - \det(R_2) \\
R_2^{-1} r_2 \cdot \det(R_2) & B & R_2^{-1} \nu_2 \cdot \det(R_2) \\
- \det(R_2) & \nu_2^T R_2^{-1} \cdot \det(R_2) & 0
\end{pmatrix},
\]

where \( B_{(i-1,j-1)} = \{ (-1)^{i+j} \nu_{2(i)}^T M_{j}^{-1} r_{2(j)} + (-1)^{i+j} \nu_{2(j)}^T M_{i}^{-1} r_{2(i)} \} \cdot \det(M_{ij}), i, j \in \{2, 3, \ldots, p\} \), \( M_{ij} \) is the adjoint matrix of \( R_{2(j)} \) (note: \( R_{2(j)} \) is the submatrix of \( R_2 \) obtained by deleting the \( j \)-th row and \( i \)-th column), \( \nu_{2(i)} \) is the matrix \( \nu_2 \) with the \( i \)-th element deleted, \( r_{2(j)} \) is the matrix \( r_2 \) with the \( j \)-th element deleted and

\[
A_2 = -\begin{pmatrix}
0 & 0 & 0 \\
0 & R_2^{-1} \cdot \det(R_2) & 0 \\
0 & 0 & 0
\end{pmatrix}.
\]
B.3 Proof of Theorem 3.1

Assume \( \hat{\pi} = (\hat{\pi}_{1,2}, \hat{\pi}_{2,3}, \ldots, \hat{\pi}_{p-1,p}, \hat{\pi}_{1,p-1}, \hat{\pi}_{2,p}, \hat{\pi}_{1,p}) \) is mle of multivariate normal likelihood function. We are going to prove it is equal to \( \tilde{\pi} \) in Eq. 3–7.

Let \( \hat{\rho} = (\hat{\rho}_{1,2}, \hat{\rho}_{2,3}, \ldots, \hat{\rho}_{p-1,p}, \hat{\rho}_{1,p-1}, \hat{\rho}_{2,p}, \hat{\rho}_{1,p}) \) be the corresponding mle of the correlation coefficients. Since the MLE of \( \rho_{i,j} \) is \( s_{i,j} \), under the multivariate normal likelihood, we only need to prove that \( s_{i,j} \) is the estimator of \( \rho_{i,j} \), \( \tilde{\rho}_{i,j} \) from sequentially maximizing the objective functions in Eq. 3–6. We do this by induction.

1) For \( i=1 \), the lag 1 estimator of \( \rho_{t,t+i}, t = 1, 2, \ldots, p \), is obviously \( s_{t,t+i} \).

2) Suppose it is true for \( i = k, k \in \{2, 3, \ldots, p - 1\} \). Then we have objective function estimator \( \tilde{\rho}_{t,t+k} = s_{t,t+k} \) for \( t = 1, 2, \ldots, p - k, \) and \( t < k \leq p - 1 \).

3) For \( i = k + 1 \), to prove objective function estimator of \( \tilde{\rho}_{t,t+i} \) from \( \tilde{\pi}_{t,t+i} \) is \( s_{t,t+i} \), we only need to prove the corresponding \( \tilde{\pi}_{t,t+i} \) is the solution of the derivative of the corresponding terms from Eq. 3–6, i.e. \( \frac{\partial g(\pi_{t,t+i})}{\partial \pi_{t,t+i}} = 0 \).

Without loss of generality, let \( t=1 \). We need to show \( \tilde{\pi}_{1,k+1} \) satisfies

\[
-2\pi_{1,k+1}(1 - \pi_{k+1}^2) + \frac{1}{a}(D_{k+1}(1 - \pi_{k+1}^2))
\]

\[
[A_{1:k+1} + 2\pi_{1,k+1}(A_{1}S_{1,k+1})] + 2\pi_{1,k+1} \text{det} (R_{1:k+1} - A_{1:k+1}S_{1,k+1}) = 0 \quad (B-2)
\]

where \( S_{1:k+1} = S[1 : k + 1, 1 : k + 1], \)
 \( R_{1:k+1} = R[1 : k + 1, 1 : k + 1], \) and \( A_{1:k+1} \) is the adjoint matrix of \( R_{1:k+1} \). The submatrix of the objective function estimator of the correlation matrix:

\[
\tilde{R}_{(1:k+1)} =
\]
Plugging $\widehat{\pi}_{1,k+1}$ into Eq. B–3, $\widehat{R}_{[1:k+1]} = S_{1k+1}$. Now, we simplify terms in Eq. B.3,

\[
tr(A_1S) = 4\widehat{\rho}_2^T \widehat{R}_2^{-1} \widehat{r}_2 \det(\widehat{R}_2) - 2s_{(1,k+1)} \cdot \det(\widehat{R}_2) + 2(k-2)\widehat{\rho}_2^T \widehat{R}_2^{-1} \widehat{r}_2 \cdot \det(\widehat{R}_2)
\]

\[
= 2[-s_{(1,k+1)} + k \cdot \widehat{\rho}_2^T \widehat{R}_2^{-1} \widehat{r}_2] \cdot \det(\widehat{R}_2)
\]

\[
tr(A_2S) = -\det(\widehat{R}_2)tr(\widehat{R}_2^{-1}S_2) = -(k-1) \det(\widehat{R}_2).
\]

and

\[
tr(A_1S_{1k+1}) + 2\rho_{(1,k+1)}tr(A_2S_{1k+1})
\]

\[
= 2[-s_{(1,k+1)} + k \cdot \widehat{\rho}_2^T \widehat{R}_2^{-1} \widehat{r}_2] \cdot \det(\widehat{R}_2) - 2(k-1)s_{(1,k+1)} \det(\widehat{R}_2)
\]

\[
= -2k[s_{(1,k+1)} - \widehat{\rho}_2^T \widehat{R}_2^{-1} \widehat{r}_2] \det \widehat{R}_2
\]

\[
tr(A_{[1:k+1]}S_{1k+1}) = (k+1) \det(\widehat{R}_{[1:k+1]})
\]

\[
= (k+1)\widehat{\alpha}(1 - \widehat{\pi}_{1,k+1}^2)
\]

\[
D_{1k+1}^2 = [1 - \widehat{\rho}_2^T \widehat{R}_2^{-1} \widehat{r}_2][1 - \widehat{\rho}_2^T \widehat{R}_2^{-1} \widehat{r}_2].
\]

So,

\[
\frac{\partial g(\pi_{t, t+i})}{\partial \pi_{t, t+i}} = -2\widehat{\pi}_{1,k+1}(1 - \widehat{\pi}_{1,1+k}^2) + \frac{1}{a}[\widehat{D}_{1k+1}[tr(\widehat{A}_1S_{1k+1}) + 2\widehat{\rho}_{1,k+1}tr(\widehat{A}_2S_{1k+1})]
\]

\[
+ 2\widehat{\pi}_{1,k+1}tr(\widehat{R}_{[1:k+1]}^{-1}S_{1k+1})]
\]

\[
= (1 - \widehat{\pi}_{1,1+k}^2) \cdot \{ -2\widehat{\pi}_{1,k+1}
\]
\[
\begin{align*}
+ \frac{-2\hat{D}_{ik+1} \det(\hat{R}_2)}{\hat{a}} [s_{(1,k+1)} - 2k \cdot \hat{v}_2^T \hat{R}_2^{-1} \hat{r}_2 + 2(2-k)s_{(1,k+1)}] \\
+ 2(k+1)\hat{\pi}_{1,1+k} \\
= (1 - \hat{\pi}_{1,1+k}^2) \times \{2k\hat{\pi}_{1,k+1} + \frac{-2k\hat{D}_{ik+1} \det(\hat{R}_2)}{\hat{a}} \\
\times \frac{s_{(1,k+1)} - \hat{v}_2^T \hat{R}_2^{-1} \hat{r}_2}{\sqrt{[1 - \hat{v}_2^T \hat{R}_2^{-1} \hat{v}_2][1 - \hat{r}_2^T (\hat{R}_2^{-1} \hat{r}_2)]}} \} \\
= (1 - \hat{\pi}_{1,1+k}^2) \{2k\hat{\pi}_{1,k+1} - 2k\hat{\pi}_{1,k+1} \}
\end{align*}
\]

Similarly, we can prove that it is true for \( \{\pi_{i,k+1}, i \in \{2, 3, \ldots, p - k - 1\}\} \). Therefore, by induction it is true.

Now define \( \hat{\pi}^{1,\rho} = (\hat{\pi}_{1,2}, \hat{\pi}_{2,3}, \ldots, \hat{\pi}_{p-1,p}, \ldots, \hat{\pi}_{1,p-1}, \hat{\pi}_{2,p}, \hat{\pi}_{1,p}) \) as in Eq. 3–7 (maximizer of the objective function). We are going to prove it is also mle of the likelihood function.

Let \( \hat{\rho}^{1,\rho} = (\hat{\rho}_{1,2}, \hat{\rho}_{2,3}, \ldots, \hat{\rho}_{p-1,p}, \ldots, \hat{\rho}_{1,p-1}, \hat{\rho}_{2,p}, \hat{\rho}_{1,p}) \) be the corresponding estimators of the correlation coefficients. We again use an induction argument.

1) For \( k = 1 \), \( \hat{\pi}_{j,j+1} \), the maximizer of \( G(\pi_{j,j+1}) \), is also the mle of the corresponding correlation coefficients \( \rho_{j,j+1} = \hat{\pi}_{j,j+1} \).

2) assume it is also true for \( k = t \in \{1, \ldots, p - 2\} \). That is

\[
\hat{\pi}^{j+1} = (\hat{\pi}_{j+1,1}, \hat{\pi}_{j+1,2}, \ldots, \hat{\pi}_{j+1,j}, \ldots, \hat{\pi}_{j+1,j-t+1}, \hat{\pi}_{j+1,j-t}, \hat{\pi}_{j+1,j})
\]

is the maximizer of the multivariate normal likelihood with \( j = 1, 2, \ldots, p - t \).

Therefore, the corresponding correlation coefficient estimators

\[
\hat{\rho}^{j+1} = (\hat{\rho}_{j+1,1}, \hat{\rho}_{j+1,2}, \ldots, \hat{\rho}_{j+1,j}, \ldots, \hat{\rho}_{j+1,j-t+1}, \hat{\rho}_{j+1,j-t}, \hat{\rho}_{j+1,j})
\]

are the maximizers of the multivariate normal likelihood on the \( \rho \) scale, \( \hat{\rho}^{j+1} = \hat{\rho}^{j+1} \).

3) for \( k = t+1 \), let \( \hat{\pi}_{j,j+t+1} = \{\hat{\pi}_{j,j+1}, \ldots, \hat{\pi}_{j,j+t+1}, \hat{\rho}_{j,j+t}, \hat{\rho}_{j,j+t+1}, \hat{\rho}_{j,j+t+1}\} \),

and \( \hat{\pi}_{j,j+t+1} = \{\hat{\pi}_{j,j+1}, \ldots, \hat{\pi}_{j,j+t+1}, \hat{\rho}_{j,j+t}, \hat{\rho}_{j,j+t+1}, \hat{\rho}_{j,j+t+1}\} \).
Since
\[
\tilde{\rho}_{j,j+t+1} = (\tilde{\rho}_{j,j+1}, \tilde{\rho}_{j+1,j+2}, \ldots, \tilde{\rho}_{j+t,j+t+1}, \ldots, \tilde{\rho}_{j,j+t+1})
\]
\[
= (\hat{\rho}_{j,j+1}, \hat{\rho}_{j+1,j+2}, \ldots, \hat{\rho}_{j+t,j+t+1}, \ldots, \hat{\rho}_{j,j+t+1}, \tilde{\rho}_{j,j+t+1})
\]
\[
= \{\hat{\rho}_{j,j+t+1}, \tilde{\rho}_{j,j+t+1}\}
\]
and
\[
\hat{\rho}_{j,j+t+1} = (\hat{\rho}_{j,j+1}, \hat{\rho}_{j+1,j+2}, \ldots, \hat{\rho}_{j+t,j+t+1}, \ldots, \hat{\rho}_{j,j+t+1})
\]
\[
= \{\hat{\rho}_{j,j+t+1}, \tilde{\rho}_{j,j+t+1}\}
\]
is the maximizer of \(L(\prod_{j,j+t+1})\).

Moreover,
\[
\tilde{\nu}_{j,j+t+1} = \max_{\nu} \{G(\nu_{j,j+t+1})\} = \max_{\nu} \{L(\prod_{j,j+t+1} | \hat{\prod}_{j,j+t+1})\}
\]
\[
= \hat{\nu}_{j,j+t+1}.
\]

Therefore,
\[
\tilde{\rho}_{j,j+t+1} = \hat{\rho}_{j,j+t+1}.
\]

Correspondingly,
\[
\tilde{\nu}_{j,j+t+1} = \max_{\nu} \{G(\nu_{j,j+t+1})\} = \max_{\nu} \{L(\prod_{j,j+t+1} | \hat{\prod}_{j,j+t+1})\}
\]
\[
= \hat{\nu}_{j,j+t+1}.
\]

Hence, by induction, \(\hat{\nu}_{1j}^{IP}\) is mle of the likelihood.
Proof of Result 4. First, let \( R = (\rho_{j,k})_{p \times p} \) be a correlation matrix and denote \( R_i = R^T_{[i:p-1]} \) and \( r_i = R_{[i:p-1],} \).

We partition \( R \) as
\[
R = \begin{pmatrix}
1 & \rho_{1,2} & \cdots & \rho_{1,p-1} & \rho_{1,p} \\
\rho_{2,1} & 1 & \cdots & \rho_{2,p-1} & \rho_{2,p} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\rho_{p-1,1} & \rho_{p-1,2} & \cdots & 1 & \rho_{p-1,p} \\
\rho_{p,1} & \rho_{p,2} & \cdots & \rho_{p,p-1} & 1
\end{pmatrix} = \begin{pmatrix}
R_1 & r_1 \\
r_1^T & 1_{1 \times 1}
\end{pmatrix},
\]

where \( r_1^T = \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \rho_{p-1} & \rho_p \end{pmatrix}^T = \begin{pmatrix} \rho_1 & r_2^T \end{pmatrix}^T \).

Thus, by Result 1
\[
R^{-1} = A_{22 \cdot 1}^{-1} \times \begin{pmatrix}
R_1^{-1}r_1r_1^TR_1^{-1} & -R_1^{-1}r_1 \\
-R_1^{-1}r_1^TR_1^{-1} & 1_{1 \times 1}
\end{pmatrix} = \begin{pmatrix}
R_1 & 0_{(p-1) \times 1} \\
0_{1 \times (p-1)} & 0_{1 \times 1}
\end{pmatrix},
\]

where \( A_{22 \cdot 1} = 1 - r_1^TR_1^{-1}r_1 = \prod_{j=1}^{p-1}(1 - \pi_{j,p}^2), \) and \( \det(R) = A_{22 \cdot 1} \det(R_1) \).

Furthermore, partition \( R_1 \) as
\[
R_1 = \begin{pmatrix}
1_{1 \times 1} & \nu_2^T \\
\nu_2 & R_2
\end{pmatrix}.
\]

We can then show that
\[
R_1 = A_{11 \cdot 2}^{-1} \left( \begin{pmatrix}
1_{1 \times 1} & -\nu_2^T R_2^{-1} \\
-R_2^{-1} \nu_2 & R_2^{-1} \nu_2 v_2^T R_2^{-1}
\end{pmatrix} + \begin{pmatrix}
0_{1 \times 1} & 0_{1 \times (p-2)} \\
0_{(p-2) \times 1} & R_2^{-1}
\end{pmatrix} \right),
\]

\[
R_1 r_1 = \frac{\rho_{1,p} - r_2^T R_2^{-1} r_2}{1 - \nu_2^T R_2^{-1} \nu_2} \begin{pmatrix}
1_{1 \times (p-2)} \\
-R_2^{-1} \nu_2
\end{pmatrix} + \begin{pmatrix}
0_{1 \times (p-2)} \\
R_2^{-1} r_2
\end{pmatrix}.
\]
\[
\pi_1, \rho \sqrt{1 - r_2^T R_2^{-1} r_2 \over 1 - \nu_2^T R_2^{-1} \nu_2} \begin{pmatrix}
1_{1 \times (p-2)} \\
R_2^{-1} \nu_2
\end{pmatrix} + \begin{pmatrix}
0_{1 \times (p-2)} \\
R_2^{-1} r_2
\end{pmatrix}
\]

\[
R_1^{-1} r_1^T R_1^{-1} = \pi_1, \rho \cdot \frac{1 - r_2^T R_2^{-1} r_2}{1 - \nu_2^T R_2^{-1} \nu_2} \begin{pmatrix}
1_{1 \times 1} & -\nu_2^T R_2^{-1} \\
-\nu_2 R_2^{-1} \nu_2 & R_2^{-1} \nu_2 R_2^{-1} \nu_2^T R_2^{-1}
\end{pmatrix}
\]

\[
+ \pi_1, \rho \cdot \sqrt{1 - r_2^T R_2^{-1} r_2 \over 1 - \nu_2^T R_2^{-1} \nu_2} \begin{pmatrix}
0_{1 \times 1} & r_2^T R_2^{-1} \\
R_2^{-1} r_2 & -R_2^{-1} (\nu_2 r_2^T + r_2 \nu_2^T) R_2^{-1}
\end{pmatrix}
\]

\[
+ \begin{pmatrix}
0_{1 \times 1} & 0_{1 \times (p-2)} \\
0_{(p-2) \times 1} & R_2^{-1} r_2^T R_2^{-1}
\end{pmatrix}
\]

Therefore,

\[
\begin{pmatrix}
R_1^{-1} r_1^T R_1^{-1} - R_1^{-1} r_1 \\
-R_1^{-1} R_1^{-1} & 1_{1 \times 1}
\end{pmatrix}
\]

\[
= \pi_1, \rho \cdot \frac{\prod_{k=2}^{p-1} (1 - \pi_{k, \rho})}{\prod_{k=1}^{p-1} (1 - \pi_{1, \lambda})} \begin{pmatrix}
1_{1 \times 1} & -\nu_2^T R_2^{-1} \\
-\nu_2 R_2^{-1} \nu_2 & R_2^{-1} \nu_2 R_2^{-1} \nu_2^T R_2^{-1}
\end{pmatrix}
\]

\[
+ \pi_1, \rho \cdot \sqrt{\prod_{k=1}^{p-1} (1 - \pi_{k, \rho}) \over \prod_{k=1}^{p-1} (1 - \pi_{1, \lambda})} \begin{pmatrix}
0_{1 \times 1} & r_2^T R_2^{-1} \\
R_2^{-1} r_2 & -R_2^{-1} (\nu_2 r_2^T + r_2 \nu_2^T) R_2^{-1}
\end{pmatrix}
\]

\[
+ \begin{pmatrix}
0_{1 \times 1} & 0_{1 \times (p-2)} \\
0_{(p-2) \times 1} & R_2^{-1} r_2^T R_2^{-1}
\end{pmatrix}
\]

\[
+ \begin{pmatrix}
0_{1 \times 1} & 0_{1 \times (p-2)} \\
0_{(p-2) \times 1} & R_2^{-1} r_2^T R_2^{-1}
\end{pmatrix}
\]

\[
+ \begin{pmatrix}
0_{1 \times 1} & -r_2^T R_2^{-1} \\
R_2^{-1} r_2 & 1_{1 \times 1}
\end{pmatrix}
\]

110
since the partial autocorrelation matrix $\Pi = (\pi_{j,k})_{p \times p}$ has $k_0$ bands, $\pi_{j,j+k} = 0$ for $k > k_0$.

Hence,
\[
\begin{pmatrix}
R_1^{-1} r_1 r_1^T R_1^{-1} - R_1^{-1} r_1 \\
- r_1^T R_1^{-1} 1_{1 \times 1}
\end{pmatrix} = \begin{pmatrix}
0_{1 \times 1} & 0_{1 \times (p-2)} & 0_{1 \times 1} \\
0_{(p-2) \times 1} & R_2^{-1} r_2 r_2^T R_2^{-1} - R_2^{-1} r_2 \\
0_{1 \times 1} & - r_2^T R_2^{-1} 1_{1 \times 1}
\end{pmatrix}, \text{ (for } p - 1 > k_0).\]

Similarly, for $p - i > k_0$, we obtain
\[
\begin{pmatrix}
R_i^{-1} r_i r_i^T R_i^{-1} - R_i^{-1} r_i \\
- r_i^T R_i^{-1} 1_{1 \times 1}
\end{pmatrix} = \begin{pmatrix}
0_{1 \times 1} & 0_{1 \times (p-i-1)} & 0_{1 \times 1} \\
0_{(p-i-1) \times 1} & R_{i+1}^{-1} r_{i+1} r_{i+1}^T R_{i+1}^{-1} - R_{i+1}^{-1} r_{i+1} \\
0_{1 \times 1} & - r_{i+1}^T R_{i+1}^{-1} 1_{1 \times 1}
\end{pmatrix}.
\]

Therefore,
\[
\begin{pmatrix}
R_1^{-1} r_1 r_1^T R_1^{-1} - R_1^{-1} r_1 \\
- r_1^T R_1^{-1} 1_{1 \times 1}
\end{pmatrix} = \begin{pmatrix}
0_{(p-k_0-1) \times (p-k_0-1)} & 0_{(p-k_0-1) \times k_0} & 0_{1 \times 1} \\
0_{k_0 \times (p-k_0-1)} & R_{p-k_0}^{-1} r_{p-k_0} r_{p-k_0}^T R_{p-k_0}^{-1} - R_{p-k_0}^{-1} r_{p-k_0} \\
0_{1 \times (p-k_0-1)} & - r_{p-k_0}^T R_{p-k_0}^{-1} 1_{1 \times 1}
\end{pmatrix},
\]
where $R_{p-k_0} = R_{[p-k_0:p-1,p-k_0:p-1]}$ is a principle submatrix of $R$ with rows and columns from $(p-k_0)$ to $(p-1)$, and $r_{p-k_0} = R_{[p-k_0:p-1,1]}$.

Therefore, under the assumption of $k_0$ bands, we obtain
\[
R^{-1} = A_{2 \times 1}^{-1} = \prod_{j=1}^{p-1} \begin{pmatrix}
0_{(p-k_0-1) \times (p-k_0-1)} & 0_{1 \times k_0} & 0_{(p-k_0-1) \times 1} \\
0_{k_0 \times (p-k_0-1)} & R_{p-k_0}^{-1} r_{p-k_0} r_{p-k_0}^T R_{p-k_0}^{-1} - R_{p-k_0}^{-1} r_{p-k_0} \\
0_{1 \times (p-k_0-1)} & - r_{p-k_0}^T R_{p-k_0}^{-1} 1_{1 \times 1}
\end{pmatrix} + \begin{pmatrix}
R_1^{-1} & 0_{(p-1) \times 1} \\
0_{1 \times (p-1)} & 0_{1 \times 1}
\end{pmatrix}.
\]
Now, we re-write \( R \) similar to \( R \) to obtain

\[
R^{-1} = \prod_{j=p-k_0}^{p-1} (1 - \pi_{j,p}^2)^{-1} \left( \begin{array}{ccc} 0_{(p-k_0-0) \times (p-k_0-1)} & 0_{(p-k_0-1) \times k_0} & 0_{(p-k_0-1) \times 1} \\ 0_{k_0 \times (p-k_0-1)} & M_{jp} M_{jp}^T & -M_{jp} \\ 0_{1 \times (p-k_0-1)} & -M_{jp}^T & 1_{1 \times 1} \end{array} \right) 
+ \cdots 
+ \prod_{j=s-k_0}^{s-1} (1 - \pi_{j,s}^2)^{-1} \left( \begin{array}{cccc} 0_{(p-k_0-s-1) \times (p-k_0-s-1)} & 0_{(p-k_0-s-1) \times k_0} & 0_{(p-k_0-s-1) \times 1} & 0_{(p-k_0-s-1) \times s} \\ 0_{k_0 \times (p-k_0-s-1)} & M_{js} M_{js}^T & -M_{js} & 0_{k_0 \times s} \\ 0_{1 \times (p-k_0-s-1)} & -M_{js}^T & 1_{1 \times 1} & 0_{1 \times s} \\ 0_{s \times (p-k_0-s-1)} & 0_{s \times k_0} & 0_{s \times 1} & 0_{s \times s} \end{array} \right) 
+ \cdots 
+ \prod_{j=1}^{k_0} (1 - \pi_{j,k_0+1}^2)^{-1} \left( \begin{array}{cccc} 0_{1 \times 1} & 0_{1 \times k_0} & 0_{1 \times 1} & 0_{1 \times (p-k_0-2)} \\ 0_{k_0 \times 1} & M_{j(k_0+1)} M_{j(k_0+1)}^T & -M_{j(k_0+1)} & 0_{1 \times (p-k_0-2)} \\ 0_{1 \times 1} & -M_{j(k_0+1)}^T & 1_{1 \times 1} & 0_{1 \times (p-k_0-2)} \\ 0_{(p-k_0-2) \times 1} & 0_{(p-k_0-2) \times k_0} & 0_{(p-k_0-2) \times 1} & 0_{(p-k_0-2) \times (p-k_0-2)} \end{array} \right) 
+ \left( \begin{array}{cc} R^{-1}_{[1:k_0]} & 0_{k_0 \times (p-k_0)} \\ 0_{(p-k_0) \times k_0} & 0_{(p-k_0) \times (p-k_0)} \end{array} \right),
\]

where \( M_{jp} = R^{-1}_{[p-k_0:p-1,p-k_0:p-1]} R_{[p-k_0:p-1,p]} \), \( M_{js} = R^{-1}_{[s-k_0:s-1,s-k_0:s-1]} R_{[s-k_0:s-1,s-1]} \), and \( M_{j(k_0+1)} = R^{-1}_{[2:k_0+1,2:k_0+1]} R_{[2:k_0+1,k_0+2]} \). This shows that \( R^{-1} \) is a sum of \( p \times p \) matrices including only \((k_0 + 1) \times (k_0 + 1)\) non-zero principle sub-matrices. As a result, we
only need to invert \((k_0 - 1)\)-dimensional matrices to move from a \(k_0\) band \(\Pi\) to \(R^{-1}\). Furthermore, both \(R^{-1}\) and the precision matrix, \(\Sigma^{-1} = D^{-1}R^{-1}D^{-1}\) are \(k_0\)-band matrices.

**Proof of Result 5.**

Using Result 4, under the condition of a \(k_0\) band partial autocorrelation matrix,

\[
h(\hat{\Pi}_{k_0}, \hat{\sigma}_0) = \exp(-\frac{1}{2} \text{trace}(D^{-1}R^{-1}D^{-1}S))
\]

is only related to \((k_0 + 1) \times (k_0 + 1)\) principle sub-matrices of \(S\), i.e., it is only affected by sample partial autocorrelations with lag not greater than \(k_0\).

**Proof of Lemma 3.1.1**

Since \(Y_1, Y_2, \ldots, Y_n\) are i.i.d. multivariate normal random vectors \(N(0, DRD)\), the sample covariance \(S\) follows a Wishart distribution \(W_n(\Sigma)\) for \(n \geq p + 1\) with pdf,

\[
p(s) \propto |S|^\frac{n-p-1}{2} \exp\left(-\frac{1}{2} \text{trace}(D^{-1}R^{-1}D^{-1}S)\right).
\]

Let \(S = \{\hat{\sigma}_{j,l} : j, l = 1, \ldots, p\}\), \(A = (\hat{\sigma}_{1,2}, \hat{\sigma}_{2,3}, \hat{\sigma}_{1,3}, \ldots, \hat{\sigma}_{1,p}, \hat{\sigma}_{1,1}, \ldots, \hat{\sigma}_{p,p})\), \(B = (\hat{\rho}_{1,2}, \hat{\rho}_{2,3}, \hat{\rho}_{1,3}, \ldots, \hat{\rho}_{1,p}, \hat{\sigma}_{1,1}, \ldots, \hat{\sigma}_{p,p})\), and \(\hat{\Pi} = (\hat{\pi}_{1,2}, \hat{\pi}_{2,3}, \hat{\pi}_{1,3}, \ldots, \hat{\pi}_{1,p}, \hat{\sigma}_{1,1}, \ldots, \hat{\sigma}_{p,p})\).

The Jacobian from \(A\) to \(B\) is

\[
J = \begin{pmatrix} \hat{J}_{11} & 0 \\ 0 & I_{p \times p} \end{pmatrix}, \text{ where } \hat{J}_{11} = \text{diag}(vh(\hat{\sigma}_0 \otimes \hat{\sigma}_0)) \text{ and } \hat{\sigma}_0 = (\hat{\sigma}_{1,1}, \ldots, \hat{\sigma}_{p,p}).
\]

According to Joe (2006), the determinant of the Jacobian for \(J_{B \rightarrow \pi}\) is

\[
|J_{B \rightarrow \pi}| = \prod_{j=1}^{p-1}(1 - \hat{\pi}_{j,j+1}^2)^\frac{p-2}{2} \cdot \prod_{k=2}^{p-1} \prod_{j=1}^{p-k}(1 - \hat{\pi}_{j,j+k}^2)^\frac{p-1-k}{2}.
\]

Also, recall \(|\hat{R}| = \prod_{k=1}^{p-1} \prod_{j=1}^{p-k}(1 - \hat{\pi}_{j,j+k}^2)\) (Result 3).

Therefore,

\[
p(B) \propto |J||S|^\frac{n-p-1}{2} \exp\left(-\frac{1}{2} \text{trace}(D^{-1}R^{-1}D^{-1}S)\right) \text{ and } \]

\[
p(\pi, \sigma_0) \propto |J_{11}| |\hat{J}_{B \rightarrow \pi}| |\hat{D} \hat{R} \hat{D}|^\frac{n-p-1}{2} \exp\left(-\frac{1}{2} \text{trace}(D^{-1}R^{-1}D^{-1}S)\right).
\]
We can simplify this as follows,

\[
p(\pi, \sigma_0) \propto |\hat{J}_{11}| \hat{D}^{n-p-1} \left[ \prod_{j=1}^{p-1} (1 - \pi_{j,j+1})^{\frac{n-2}{2}} \cdot \prod_{k=2}^{p-k} (1 - \hat{\pi}_{j,j+k})^{\frac{n-k}{2}} \right].
\]

\[
\prod_{k=1}^{p-k} \left[ \prod_{j=1}^{p-2} (1 - \pi_{j,j+1}) \right]^{\frac{n-p-1}{2}} h(\hat{\pi}, \hat{\sigma})
\]

\[
= (1 - \hat{\pi}_{1,1, p})^{n-3} \prod_{k=2}^{p-k} (1 - \hat{\pi}_{j,j+k})^{\frac{n-k-2}{2}} \cdot \prod_{j=1}^{p-1} \prod_{k=2}^{p-k} (1 - \hat{\pi}_{j,j+k})^{\frac{n-k}{2}} |\hat{J}_{11}| \hat{D}^{n-p-1} h(\hat{\pi}, \hat{\sigma})
\]

\[
= (1 - \hat{\pi}_{1,1, p})^{n-3} \prod_{k=2}^{p-k} (1 - \hat{\pi}_{j,j+k})^{\frac{n-k-2}{2}} \cdot h^*(\hat{\pi}, \hat{\sigma}),
\]

where \( h^*(\hat{\pi}, \hat{\sigma}) = \prod_{j=1}^{p-1} (1 - \pi_{j,j+1})^{\frac{n-2}{2}} \cdot \prod_{k=2}^{p-k} (1 - \hat{\pi}_{j,j+k})^{\frac{n-k-2}{2}} |\hat{J}_{11}| \hat{D}^{n-p-1} h(\hat{\pi}, \hat{\sigma}) \).

Hence, all sample partial autocorrelations with lags not less than \( k_0 \) are independent with marginal distributions given by

\[
f(\hat{\pi}_{j,j+k}) \propto \begin{cases} 
(1 - \hat{\pi}_{j,j+k})^{\frac{n-k-2}{4}} & \text{for } k \in \{k_0 + 1, \ldots, p - 2\} \\
(1 - \hat{\pi}_{1,1, p})^{\frac{n-p-1}{2}} & \text{for } k_0 < k = p - 1
\end{cases}
\]
Proof:

Let $A$ be a $n \times n$ symmetric matrix with elements, $\{a_{ij}\}$. Define

$$vh(A) = \begin{pmatrix} a_{21} & a_{31} & \cdots & a_{n1} & a_{32} & \cdots & a_{n2} & \cdots & nn-1 \end{pmatrix}^T,$$

(C–1)

to be a vector with all the elements in the lower triangular part of a square matrix (without the main diagonal). Let $Y_1, Y_2, \ldots, Y_n$ be iid $m$-dimensional multivariate normal random vectors with mean zero and covariance matrix $\Sigma_m = D_mD_m$, where $D_m$ is diagonal matrix with standard deviation on its main diagonal and $R_m$ is the correlation matrix corresponding to $\Pi_m$. For simplicity, we drop subscripts $m$ of $\rho_m$ and $\pi_m$ in the following proof without lack of clarity.

To derive the asymptotic covariance matrix of the lag $j$ estimated partial autocorrelations, $\hat{\pi}_j = (\hat{\pi}_{1j+1}, \hat{\pi}_{2j+2}, \ldots, \hat{\pi}_{m-jm})$, the MLE of $\pi_j = (\pi_{1j+1}, \pi_{2j+2}, \ldots, \pi_{m-jm})$, we introduce some additional notation which relates to the transformation from

$$(vh(\Sigma)^T, \sigma_0^T)^T \rightarrow (vh(R)^T, \sigma_0^T)^T \rightarrow (vh(\pi)^T, \sigma_0^T)^T,$$

where $\sigma_0 = (\sigma_{11}, \ldots, \sigma_{mm})$.

Let

$$J_{11}^\rho = \text{diag}(vh(\sigma_0 \otimes \sigma_0^{-\frac{1}{2}}))$$

and
Also, let

\[
\Omega(\sigma) = \begin{pmatrix}
\Omega_{11}^{*} & \Omega_{12}^{*} \\
\Omega_{21}^{*} & \Omega_{22}^{*}
\end{pmatrix}
\]

be the partitioned covariance matrix of \( \sigma^* = (\nu h(\hat{\Sigma})^T, \hat{\sigma}_0) \)

According to a Theorem in Magnus and Neudecker (1988) and results discussed in Wang and Daniels (2012), the asymptotic covariance matrix of \( \hat{\rho} = \nu h(\hat{R}) \) is

\[
\Omega(\rho_0) = J^p_{11} \Omega_{11}^{*} J^p_{11} + J^p_{12} \Omega_{22}^{*} J^p_{12} + (J^p_{11} \Omega_{12}^{*} J^p_{12} + J^p_{12} \Omega_{21}^{*} J^p_{11}) = (\omega_{ij,ij}),
\]

where \( \omega_{ij,ij} = \omega_{ij,ij}^{(1)} + \omega_{ij,ij}^{(2)} + \omega_{ij,ij}^{(3)} \) which is the covariance between \( \rho_{ij} \) and \( \rho_{ij,ij} \),

\[
J^p_{11} \Omega_{11}^{*} J^p_{11} = (\omega_{ij,ij}^{(1)}),
\]

\[
J^p_{12} \Omega_{22}^{*} J^p_{12} = (\omega_{ij,ij}^{(2)}),
\]

and

\[
(J^p_{11} \Omega_{12}^{*} J^p_{12} + J^p_{12} \Omega_{21}^{*} J^p_{11}) = (\omega_{ij,ij}^{(3)}).
\]

Then,
Let $\lambda_0$ be the constant stated in Proposition 4.1. Then, we have

\[
\begin{align*}
\omega^{(1)}_{j,j_1} &= \rho_{i_1} \rho_{j_1} + \rho_{j_1} \rho_{i_1}, \\
\omega^{(2)}_{j,j_1} &= -\left[\rho_{i_1} \rho_{i_1} + \rho_{j_1} \rho_{j_1} + \rho_{j_1} \rho_{i_1} + \rho_{i_1} \rho_{j_1} + \rho_{j_1} \rho_{j_1}\right], \\
\omega^{(3)}_{j,j_1} &= \rho_{j_1} \rho_{i_1} \left(\rho_{i_1}^2 + \rho_{i_1}^2 + \rho_{j_1}^2 + \rho_{j_1}^2\right).
\end{align*}
\]

Let $\lambda_0$ be the constant stated in Proposition 4.1. Then, we have

\[
\begin{align*}
\omega^{(1)}_{j+s,j+t+1,j_1+s,j_1+t+1} &= \rho_{j+s,j+t+1} + \rho_{j+s,j+t+1} + \rho_{j+s,j+t+1} + \rho_{j+s,j+t+1} + \rho_{j+s,j+t+1} \\
|\omega^{(1)}_{j+s,j+t+1,j_1+s,j_1+t+1}| &\leq o(\lambda_0^{j-j}) \cdot \rho_{j+s,j+t+1} + o(\lambda_0^{j-j}) \cdot \rho_{j+s,j+t+1} \\
|\omega^{(2)}_{j+s,j+t+1,j_1+s,j_1+t+1}| &\leq o(\lambda_0^{j-j}) \cdot o(\lambda_0^{j-j}) + o(\lambda_0^{j-j}) \cdot o(\lambda_0^{j-j}) \\
|\omega^{(3)}_{j+s,j+t+1,j_1+s,j_1+t+1}| &\leq o(\lambda_0^{j-j}) \cdot o(\lambda_0^{j-j}) + o(\lambda_0^{j-j}) \cdot o(\lambda_0^{j-j})
\end{align*}
\]
Therefore,

\[ \omega(j+s,j+t+1; j_1+s,j_1+t+1) \leq o(\lambda_0^{2|j_1-j|}). \]

Now, let's focus on covariance of \( \hat{\pi}'s \) within lag \( k \), i.e., covariance of between \( \hat{\pi}_{jj+k} \) and \( \hat{\pi}_{j_1j_1+k} \).

\[
\text{cov}(\hat{\pi}_{jj+k}, \hat{\pi}_{j_1j_1+k}) = \frac{\partial \pi_{jj+k}}{\partial \rho} \frac{\partial \pi_{j_1j_1+k}}{\partial \rho} \leq (\lambda_0^{2|j_1-j|}).
\]

\[ (C-4) \]

**C.2 Estimate Asymptotic Optimal Bandwidth**

To estimate \( \theta_{r,s} = \int \pi^*(r)(x) \pi^*(s)(x) v(x) \, dx \), we fit \( \pi^*(x) \) with local regression using a polynomial \( q > p \), and let

\[
\hat{\pi}^*(j)(x, h, q) = e_j^T (X_q^T W_q X_q,v)^{-1} X_q^T W_q \hat{\pi},
\]

where \( g_{r,s} \) is a bandwidth for estimating \( \theta_{r,s} \).

Here and in what follows, we denote \( e_j \) is a column vector with 1 as its jth entry and all other entries zero.
and then estimate $\hat{\theta}_{s,t}$ as

$$
\hat{\theta}_{r,s} = \frac{1}{m - k} \sum_{i=1}^{m-k} \hat{\pi}^*(r) \hat{\pi}^*(s),
$$

(C–5)

Now, let $K_{r,q}(u) = r![[M_{r,q}(u)]/|N_q|]K(u)$, where $N_q$ is a $(q+1) \times (q+1)$ matrix having $(i,j)$ entry equal to $\int u^i \pi^j K(u) du$ and $M_{r,q}(u)$ is the same as $N_q$, except that the $(r+1)$th row is replaced by $(1, u, \ldots, u^q)$. The kernel $K_q$ is defined to be $K_{0,q}$. Finally, let

$$(L_1 * L_2)(x) = \int L_1(u) L_2(x - u) du$$

denote the convolution of two real valued functions $L_1$ and $L_2$. Then, the minimizer of the mean square error of $\hat{\theta}_{2,2}$, called $g_{2,2}$ is

$$
g_{2,2} \approx C_1(K) \left[ \frac{\sigma^2_{\pi,k}(b-a)}{\theta_{24}(m-k)} \right]^{\frac{1}{2}},
$$

(C–6)

where

$$
C_1(K) = \begin{cases} 
\left[ \frac{12R(K_{2,3})}{\mu_4(K_{2,3})} \right]^{\frac{1}{2}} & \theta_{24} < 0 \\
\left[ \frac{30R(K_{2,3})}{\mu_4(K_{2,3})} \right]^{\frac{1}{2}} & \theta_{24} > 0
\end{cases}
$$

Now we need to estimate $\sigma^2_{\pi,k}$ and $\theta_{24}$ in Eq. C–6. We estimate $\sigma^2_{\pi,j}$ via a blocked quartic polynomial fit as follows,

let $N$ be the number of subsamples, and let $\mathcal{X}_j = \{X_{(j-1)t+1}, \ldots, X_{jt}\}$, $t = 1, 2, \ldots, \left\lfloor \frac{m-k}{N} \right\rfloor$, denote the $j$th subsample of ordered $X_i$'s. Let $\hat{\pi}^*_{j,Q}(x)$ be the least squares quartic fit obtained from data containing in $\mathcal{X}_j$. For $\max(r,s) \leq 4$, the "blocked quartic estimator" for $\theta_{r,s}$ is

$$
\hat{\theta}^Q_{r,s}(N) = \frac{1}{m - k} \sum_{i=1}^{m-k} \sum_{j=1}^{N} (\hat{\pi}^*_{j,Q}(r)(X_i) \hat{\pi}^*_{j,Q}(s)(X_i)) I\{X_i \in \mathcal{X}_j\}.
$$

and

$$
\hat{\sigma}^2_{Q}(N) = \frac{1}{(m - k) - 5N} \sum_{i=1}^{m-k} \sum_{j=1}^{N} (\hat{\sigma}^*_{j,Q}(X_i))^2 I\{X_i \in \mathcal{X}_j\}
$$

The optimal $N \in \{1, 2, \ldots, N^*\}$ is chosen by minimizing Mallows’ $C_p(N)$,

$$
C_p(N) = \frac{RSS(N)}{RSS(N_{\max})/((m - k) - 5N_{\max})} - ((m - k) - 10N),
$$

119
where $RSS(N)$ is residual sum of squares for a blocked quartic fit with $N$ blocks, $N \in \{1, \ldots, N_{\text{max}}\}$, and $RSS(N_{\text{max}})$ is residual sum of squares for a blocked quartic fit with $N_{\text{max}}$,

$$N_{\text{max}} = \max\{\min\left(\frac{m-k}{20}\right), N^*, 1\}, \quad (C-7)$$

where $N^*$, is some positive integer. Ruppert et al. (1995) suggests $N^* = 5$.

After estimating $g_{2,2}$, we use corresponding estimator of $\pi^*$ to calculate the optimal bandwidth, $\lambda_\sigma$, for $\sigma^2$ in Eq. 4–8

$$\lambda_\sigma = C_2(K) \left[ \frac{\sigma^4}{\hat{\theta}^2_{q+1,q+1} (m-k)^2} \right]^{1/4q+5},$$

where $C_2(K) = \left[ \frac{((q+1)t^{4R(K_q+K_q-2K_q)})}{2(q+1)\mu_{q+1}(K_q)^4} \right]^{1/4q+5}$ with $p=2$. Then, we estimate $\sigma^2_{\pi,k}$ as

$$\hat{\sigma}^2_{\pi,k}(\lambda) = \nu^{-1} \sum_{i=1}^{m-k} \{ \pi_i - \hat{\pi}^*(x_i; \lambda) \}^2,$$

where $\nu = (m-k) - 2 \sum_i \omega_{ii} + \sum_{i,j} \omega_{ij}^2$ and $\omega_{ij} = e_1^T(X^{T}_{q,x} W_{x_i} X_{q,x_i})^{-1} X_{q,x_i} W_x e_j$.

Then we use truncated data within $100\alpha\%$ is used to adjust boundary bias and the estimator of $\theta_{r,s}$ is replaced by

$$\hat{\theta}_{r,s}^\alpha(g) = \frac{1}{m-k} \sum_{i=1}^{m-k} \hat{\pi}^{*(r)}(X_i) \hat{\pi}^{*(s)}(X_i) I[\alpha < X_i < (1-\alpha)].$$

As Ruppert et al. (1995) suggest, we propose following steps to estimate optimal asymptotic bandwidth $h$,

1. Find $\theta^2_{2,2}(N)$ and $\sigma^2_{2}(N)$ according to a block quartic fit with $\hat{N}$ chosen by minimizing Mallows’ $C_p$ and $N_{\text{max}}$ given by $(C-7)$ with $N^* = 5$.

2. Estimate $\theta_{2,2}$ as $\hat{\theta}_{2,2}^{0.05}(\hat{g}_{2,2})$, where

$$\hat{g}_{2,2} = C_1(K) \left[ \frac{\hat{\sigma}^2_{\hat{N}}(\hat{N})}{|\hat{\theta}^2_{2,4}(\hat{N})|(m-k)} \right]^{1/7}. $$
and estimate $\sigma_{\pi,k}^2$ as $\hat{\sigma}_{1}^2(\lambda_\sigma)$, where

$$\hat{\lambda}_\sigma = C_2(K) \left[ \frac{\hat{\sigma}_Q^2(\hat{N})}{(\hat{\theta}_{2,2}^{0.05}(\hat{g}_{2,2}))^2(m-k)^2} \right]^{1/7}.$$

we use the differences $\hat{\pi}^{(i)} - \hat{\pi}_*^{(i)}(\lambda_\sigma)$ as estimated residuals (and calculate the covariance function estimator based on these residuals, Herrmann et al. (1992)) then estimate the long run covariance $S_{\pi,k}$ by

$$\hat{S}_{\pi,k}(\lambda_\sigma) = \sum_{j=1}^{m-k-1} \left[ \frac{1}{m-j-k+1} \sum_{i=1}^{m-j-k} (\hat{\pi}^{(i)} - \hat{\pi}_*^{(i)}(\hat{\lambda}_\sigma))(\hat{\pi}^{(i+j)} - \hat{\pi}_*^{(i+j)}(\hat{\lambda}_\sigma)) \right].$$

3. Estimate the bandwidth $h$ with Epanechnikov kernel, $K(x) = \frac{3}{4}(1-x^2)I(|x| \leq 1)$ as

$$\hat{h}_{m,k}^{\text{rob}} = \left\{ \frac{15(\hat{\sigma}_1^2(\lambda_\sigma) + 2\hat{S}_{\pi,k}(\lambda_\sigma))}{\hat{\theta}_{2,2}^{0.05}(\hat{g}_{2,2})} \right\}^{\frac{1}{5}} \left( \frac{1}{m-k} \right)^{\frac{1}{5}}. \quad (C-8)$$
REFERENCES


Daniels, M. J. (2006), “Bayesian Modeling of Several Covariance Matrices and Some Results on Popriety of the Posterior for Linear Regression with Correlated and/or Heterogeneous Errors,” Journal of Multivariate Analysis, 97, 1185–1207.


122


BIOGRAPHICAL SKETCH

Yanpin Wang was born and raised in the Sichuan, China. She graduated from Sichuan Normal University with Bachelor degree in mathematics. Then she taught mathematics in middle school and high school in China. In 2001, she moved to USA with her Husband Jiangtao Luo and her son Binjie. Yanpin was offered graduate assistantship by the Department of Mathematics at the University of Florida in 2005. In 2007, she earned her Master Degree in mathematical science and was admitted to the Department of Statistics. After receiving Master degree in statistics, she transferred to the Biostatistics Department in the College of Public Health and Health Professions at the University of Florida as her wish. She graduates in August 2012 with her Doctorate degree in biostatistics. During her time at the University of Florida, She has been an instructor in the Department of Mathematics and a research assistant in the College of Public Health and Health Professions and the Department of Statistics. Yanpin has won several travel awards for conferences and ENAR Distinguish Student Paper Award.