Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

March 11, 2018

Abstract

With high dimensional longitudinal and functional data becoming much more common, there is a strong need for methods of estimating large covariance matrices. Estimation is made difficult by the instability of sample covariance matrices in high dimensions and a positive-definite constraint we desire to impose on estimates. A Cholesky decomposition of the covariance matrix allows for parameter estimation via unconstrained optimization as well as a statistically meaningful interpretation of the parameter estimates. Regularization improves stability of covariance estimates in high dimensions, as well as in the case where functional data are sparse and individual curves are sampled at different and possibly unequally spaced time points. By viewing the entries of the covariance matrix as the evaluation of a continuous bivariate function at the pairs of observed time points, we treat covariance estimation as bivariate smoothing.

Within regularization framework, we propose novel covariance penalties which are designed to yield natural null models presented in the literature for stationarity or short-term dependence. These penalties are expressed in terms of variation in continuous time lag and its orthogonal complement. We present numerical results and data analysis to illustrate the utility of the proposed method.

keywords: non-parametric, covariance, longitudinal data, functional data, splines, reproducing kernel Hilbert space

1 Introduction

An estimate of the covariance matrix or its inverse is required for nearly all statistical procedures in classical multivariate data analysis, time series analysis, spatial statistics and, more recently, the growing field of statistical learning. Covariance estimates play a critical role in the

*The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201
†The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201

performance of techniques for clustering and classification such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), factor analysis, and principal components analysis (PCA), analysis of conditional independence through graphical models, classical multivariate regression, prediction, and Kriging. Covariance estimation with high dimensional data has has recently gained growing interest; it is generally recognized that there are two primary hurdles responsible for the difficulty in covariance estimation: the instability of sample covariance matrices in high dimensions and a positive-definite constraint we wish estimates to obey.

Prevalent technological advances in industry and many areas of science make high dimensional longitudinal and functional data a common occurrence, arising in numerous areas including medicine, public health, biology, and environmental science with specific applications including fMRI, spectroscopic imaging, gene microarrays among many others, presenting a need for effective covariance estimation in the challenging situation where parameter dimensionality p is possibly much larger than the number of observations, n.

We consider two types of potentially high dimensional data: the first is the case of functional data or times series data, where each observation corresponds to a curve sampled densely at a fine grid of time points; in this case, it is typical that the number of time points is larger than the number of observations. The second is the case of sparse longitudinal data where measurement times may be almost unique yet sparsely distributed within the observed time range for each individual in the study. In this case, the nature of the high dimensionality may not be a consequence of having more measurements per subject than the number of subjects themselves, but rather because when pooled across subjects, the total number of unique observed time points is greater than the number of individuals. Several approaches have been taken in effort to overcome the issue of high dimensionality in covariance estimation. Regularization improves stability of covariance estimates in high dimensions, particularly in the case where the parameter dimensionality p is much larger than the number of observations n. Regularization of the covariance matrix and its Cholesky decomposition has been explored extensively through various approaches including banding, tapering, kernel smoothing, penalized likelihood, and penalized regression; see citetpourahmadi2011covariance for a comprehensive overview.

To overcome the hurdle of enforcing covariance estimates to be positive definite, several have considered modeling various matrix decompositions including variance-correlation decomposition, spectral decomposition, and Cholesky decomposition. The Cholesky decomposition has received particular attention, as it which allows for a statistically meaningful interpretation as well as an unconstrained parameterization of elements of the covariance matrix. This parameterization allows for estimation to be accomplished as simply as in least squares regression. If we assume that the data follow an autoregressive process with (possibly) heteroskedastic errors, then the two matrices comprising the Cholesky decomposition, the Cholesky factor (which diagonalizes the covariance matrix) and diagonal matrix itself, hold the autoregressive coefficients and the error variances, respectively.

In longitudinal studies, the measurement schedule could consist of targeted time points or could consist of completely arbitrary (random) time points. If either the measurement schedule

has targeted time points which are not necessarily equally spaced or if there is missing data, then we have what is considered incomplete and unbalanced data. If the measurement schedule has arbitrary or almost unique time points for every individual so that at a given time point there could be very few or even only a single measurement, we must consider how to handle what we consider as sparse longitudinal data. We view the response as a stochastic process with corresponding continuous covariance function and the generalized autoregressive parameters as the evaluation of a continuous bivariate function at the pairs of observed time points rather than specifying a finite set of observations to be multivariate normal and estimating the covariance matrix. This is advantageous because it is unlikely that we are only interested in the covariance between pairs of observed design points, so it is reasonable to approach covariance estimation in a way that allows us to obtain an estimate of the covariance between two measurements at any pair of time points within the time interval of interest.

Through the Cholesky decomposition, we formulate covariance estimation as a penalized regression problem and propose novel covariance penalties designed to yield natural null models presented in the literature. By transforming the axes of the design points, we express these penalties in terms of two directions: the lag component and the additive component and characterize the solution coefficient function in terms of a functional ANOVA decomposition. Some have sidestepped the issue of high dimensionality by prescribing simple parametric models for the elements of the Cholesky decomposition. **?**, Pourahmadi [1999], and Pourahmadi and Daniels [2002] have elicited stationary parametric models for the generalized autoregressive coefficients, letting the GARPs depend only on the distance between two time points. To induce the structural simplicity of such stationary models with the flexibility of a nonparametric approach, we penalize all functional components but that corresponding to the lag component so that the set of null models is comprised of stationary models. Huang et al. [2007] follow the hueristic argument presented in Pourahmadi [1999] that the generalized autoregressive parameters are monotone decreasing in as lag increases and set off-diagonal elements of either the covariance matrix or the Cholesky factor corresponding to large lags to zero. Rather than shrinking element of the Cholesky factor to zero after particular value of l, we choose to enforce structure of the Cholesky factor such that the null models coincide with parsimonious models commonly used in time series analysis and with simple parametric models proposed in the nonparametric covariance estimation literature.

The remainder of the chapter serves as a brief survey of developments in covariance estimation. We will highlight a number of approaches to parsimonious covariance modeling, but our attention will be delegated to recent progress in parsimonious covariance models for longitudinal data. The review will conclude with the presentation of matrix factorizations for reparameterizing elements of the covariance matrix, translating covariance estimation into a generalized linear modeling problem.

2 Covariance estimation: a review

Estimation of the covariance matrix is fundamental to the analysis of multivariate data, and the most commonly used estimator is the sample covariance matrix, S. While it is both positive-definite and an unbiased estimator of Σ, it is unstable large dimension M. Approaches rooted in decision theory yield stable estimators which are scalar multiples of the sample covariance matrix; these estimators distort the eigenstructure of Σ unless the sample size is greater than the dimension, $N >> M$ (Dempster [1972].) There is a vast body of work which addresses the efficient estimation of the covariance matrix of a normal distribution by correcting the eigenstructure distortion or reducing the number of parameters to be estimated. See Stein [1975], Lin [1985], Yang and Berger [1994], Daniels and Kass [1999], Champion [2003]

The sample covariance matrix S, which is used in virtually all multivariate techniques, is both unbiased and positive-definite. The flexible estimator is also computationally convenient, however it is neither parsimonious nor, in high dimensions, a stable estimator. Given a sample of size N $Y_1, \ldots, Y_N$, from an M-dimensional Normal distribution with mean μ and covariance matrix Σ, the sample covariance matrix

$$S = (N-1)^{-1} \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right) \left(Y_i - \bar{Y}\right)' \tag{1}$$

is a straightforward estimator of the $\frac{M(M+1)}{2}$ parameters of the unstructured covariance matrix Σ. The number of parameters of $\Sigma = (\sigma_{ij})$ grows quadratically in the dimension M, and the parameters must satisfy the positive-definiteness constraint

$$v'\Sigma v = \sum_{i,j=1}^{M} v_i v_j \sigma_{ij} \geq 0 \tag{2}$$

for all $v \in \Re^M$. The challenge presented by these hurdles have motivated a growing body of research in statistics and its areas of application aimed at effectively estimating covariance matrices.

Our review of work in this area focuses on developments made from two connected perspectives: regularization or sparsity in covariance matrices for high-dimensional data, and generalized linear models (GLM) or parsimony and use of covariates in low dimensions. A recurring technique in both perspectives is the reduction of covariance estimation to estimating a single of sequence of regression. The generalized linear model (GLM) framework McCullagh and Nelder [1989] merges numerous seemingly disconnected approaches to model the mean of a distribution, and can accommodate many types of including normal, probit, logistic and Poisson regressions, survival data, and log-linear models for contingency tables. The key to the power of the GLM paradigm is the use of a link function to induce unconstrained reparameterization for the mean of a distribution, and hence the ability to reduce the dimension of the parameter space via modeling the covariate effect additively by increasing the number of parameters gradually one at a time corresponding to

inclusion of each covariate. The extension of the GLM has lead to large class of models including nonparametric and generalized additive models, Bayesian GLM, and generalized linear mixed models. See Hastie and Tibshirani [1990], Dey et al. [2000], McCulloch and Neuhaus [2001]. An analogous framework for modeling covariance matrices facilitates further developments in covariance estimation from the Bayesian, nonparametric and other paradigms.

2.1 Structured parametric covariances

2.2 Structured parametric covariances

In the applied statistics literature, particularly for repeated measure data, it is quite common to pick a stationary covariance matrix for the covariance structure. Typical choices are simple models which depend on a small number of parameters such as compound symmetry and autoregressive models of order k, where k is small.We will review a selection of modeled frequently encountered in the applied statistics literature in sections to follow. This approach is attractive because it is computationally inexpensive, and software packages implementing fitting procedures for a growing number of simple models are readily accessible. The compound symmetric model was at one time a very popular choice for parametric covariance structure, specifying

$$\sigma_{ij} = \begin{cases} \rho, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \tag{3}$$

where σ_{ij} denotes the (i, j) element of Σ. With only two parameters to be estimated, this model is highly parsimonious, but has received less attention with the development of models that allow for heterogeneous variances and non-constant correlation.

The first order autoregressive model for response variable y_t associated with measurement time t specifies

$$y_t = \begin{cases} \mu_t + \epsilon_t, & t = 1, \\ \\ \mu_t + \rho\left(y_{t-1} - \mu_{t-1}\right) + \epsilon_t, & t > 1, \end{cases} \tag{4}$$

where $|\rho| < 1$, and the innovations $\{\epsilon_t\}$ are independently distributed according to $N\left(0, \sigma_t^2\right)$ with $\sigma_1^2 = \sigma^2/\left(1 - \rho^2\right)$, and $\sigma_t^2 = \sigma^2$ for $t = 2, \ldots, M$. The corresponding dependence components of the covariance structure are monotonically decreasing in $l = |i - j|$; specifically,

$$\sigma_{ij} = \begin{cases} \rho^{|i-j|}, & i \neq j, \\ \\ \sigma^2, & i = j, \end{cases} \tag{5}$$

The AR(1) model generalizes to any arbitrary order p by simply adding additional predecessors to

the covariates in the linear model for y_t:

$$y_t = \begin{cases} \mu_t + \epsilon_t, & t = 1, \\ \mu_t + \sum_{j=1}^{p^*} \phi_j \left(y_{t-j} - \mu_{t-j} \right) + \epsilon_t, & t > 1, \end{cases}$$

where $p^* = \min(p, t-1)$, and the $\{\epsilon_t\}$ are independent mean zero Normal random variables. The variance of $\{\epsilon_t\}$ is constant for $t > p$, and for $t \leq p$, the variance is specified so as to ensure that the variance is constant across all responses y_t and the covariance between y_i and y_j depends only on $|i-j|$.

The response specification for q^{th} order moving average model is given by

$$y_t = \sum_{j=0}^{q} \theta_j \epsilon_{t-j}, \tag{6}$$

where the $\{\epsilon_t\}$ are independently and identically distributed mean zero Normal random variables with variance σ^2. This model corresponds to covariance structures with elements given by

$$\sigma_{ij} = \begin{cases} \left(\theta_{i-j} + \theta_1 \theta_{i-j+1} + \cdots + \theta_{q-i+j} \theta_q \right) / \left(1 + \sum_{j=1}^{q} \theta_j^2 \right), & |i-j| \leq q, \\ 0, & |i-j| > q, \\ \sigma^2 \sum_{j=0}^{q} \theta_j^2, & i = j, \end{cases}$$

Thus, variances are constant and correlations between y_t and y_{t-l} vanish beyond a finite, constant lag l. Here $\rho_1, \ldots, \rho_q$are arbitrary parameters subject only to positive definiteness constraints. This model generalizes to a q^{th}-order Toeplitz model, which specifies

$$\sigma_{ij} = \begin{cases} \rho_{i-j} & |i-j| \leq q, \\ 0 & |i-j| > q, \\ \sigma^2 & i = j, \end{cases} \tag{7}$$

or covariance matrix of the form

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \ldots & m_{p-1} \\ m_1 & m_0 & m_1 & \ldots & m_{p-2} \\ m_2 & m_1 & m_0 & \ldots & m_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p-1} & m_{p-2} & m_{p-3} & \ldots & m_0 \end{bmatrix}, \tag{8}$$

where $m_j = 0$ for all $j > q$.

In turn, one can further generalize to a q^{th}-order banded model by specifying that the covariances on off-diagonals of the correlation matrix beyond the q^{th} off-diagonal are zero, and otherwise not imposing any structural restrictions on the remaining elements of the covariance matrix beyond those required for positive definiteness. The tradeoff of the additional flexibility of the general banded model over the MA and Toeplitz models is that the number of parameters in a general q-banded covariance structure is $O(n)$ rather than $O(1)$.

The aforementioned models are stationary, specifying constant variance and with equal same-lag correlations among responses when the data are observed on a regular grid. Heterogeneous extensions of these models specify the same form of the correlation but allow time-dependent response variances. Completely general time dependence (subject to positive definiteness constraints) requires the covariance structure to be characterized by $O(n)$ parameters, while specifying linear or quadratic dependence on time leads to more parsimonious heterogeneous models.

An ARIMA(p, d, q) model generalizes a stationary autoregressive moving average (ARMA) model by postulating that not the observations themselves, but rather the d^{th}-order differences among consecutive measurements follow a stationary ARMA(p, q) model. A special case is the ARIMA$(0, 1, 0)$ model - the random walk:

$$y_t = \mu_t + \sum_{j=1}^{t} \epsilon_j, \quad t = 1, \ldots, M, \tag{9}$$

where the ϵ_t are independent mean zero Normal random variables with variance σ_ϵ^2. The variance of the process increases linearly in time, and the correlation between y_t and y_{t-l} also increases, but nonlinearly, in time:

$$\sigma_{ij} = \begin{cases} \sqrt{i/j} & i \neq j \\ j\sigma_\epsilon^2 & i = j, \end{cases} \tag{10}$$

This model is applicable to longitudinal data only when data are observed on a regular grid, however, its continuous time analogue permits this restriction to be relaxed. An important special case is the continuous time analogue to the random walk, the Weiner process, which has covariance function $Cov\left(y\left(t_i\right), y\left(t_j\right)\right) = \sigma^2 \min\left(t_i, t_j\right)$.

Random coefficient models are a broad class of models often used for clustered or longitudinal data. They offer reasonable flexibility for characterizing dependency structure but remain parsimonious because the number of model parameters is unrelated to the number of repeated measurements and can be applied to non-rectangular data. The formulation of the covariance structure for these models is most usually a consideration of regressions that vary across subjects rather than a consideration of within-subject similarity, which is why they are most often considered distinct

from parametric covariance models. Still, they yield parametric covariance structures that generally have non-constant variances and non-stationary correlations. A general form of the random coefficient model is given by

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \quad i = 1, \ldots, M, \tag{11}$$

where the Z_i are specified matrices, the γ_i are vectors of random coefficients distributed independently as $N(0, G_i)$, the G_i are positive definite but otherwise unstructured matrices, and the ϵ_i are distributed independently (of the γ_i and of each other) as $N\left(0, \sigma^2 \mathrm{I}_{n_i}\right)$. The G_i are usually assumed to be equal, so the covariance matrix of y_i is taken to be $\Sigma_i = Z_i G Z_i' + \sigma^2 \mathrm{I}_{n_i}$. Special cases include the linear random coefficients (RCL) and quadratic random coefficients (RCQ) models. In the linear case, $Z_i = \left[1_{m_i}, \left(t_{i1}, \ldots, t_{i,m_i}\right)'\right]$ and

$$G = \begin{bmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{10} & \sigma_{11} \end{bmatrix}$$

In the quadratic case, $Z_i = \left[1_{m_i}, \left(t_{i1}, \ldots, t_{i,m_i}\right)', \left(t_{i1}^2, \ldots, t_{i,m_i}^2\right)'\right]$. It is worth noting that when $Z_i = 1_{m_i}$, the random coefficient model corresponds to the compound symmetric model 5. The covariance structure for a subject having measurements $y_1, \ldots, y_{m_i}$ taken at equally spaced measurement times $t_1 = 1, \ldots, t_{m_i} = m_i$ is given by

$$\sigma_{ij} = \begin{cases} \frac{\sigma_{00}+\sigma_{01}(i+j)+\sigma_{11}ij}{\sqrt{\sigma^2+\sigma_{00}+2i\sigma_{01}+\sigma_{11}i^2}\sqrt{\sigma^2+\sigma_{00}+2j\sigma_{01}+j^2\sigma_{11}}} & i \neq j \\ \sigma^2 + \sigma_{00} + 2\sigma_{01}j + \sigma_{11}j^2 & i = j, \end{cases} \tag{12}$$

These models are can permit variance and covariances which exhibit several kinds time dependency, including increasing or decreasing variances and correlations of which some are negative while others are positive. However, this model does not permit variances which are concave-down in time, and it precludes the variances from being constant if the same-lag correlations are different.

The previous list is far from an exhaustive list of parametric covariance structures - we will later reference structures which we have not discussed here, such as antedependence models. For example, see Jennrich and Schluchter [1986] for additional models for repeated measures data. While these models are computationally attractive and the choices for parametric model structure are seemingly unlimited, specifying the appropriate parametric covariance structure is a challenge even for the experts, and model misspecification can lead to considerably biased estimates. To strike a balance between the variability of the sample covariance matrix and the bias of the estimated structured covariance matrix, it is prudent to rely on the data to formulate structures for the unknown underlying dependence in the data.

2.3 Shrinkage estimators based on the sample covariance matrix

2.3.1 Shrinking the spectrum and the correlation matrix

Stein [1975] observed that the sample covariance matrix systematically distorts the eigenstructure of Σ, especially when M is large. His work spurred efforts in the improvement of S, which he did by simply shrinking its eigenvalues. He considered estimators of the form

$$\hat{\Sigma} = \Sigma(S) = P\Phi(\lambda)P', \tag{13}$$

where $\lambda = (\lambda_1, \ldots, \lambda_M)'$, $\lambda_1 > \cdots > \lambda_M$ are the ordered eigenvalues of S, P is the orthogonal matrix whose i^{th} column is the normalized eigenvector of S corresponding to λ_i, and $\Phi(\lambda) = diag(\phi_1, \ldots, \phi_M)$ is the diagonal matrix where $\phi_j(\lambda)$ is an estimate of the j^{th} largest eigenvalue of Σ. Letting $\phi_j(\lambda) = \lambda_j$ corresponds to the usual unbiased estimator S. It is known that λ_1 and λ_M are biased low and high, respectively, so Stein chooses $\Phi(\lambda)$ to shrink the eigenvalues toward central values to counteract the biases of the sample eigenvalues. The modified estimators of the eigenvalues of Σ are given by $\phi_j = \frac{N\lambda_j}{\alpha_j}$, where

$$\alpha_j(\lambda) = N - M + 2\lambda_j \sum_{i \neq j} \frac{1}{\lambda_j - \lambda_i}. \tag{14}$$

The Stein estimators ϕ_j differ from the sample eigenvalues when are nearly equal and N/M is not small. The work of Lin [1985] includes an algorithm to modify any ϕ_j's which are negative and or do not not satisfy $\phi_1 < \cdots < \phi_M$.

2.3.2 Ledoit-Wolf shrinkage estimator

The estimator proposed by Ledoit and Wolf [2004] is motivated by the fact that the sample covariance matrix is unbiased but has high variance - the risk associated with S is considerable when $M >> N$, and even in cases when the dimension is close to the sample size. In contrast, very little estimation error is associated with a highly structured estimator of a covariance matrix, like those presented in Section 2.2, but when the model is misspecified, these can exhibit severe bias. A natural inclination is to define an estimator as a linear combination of the two extremes, letting

$$\hat{\Sigma} = \alpha_1 I + \alpha_2 S, \tag{15}$$

where α_1, α_2 are chosen to optimize the Frobenius norm of $\hat{\Sigma} - S$ or the slightly modified Frobenius norm:

$$L\left(\hat{\Sigma}, \Sigma\right) = M^{-1}||\hat{\Sigma} - \Sigma||^2 = M^{-1}\text{tr}\left(\hat{\Sigma} - \Sigma\right)^2.$$

They show that the optimal α_i depend on only four characteristics of the true covariance matrix:

$$\begin{aligned} \mu &= \text{tr}\left(\Sigma\right)/M, \\ \alpha^2 &= ||\Sigma - \mu I||^2, \\ \beta^2 &= ||S - \Sigma||^2, \\ \delta^2 &= ||S - \mu I||^2. \end{aligned} \tag{16}$$

Ledoit and Wolf [2004] give consistent estimators of these quantities, so that substitution of these in $\hat{\Sigma}$ produces a positive definite estimator of Σ. They demonstrate the superiority of their estimator to several others including the sample covariance matrix and the empirical Bayes estimator (Haff [1980]).

2.3.3 Elementwise shrinkage

A broad class of estimators that aim to stabilize the sample covariance matrix do so by applying shrinkage elementwise to the same covariance matrix. Shrinking the elements of the sample covariance matrix has been approached in a multitude of ways, including banding, tapering, and thresholding. These estimators are computationally inexpensive, with the exception of cross validation necessary for smoothing parameter selection. The tradeoff accompanying the ease of computation is that, because transformations of sample estimates are elementwise, the resulting estimators are not guaranteed to be positive definite.

2.3.4 Tapering and banding the sample covariance matrix

The sample covariance matrix is unstable when the dimension of the data M is larger than the sample size N, and even when the sample size is larger than the dimension of the data many entries of the sample covariance matrix $S = (s_{ij})$ could be small. Setting certain entries to zero is one approach to reducing parameter dimension to stabilize estimates. In time series analysis, one observes a sample size of $N = 1$: the data is a single, long realization. Assuming stationarity of the process reduces the number of distinct parameters of the $M \times M$ covariance matrix Σ from $M\left(M+1\right)/2$ to M, which could be large yet. Moving average (MA) and autoregressive (AR) models reduce the number of parameters in the same way as banding a covariance or inverse covariance matrix. Bickel and Levina [2008]; Wu and Pourahmadi [2009]. For a given sample covariance matrix $S = (s_{ij})$ and integer k, $0 < k < M$, the k-banded sample covariance matrix is given by

$$B_k\left(S\right) = \left[s_{ij} 1\left(|i-j| \leq k\right)\right] \tag{17}$$

This kind of regularization is ideal when the indices have been arranged so that

$$|i-j| > k \Rightarrow \sigma_{ij} = 0,$$

which is applicable if, for example, y_t, $t = 1, \ldots, M$ follow a finite heterogeneous moving average process

$$y_t = \sum_{j=1}^{k} \theta_{t,t-j}\epsilon_j,$$

where the ϵ_j's are iid mean zero errors having finite variance. Banding estimators are a special case of tapering estimators, which have the form

$$\hat{\Sigma} = R * S \tag{18}$$

where R is a positive definite tapering matrix, and the $(*)$ operator denotes the Schur matrix multiplication (the element-wise matrix product). The Schur product of two positive definite matrices is also guaranteed to be positive definite, so the tapering estimator's positive definiteness is dependent on the choice of tapering matrix R. Banding the sample covariance matrix is equivalent to premultiplying S by

$$R = (r_{ij}) = (1\,(|i-j| \leq k)),$$

which is not positive definite. However, several have used the same concept on the lower triangular matrix of the Cholesky decomposition of Σ^{-1}, including Wu and Pourahmadi [2003], Huang et al. [2006], Levina et al. [2008]. Banding the Cholesky factor mitigates the need for the tapering matrix to be positive definite, since the parameters of the reparameterization are completely free while still guaranteeing that the estimate is positive definite. Detailed discussion follows in Section 2.4.4.

When N, M, and k are large, asymptotic analysis of banding estimators is available. Bickel and Levina [2008] establish consistency of the banded estimator in the operator norm, and uniform consistency over the class of "approximately bandable" matrices under a normal likelihood. Convergence requires that $\log M/N \to 0$, and they derive an explicit rate of convergence which depends on the rate at which k grows. Cai et al. [2010] proposed the following tapering estimator of the sample covariance matrix:

$$S^{\omega} = \left[\omega_{ij}^{k} s_{ij}\right], \tag{19}$$

where the ω_{ij}^{k} are given by

$$\omega_{ij}^{k} = k_h^{-1}\left[(k - |i-j|)_+ - (k_h - |i-j|)_+\right],$$

The weights ω_{ij}^{k} are indexed with superscript to indicate that they are controlled by a tuning parameter, k, which can take integer values between 0 and M, the dimension of the covariance matrix. Without loss of generality, we assume that $k_h = k/2$ is even. The weights may be rewritten as

$$\omega_{ij} = \begin{cases} 1, & ||i-j|| \leq k_h \\ 2 - \frac{i-j}{k_h}, & k_h < ||i-j|| \leq k, \\ 0, & \text{otherwise} \end{cases}$$

This expression of the weights makes it clear how the selection of k controls the amount of shrinkage applied to a particular element of the sample covariance matrix. Elements of S belonging to

the subdiagonals closest to the main diagonal are left unregularized. The shrinkage applied to elements increases as we move away from the diagonal: a multiplicative shrinkage factor of $2 - \frac{i-j}{k_h}$ is applied to elements belonging to subdiagonals $k_h, \ldots, k-1, k$, and elements further than k subdiagonals from the main diagonal are shrunk to zero. Cai et al. [2010] derived optimal rates of convergence under the operator norm for their estimator and presented simulations demonstrating that it nearly uniformly outperforms the banding estimator of Bickel and Levina [2008].

2.3.5 thresholding the sample covariance matrix

When both N and M are large, it is reasonable to assume that Σ is sparse, so that many elements of the covariance matrix are equal to 0. In this case, setting certain elements of sample estimates to zero can improve the quality of estimators. Thresholding was originally a method developed in nonparametric function estimation, but recently Bickel et al. [2008] and Rothman et al. [2009] have utilized thresholding for estimating large covariance matrices. For $\lambda > 0$, a thresholding operator $s_\lambda(z) : \Re \to \Re$ satisfies

- $s_\lambda(z) \leq z$;
- $s_\lambda(z) = 0$ for $|z| \leq \lambda$;
- $|s_\lambda(z) - z| \leq \lambda$

Shrinkage and thresholding estimators can be viewed as the solution to the problem of minimizing a penalized quadratic loss function, and since the thresholding operator is applied elementwise to the sample covariance S, these optimization problems are univariate. A generalized thresholding estimator $s_\lambda(z)$ is the solution to

$$s_\lambda(z) = \arg\min_\sigma \left[\frac{1}{2}(\sigma - z)^2 + J_\lambda(\sigma)\right] \tag{20}$$

For detailed discussion of the connection between penalty functions and the resulting thresholding rules, see Antoniadis and Fan [2001]. Soft thresholding results from minimizing 20 using the lasso penalty, $J_\lambda = \lambda|\sigma|$, which corresponds to thresholding rule

$$s_\lambda(z) = \text{sign}(\sigma)(\sigma - \lambda)_+. \tag{21}$$

Rothman et al. [2009] presented a class of generalized thresholding estimators, including the soft-thresholding estimator given by

$$S^\lambda = \left[\text{sign}(s_{ij})(s_{ij} - \lambda)_+\right],$$

where σ^*_{ij} denotes the i-j^{th} entry of the sample covariance matrix, and λ is a penalty parameter controlling the amount of shrinkage applied to the empirical estimator. These estimators are simple to compute compared to competitor estimates like the penalized likelihood with LASSO penalty, but they suffer from the lack of guaranteed positive definiteness. However, similar to the result for banded estimators, Bickel et al. [2008] have established the consistency of the threshold

estimator in the operator norm, uniformly over the class of matrices that satisfy a certain sparsity requirement.

Alternately, for estimating the covariance of a random vector which is assumed to have a natural (time) ordering, several have proposed applying kernel smoothing methods directly to elements of the sample covariance matrix or a function of the sample covariance matrix. Zeger and Diggle [1994] introduced a nonparametric estimator obtained by kernel smoothing the sample variogram and squared residuals. Yao et al. [2005] applied a local linear smoother to the sample covariance matrix in the direction of the diagonal and a local quadratic smoother in the direction orthogonal to the diagonal to account for the presence of additional variation due to measurement error. The latter work is one of the few nonparametric methods utilizing smoothing in both dimensions of the covariance matrix, which was an inspiration of sorts for the work we present in Chapter 2. Like other elementwise shrinkage estimators, however, their proposed estimator is not guaranteed to be positive definite.

2.3.6 Tuning parameter selection for element-wise shrinkage estimators

The performance of any regularized estimator depends heavily on the quality of tuning parameter selection. The Frobenius is a natural measure of the accuracy of an estimator; it quantifies the sum over the unique elements of Σ of the the first term in 20,

$$||\hat{\Sigma}^{\lambda} - \Sigma||^2 = \left(\sum_{i,j} \left(\hat{\sigma}_{ij}^{\lambda} - \sigma_{ij} \right)^2 \right)^{1/2} \tag{22}$$

If Σ were available, one would choose the value of the tuning parameter λ which minimizes **??**. In practice, one tries to first approximate the risk, or

$$E_{\Sigma} \left[||\hat{\Sigma}^{\lambda} - \Sigma||^2 \right],$$

and then choose the optimal value of λ. As in regression methods, cross validation and a number of its variants have become popular choices for tuning parameter selection in covariance estimation, though unanimous agreement on which precise procedure is optimal is fleeting. K-fold cross validation requires first splitting the data into folds $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$. The value of the tuning parameter is selected to minimize

$$\text{CV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(-k)} - \tilde{\Sigma}^{(k)}||_F^2, \tag{23}$$

where $\tilde{\Sigma}^{(k)}$ is the unregularized estimator based on based on $\mathcal{D}_k$, and $\hat{\Sigma}^{(-k)}$ is the regularized estimator under consideration based on the data after holding $\mathcal{D}_k$ out. Using this approach, the size of the training data set is approximately $(K-1)N/K$, and the size of the validation set is approximately N/K (though these quantities are only relevant when subjects have equal numbers of observations). For linear models, it has been shown that cross validation is asymptotically

consistent is the ratio of the validation data set size over the training set size goes to 1. See Shao [1993]. This result motivates the reverse cross validation criterion, which is defined as follows:

$$\text{rCV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(k)} - \tilde{\Sigma}^{(-k)}||_F^2, \tag{24}$$

where $\tilde{\Sigma}^{(-k)}$ is the unregularized estimator based on based on the data after holding out $\mathcal{D}_k$, and $\hat{\Sigma}^{(k)}$ is the regularized estimator under consideration based on $\mathcal{D}_k$.

2.4 Matrix decompositions

The most methodic and successful approaches to covariance modeling is to decompose the covariance matrix into its variance and dependence components. The following section demonstrates the role of multiple matrix parameterizations in removing the positive definite constraint that poses a challenge in most covariance estimation settings.

2.4.1 The variance-correlation decomposition

The variance-correlation decomposition of Σ is perhaps the most familiar of the following three parameterizations, which parameterizes the covariance matrix according to

$$\Sigma = DRD, \tag{25}$$

where $D = \text{diag}\left(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{MM}}\right)$ denotes the diagonal matrix with diagonal entries equal to tje square-roots of those of Σ, and R is the corresponding correlation matrix. This parameterization enjoys attractive practicality because the standard deviations are on the same scale as the responses, and because the estimation of D and R can be separated by eithering iteratively fixing one sequence of parameters to estimate the other. Moreover, one set of parameters may be more important than the others in some applications; the dynamic correlation model presented in Engles (2002) is actually motivated by the fact that variances (volatilities) of individual assets are more important than their time-varying correlations.

While the diagonal entries of D are constrained to be nonnegative, their logarithms are unconstrained. However, the correlation matrix R is positive-definite constrained to have unit diagonal entires and off-diagonal entries to be less than or equal to 1 in absolute value. Because of these constraints, the variance-correlation decomposition does not lend to modeling its components with the use of covariates.

2.4.2 Gaussian graphical models

The marginal (pairwise) dependence among the entries of a random vector are captured by the off-diagonal entries of Σ or the entries of the correlation matrix $R = (\rho_{ij})$. However, the conditional

dependencies can be found in the off-diagonal entries of the precision matrix $\Sigma^{-1} = (\sigma^{ij})$. More precisely, for Y a mean zero normal random vector with a positive-definite covariance matrix, if the (i, j) component of the precision matrix is zero, then given the other variables, y_i and y_j are conditionally independent (Anderson [1984]).

Graphical models are a common way of representing the conditional independence structure in Y, with the nodes of the graph corresponding to variables. The absence of an edge between variables i and j, or a zero in the (i, j) position of the inverse covariance matrix indicates that the two variables are conditionally independent. The entries of the variance-correlation decomposition of the precision matrix

$$\Sigma^{-1} = \left(\sigma^{ij}\right) = \tilde{D}\tilde{R}\tilde{D} \tag{26}$$

can be interpretted as certain coefficients of a regression model. A number of regression-based approaches to modeling the precision structure have spawned from the work of Meinhausen and Buhlmann [2006]. Their method is based on solving M separate LASSO regression problems. The entries of $\left(\tilde{R}, \tilde{D}\right)$ have direct statistical interpretations in terms of partial correlations, and variance of predicting a variable given the rest. Regression calculations can be used to show that the partial correlation coefficient between y_i and y_j after removing the linear effect of the $M - 2$ remaining variables is given by

$$\tilde{\rho}_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}. \tag{27}$$

The partial variance of y_i after removing the linear effect of the remaining $M-$ variables is given by

$$\tilde{d}_{ii}^2 = \frac{1}{\sigma^{ii}}. \tag{28}$$

To connect these parameters to those of a regression model, consider partitioning random vector $Y = (y_1, \ldots, y_M)'$ into two components $(Y_1', Y_2')'$ of dimensions M_1 and M_2, and similarly partitioning its covariance and precision matrices:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{29}$$

Let $\Phi_{2|1}$ denote the $M_2 \times M_1$ matrix of regression coefficients resulting from the least squares regression of Y_2 on Y_1, and let $e_{2|1} = Y_2 - \Phi_{2|1}Y_1$ denote the corresponding vector of residuals. The regression coefficients $\Phi_{2|1}$ and residuals $e_{2|1}$ are obtained from restricting $e_{2|1}$ to be uncorrelated with Y_1:

$$\begin{aligned} \Phi_{2|1} &= \Sigma_{21}\Sigma_{11}^{-1} \\ &= -\left(\Sigma^{22}\right)^{-1}\Sigma^{21} \end{aligned} \tag{30}$$

$$\begin{aligned} Cov\left(e_{2|1}\right) &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\ &= \Sigma_{22|1} = \left(\Sigma^{22}\right)^{-1}. \end{aligned} \tag{31}$$

If we let $M_2 = 1$, then one can establish the relationship between elements of the inverse covariance matrix and these regression coefficients and conditional covariances. When $Y_1 = Y_{-(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_M)'$ and Y_2 corresponds to a single y_i, $\Sigma_{22|1}$, a scalar, is referred to as the *partial variance* of y_i given the other variables. Denote the linear least squares predictor of y_i based on $Y_{-(i)}$ by y_i^* and $\epsilon_i^* = y_i - y_i^*$ with prediction variance $Var\left(\epsilon_i^*\right) = {d^*}_i^2$. Then

$$y_i = \sum_{j \neq i} \beta_{ij} y_j + \epsilon_i^*,$$

where (31) and (32) give

$$\begin{aligned} \beta_{ij} &= -\frac{\sigma^{ij}}{\sigma^{ii}}, \quad j \neq i \\ {d^*}_i^2 &= Var\left(y_i | y_j\right) = \frac{1}{\sigma_{ii}}, \quad j \neq i, \;\; i = 1, \ldots, M \end{aligned} \tag{32}$$

Thus, the unconstrained regression coefficient of the j^{th} variable when we regressing y_i on the rest of the variables is given by the (i, j) entry of the inverse covariance matrix. The partial correlation between y_i and y_j can be defined if we consider the case where $M_2 = 2$. Letting $Y_2 = (y_i, y_j)'$, $i \neq j$ and $Y_1 = Y_{-(ij)}$ contain the remaining $M - 2$ variables, the covariance of (y_i, y_j) after removing the linear effects of $\{y_k : k \neq i, j\}$ is given by

$$\begin{aligned} \Sigma_{22|1} &= \begin{bmatrix} \sigma^{ii} & \sigma^{ij} \\ \sigma^{ji} & \sigma^{jj} \end{bmatrix}^{-1} \\ &= \frac{1}{\sigma^{ii}\sigma^{jj} - \left(\sigma^{ij}\right)^2} \begin{bmatrix} \sigma^{jj} & -\sigma^{ij} \\ -\sigma^{ij} & \sigma^{ii} \end{bmatrix} \end{aligned}$$

The regression coefficients (32) can be written in terms of the partial correlation between y_i and y_j:

$$\rho_{ij}^* = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{ij}}}. \tag{33}$$

Rewriting the β_{ij}, we have

$$\beta_{ij} = \rho_{ij}^* \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}, \tag{34}$$

which shows that the sparsity of the inverse covariance matrix mirrors that of the matrix of partial correlations. This parallel motivates estimation of the inverse covariance matrix by fitting a sequence of penalized regression models, notably the approach taken by Peng et al. [2012] which imposes a Lasso penalty on the off-diagonal elements of the partial correlation matrix.

2.4.3 The spectral decomposition

The spectral decomposition is the basis of several methods in multivariate statistics, including principal component analysis and factor analysis. See Anderson [1984], (Hotelling, 1933). The spectral decomposition of a covariance matrix Σ is given by

$$\Sigma = P\Lambda P' = \sum_{i=1}^{M} \lambda_i e_i e_i', \tag{35}$$

where Λ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_M$, and P is the orthogonal matrix of normalized eigenvectors, having e_i as its i^{th} column. The entries of Λ and P can be interpreted as thevariances and coefficients of the M principal components. The matrix P is constrained by its orthogonality, its use within the framework of GLM or alongside covariates in an effort to reduce parameter dimension is inconvenient. In spite of this, Chiu et al. [1996] proposed an new unconstrained reparameterization of a covariance matrix using the spectral decomposition, modeling the matrix logarithm:

$$\log \Sigma = P \log \Lambda P' = \sum_{i=1}^{M} \log(\lambda_i)\, e_i e_i', \tag{36}$$

This decomposition is particularly interesting because it highlights a tradeoff between the requirements for unconstrained parameterization of covariance matrices and the statistical interpretability of the corresponding parameters. The components of the matrix logarithm, $\log \lambda_i$, are free, but lack any relevant statistical interpretability. We further discuss the log-linear GLM for covariance matrices in Section 2.5.2 .

2.4.4 The Cholesky decomposition

The Cholesky decomposition of a positive-definite matrix has the form

$$\Sigma = CC', \tag{37}$$

where $C = (c_{ij})$ is a unique lower-triangular matrix with positive diagonal entries. This factorization is frequently encountered in optimization techniques and matrix computation; see Golub and Van Loan [2012]. It is difficult to attach any statistical interpretation to the entries of C in this form Pinheiro and Bates [1996]. But by transforming C to unit lower-triangular matrices, statistically interpreting of the diagonal entries of C and the resulting unit lower-triangular matrix is much easier. To do this, one must simply divide the i^{th} column of C by its i^{th} diagonal element c_{ii}. Letting $D^{1/2} = diag(c_{11}, \ldots, c_{MM})$, the standard Cholesky decomposition 37 can be written

$$\Sigma = CD^{-1/2}D^{1/2}D^{1/2}D^{-1/2}C' = LDL', \tag{38}$$

where $L = D^{-1/2}C$. This is commonly referred to as the modified Cholesky decomposition (MCD) of Σ. We can also write the modified Cholesky decomposition of the inverse covariance matrix:

$$D = T\Sigma T', \quad \Sigma^{-1} = T'D^{-1}T, \tag{39}$$

where $T = L^{-1}$. Like P as in the spectral decomposition, the lower triangular matrix T diagonalizes Σ. However, the Cholesky decomposition is perhaps more attractive since unlike the entries of the orthogonal matrix of the spectral decomposition, the entries of T are unconstrained, and furthermore, have a specific statistical interpretation.

Like the variance-correlation decomposition of the inverse covariance matrix 26, the Cholesky factor T and diagonal matrix D can be constructed using components of a regression model. Consider regressing y_t on its predecessors $y_1, \ldots, y_{t-1}$. Let $Y = (y_1, \ldots, y_M)'$ denote a mean zero random vector with positive definite covariance matrix Σ, and let $\hat{y}_t$ be the linear least-squares predictor of y_t based on previous measurements $y_{t-1}, \ldots, y_1$. Let ϵ_t denote the corresponding prediction residual having variance $\sigma_t^2 = Var(\epsilon_t)$. Standard regression machinery gives us that there exist unique scalars ϕ_{tj} so that

$$y_t = \sum_{j=1}^{t-1} \phi_{t,j} y_j + \sigma_t \epsilon_t, \quad t = 2, \ldots, M \tag{40}$$

where

$$\epsilon_t = \begin{cases} y_t - \hat{y}_t, & t > 1 \\ y_t, & t = 1 \end{cases}$$

are i.i.d. mean zero random variables with unit variance. The connection between the Cholesky decomposition and the autoregressive model (2.4.4) is established by noting that the Cholesky factor contains the negatives of the regression coefficients and the prediction error variances are the diagonal elements of D. Let $\epsilon = (y_1, \ldots, y_M)'$ denote the vector of uncorrelated prediction residuals with

$$Cov(\epsilon) = D = diag\left(\sigma_1^2, \ldots, \sigma_M^2\right)'.$$

Then model (2.4.4) can be written in vector form $\epsilon = TY$, where the (t, j) entry of T is $-\phi_{tj}$, and the (t, t) entry of D is the t^{th} prediction variance $\sigma_t^2 = var(\epsilon_t)$.

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{m1} & -\phi_{m2} & \ldots & -\phi_{m,m-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \ddots \\ y_m \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \ddots \\ \epsilon_m \end{bmatrix} \tag{41}$$

Table 1 illustrates how the components of a covariance matrix are obtained through successive regressions. Specifically, this representation demonstrates how modeling a covariance matrix is equivalent to fitting a sequence of $M - 1$ varying-coefficient and varying-order regression models. Since the ϕ_{ij}s are regression coefficients, for any unstructured covariance matrix, these and the log innovation variances are unconstrained. The regression coefficients of the model in () are referred

to as the *generalized autoregressive parameters* (GARP) and *innovation variances* (IV) (Pourahmadi [1999], Pourahmadi [2000]). The powerful implication of the parallel regression framework of decomposition (39) is the accessibility of the entire portfolio of regression methods for the service of modeling covariance matrices. Moreover, the estimator $\hat{\Sigma}^{-1} = \hat{T}'\hat{D}^{-1}T$ constructed from the unconstrained parameters ϕ_{ij}, σ_j^2 is guaranteed to be positive definite.

Table 1: Autoregressive coefficients and prediction error variances of successive regressions.

y_1	y_2	y_3	$\cdots$	y_{m-1}	y_m
1					
ϕ_{21}	1				
ϕ_{31}	ϕ_{32}	1			
$\vdots$	$\vdots$		$\ddots$		
$\vdots$	$\vdots$			$\ddots$	
ϕ_{m1}	ϕ_{m2}	$\cdots$	$\cdots$	$\phi_{m,m-1}$	1
σ_1^2	σ_1^2	$\cdots$	$\cdots$	σ_{m-1}^2	σ_m^2

2.5 Generalized linear models for covariances

Modeling covariance matrices in a systematic, data-driven manner is impeded by the positive-definiteness constraint and high-dimensionality; however, similar (albeit simpler) hurdles in modeling the mean vector μ of the distribution of a random vector $Y = (y_1, \ldots, y_M)'$ has been successfully handled in the context of regression analysis. The resulting techniques have lead to the framework of generalized linear models (GLM), which enjoys a rich and extensive theoretical foundation. The success of GLMs is in most part due to the use of a link function $g(\cdot)$ and a linear predictor $g(\cdot) = X\beta$, which induces an unconstrained parameterization and reduces the parameter space dimension simultaneously. Since the covariance matrix of a random vector Y , defined by $\Sigma = E(Y - \mu)(Y - \mu)$, is a mean-like parameter, one would like to exploit the idea of GLM along with the experience and progress in fitting the mixed-effects and time series models in developing a systematic, data-based procedure for covariance matrices.

Approaches to modeling covariances with the explicit use covariates has been extensively explored in the time series literature, while the implicit use of covariates for covariance modeling has been the focus of many in the area of variance components; see Klein [1997] and Searle et al. [2009]. Time series techniques based on spectral and Cholesky decompositions provide the necessary tools for handling the cumbersome positivedefiniteness constraint on a stationary covariance matrix or covariance function. In the GLM setting, simply applying a link function componentwise to the potentially constrained mean vector μ permits its unconstrained estimation. Unfortunately employing the same precise approach to covariance matrices isn't viable since positive-definiteness

is a simultaneous constraint on all entries of a matrix. Successfully modeling a general covariance structure almost necessitates decomposing a covariance matrix into its "variance" and "dependence" components because of its inherent complicated structure. The three major methods for performing such decompositions include the variance-correlation decomposition, the spectral decomposition, and the Cholesky decomposition. Section 2.4.4 touched on the attractive properties of the latter that lead to advantages over the other two covariance parameterizations.

2.5.1 Linear models for covariance

Gabriel [1962] was among the first to implicitly parameterize a multivariate normal distribution in terms of entries of the precision matrix Ω^{-1}. Dempster (1972) who recognized the entries of $\Sigma^{-1} = (\sigma^{ij})$ as the canonical parameters of the exponential family of normal distributions with mean zero and unknown covariance matrix Σ:

$$\log f\left(Y, \Sigma^{-1}\right) = -\frac{1}{2}\text{tr}\Sigma^{-1}\left(Y'Y\right) + \log\left|\Sigma\right|^{-1/2} - M\log\sqrt{\pi}$$

Soon thereafter, the simple structures of time series and variance components models motivated Anderson [1973] to define the class of linear covariance models:

$$\Sigma = \sum_{i=1}^{q} \alpha_q U_q \tag{42}$$

where the U_is are known symmetric matrices and the α_is are unknown parameters, restricted to ensure that Σ is positive definite. This class of models is general enough to include all linear mixed effects models as well as certain time series and graphical models. In, for q large enough, any covariance matrix admits representation of the form (**??**), since one can decompose every covariance matrix as

$$\Sigma = \sum_{i=1}^{M}\sum_{j=1}^{M} \sigma_{ij} U_{ij}, \tag{43}$$

where U_{ij} is an $M \times M$ matrix with a 1 in the (i, j) position, and zeros everywhere else. The linear model (42) can be viewed as modeling the link-transformed covariance $g\left(\Sigma\right) = \sum_{i=1}^{q} \alpha_q U_q$, where $g\left(\cdot\right)$ is the identity link. Despite the convenience of parameterization, the positive definite constraint (2) makes estimation an arduous task.

Inducing sparsity by setting certain elements of the covariance matrix or its inverse to zero is a common approach to reducing the dimensionality of a covariance structure. Inspection of model (42) and the covariance parameterization given in (43) makes it easy to see that this can be achieved by eliminating certain U_{ij} from the covariates in the linear covariance model. On the extreme end of the sparsity spectrum is the case of independent observations and Σ is diagonal, eliminating all U_{ij} from the linear model covariates for $i \neq j$. Connection between the linear covariance model and other models for covariance discussed in previous sections can be established if we consider

intermediary cases, such as classes of stationary moving average (MA) and autoregressive (AR) models introduced in the early times series literature. The $MA(q)$ model corresponds to a banded covariance matrix, setting

$$\sigma_{ij} = 0 \quad \text{for } |i-j| > q, \tag{44}$$

while the $AR(p)$ model corresponds to a banded inverse:

$$\sigma^{ij} = 0 \quad \text{for } |i-j| > p. \tag{45}$$

Of course, there are the nonstationary analogues to these classes of models, some of which were discussed in Section **??**. We will review others which are related to antedependence models and Gaussian graphical models. Random variables $y_1, \ldots, y_M$, which correspond to observation times $t_1, \ldots, t_M$, with multivariate normal joint distribution said to be p^{th}-order antedependent or $AD(p)$ Gabriel [1962] if y_t and y_{t+s+1} are independent given the intervening values $y_{t+1}, \ldots, y_{t+s}$ for $t = 1, \ldots, p?s?1$ and all $s \geq p$. A random vector $Y = (y_1, \ldots, y_p)$ is $AD(p)$ if and only if its covariance matrix satisfies (45). Closely connected are the classes of variable order AD models and varying order, varying coefficient autoregressive models Kitagawa and Gersch [1985] in which the coefficients and order of antedependence depend on time.

2.5.2 Log-linear covariance models

The constraint on the α_is in (42) was eliminated with the introduction of log-linear covariance models (Chiu et al. [1996], Pinheiro and Bates [1996].) For a general covariance matrix having spectral decomposition

$$\Sigma = P\Lambda P', \tag{46}$$

its matrix logarithm, denoted $\log \Sigma$, and defined by $log\Sigma = P \log \Lambda P'$ is a symmetric matrix with unconstrained entries taking values in $\Re$. Application of the log-link function leads to the log-linear model for Σ:

$$g(\Sigma) = \log \Sigma = \sum_{i=1}^{q} \alpha_i U_i, \tag{47}$$

where the U_is are as before in 42 and the α_is are now unconstrained. The α_is, however, now lack statistical interpretation since $g(A) = \log A$ is a highly nonlinear operation. But for diagonal Σ, $\log \Sigma = \text{diag}(\sigma_{11}, \ldots, \sigma_M M)$, and model 47 reduces to modeling of heterogeneous variances, which has been extensively studied. Detailed presentation is given in Carroll and Ruppert [1988], Verbyla [1993] and in references therein.

Rice and Silverman [1991] were the first to pursue nonparametric estimation of the spectral decomposition for functional data, which arise from experiments which produce observed responses in the form of curves. See Ramsay [2006], Ramsay and Silverman [2007]. The covariance structure is estimated via functional principal component analysis (fPCA); principal components of functional data are estimated using penalized least squares of the normalized eigenvectors, subject to the orthogonality constraint. Additionally, Boente and Fraiman [2000] proposeds kernel-based

PCA, but maintaining orthogonality of the smooth principal components remains a major computational challenge in both approaches.

The log link resolves the issued presented by the constrained parameter space associated with the identity link, leading to unconstrained parameterization of a covariance matrix. However, the parameters of the matrix logarithm lack any meaningful statistical interpretation. The hybrid link constructed from the modified Cholesky decomposition of Σ^{-1} given in 48 combines ideas in Edgeworth [1892], Gabriel [1962], Anderson [1973], Dempster [1972], Chiu et al. [1996], and Zimmerman and Núñez-Antón [1997]. It leads to unconstrained and statistically meaningful reparameterization of the covariance matrix so that the ensuing GLM overcomes most of the shortcomings of the linear and log-linear models. For an unstructured covariance matrix Σ, the nonredundant entries of the components $(T, \log D)$ of the modified Cholesky decompostion 39 can be written as the entries of

$$g(\Sigma) = 2I - T - T' + \log D. \tag{48}$$

These entries are unconstrained, allowing them to be modeled using any desired technique, including parametric, semi- and nonparametric, and Bayesian approaches. Including covariates in any proposed model for these components can be done so seamlessly. As in the usual GLM setting for estimation of the mean, one can elicit parametric models for ϕ_{tj} and $\log \sigma_t^2$. For example, one might model the nonredundant entries of T, say, linearly as in model 42 and those of $\log D$ as in, say, model 47, letting

$$\begin{aligned} \phi_{tj} &= x_{tj}'\beta, \\ \log \sigma_t^2 &= z_t'\gamma, \end{aligned} \tag{49}$$

where x_{tj} and z_t denote $q \times 1$ and $p \times 1$ vectors of known covariates, and $\beta = (\beta_1, \ldots, \beta_q)'$ and $\gamma = (\gamma_1, \ldots, \gamma_p)'$ are the parameters relating these covariates to the innovation variances and the dependence among the elements of Y. Covariates most frequently used in the analysis of real longitudinal data sets are low order polynomials of lag and time, modeling

$$\begin{aligned} z_{jk}' &= \left(1, t_j - t_k, (t_j - t_k)^2, \ldots, (t_j - t_k)^{p-1}\right)' \\ z_i' &= \left(1, t, \ldots, t^{q-1}\right)' \end{aligned} \tag{50}$$

Pourahmadi [1999], Pourahmadi [2000], and Pan and [2006] prescribe methods for identifying models such as model 49 using model selection criteria, such as AIC, and regressograms, which are a nonstationary analogue of the correlelogram one typically encounters in the time series literature. Pan and Mackenzie [2003] jointly estimate the mean and covariance of longitudinal data using maximum likelihood, iterating between estimation of the mean vector μ, the log innovation variances $\log \sigma_{ij}^2$, and the generalized autoregressive parameters ϕ_{ij}. Score functions can be computed by direct differentiation of the normal log likelihood, and optimization is achieved by solving these via iterative quasi-Newton method. Modeling the covariance in such a way is reduces a potentially high dimensional problem to something much more computationally feasible;

if one models the innovation variances $\sigma^2(t)$ similarly using a d-dimensional vector of covariates, the problem reduces to estimating $q + d$ unconstrained parameters, where much of the dimensionality reduction is a result of characterizing the GARPs in terms of only the difference between pairs of observed time points, and not the time points themselves. This model specification of ϕ is equivalent to specifying a Toeplitz structure for Σ. An $M \times M$ Toeplitz matrix Σ is a matrix with elements σ_{ij} such that $\sigma_{ij} = \sigma_{|i-j|}$ i.e. a matrix of the form (8), having entries which are constant on each subdiagonal.

The estimated covariance matrix may be considerably biased when the specified parametric model is far from the truth. To avoid model misspecification, many have alternatively proposed nonparametric and semiparametric techniques approaches to estimation. When the data $Y_1, \ldots, Y_N$ are a random sample of M-dimensional vectors from a mean zero multivariate normal population with common covariance matrix Σ parameterized as $D = T'\Sigma T$, the form of the likelihood allows for relatively simple computation of the MLE of the parameters. Up to a constant, the log likelihood is given by

$$\begin{aligned} -2\ell(Y_1, \ldots, Y_N, \Sigma) &= \sum_{i=1}^{N} \left(\log|\Sigma| + Y_i' \Sigma^{-1} Y_i' \right) \\ &= N \log|D| + N \text{tr} \Sigma^{-1} S \\ &= N \log|D| + N \text{tr} D^{-1} T S T', \end{aligned} \tag{51}$$

where $S = N^{-1} \sum_{i=1}^{N} Y_i Y_i'$. The negative log likelihood (51) is quadratic in T for fixed D, so the MLE for the ϕ_{ij} has closed form. Similarly, the MLE for D for fixed T has closed form. See Pourahmadi [2000]. While the MLE is flexible and thus exhibits low bias, this advantage can be offset with high variance, so to balance the tradeoff between bias and variance, shrinkage or regularization may be applied to estimates to improve stability of estimators.

The fact that the entries of T are unconstrained makes the Cholesky decomposition ideal for nonparametric estimation and regularization methods. Wu and Pourahmadi [2003] proposed local polynomial smoothers to individually estimate the subdiagonals of T. The idea of smoothing along the subdiagonals rather than down the rows or columns, or viewing T as a bivariate function is analogous to the successive regressions in (2.4.4). A similar procedure by Dahlhaus et al. [1997] uses varying coefficient regression models for each subdiagonal of T:

$$y_t = \sum_{j=1}^{t-1} f_{j,M}(t/M)\, y_{t-j} + \sigma_M(t/M)$$

Wu and Pourahmadi [2003] give details of smoothing and selection of the order k of the autoregression under the assumption that the N subjects share common observation times. In the first step, they derive a raw estimate of the covariance matrix and the estimated covariance matrix is subject to the modified Cholesky decomposition. In the second step, they apply local polynomial smoothing to the diagonal elements of D and the subdiagonals of T. Their procedure is not

capable of handling missing or irregular data. Huang et al. [2007] jointly model the mean and covariance matrix of longitudinal data using basis function expansions. They treat the subdiagonals of T as smooth functions, approximated by B-splines and carry out estimation maximum (normal) likelihood. Their method permits subject-specific observations times, but assumes that observation times lie on some notion of a regular grid. They treat within-subject gaps in measurements as missing data and which they handle using the E-M algorithm. Regularization is achieved through the choice of k, the number of nonzero subdiagonals, and the total number of basis functions used to approximate the k smoothed diagonals. They treat these as tuning parameters and use BIC for model selection. Due to the closer connection between entries of T and the family of regression (2.4.4), it is conceivable that T exhibits sparsity, having some of its entries could be zero or close to it. **?** propose a prior distribution that allows for zero entries in T and have obtained a parsimonious model for Σ without assuming a parametric structure. Similar results are reported in Huang et al. [2006] using penalized likelihood with L_1-penalty to estimate T for Gaussian data. Levina et al. [2008] impose a banded structure on the Cholesky factor using penalized maximum likelihood estimation. A novel penalty that they call the nexted Lasso produces an estimator with an adaptive bandwidth for each row of the Cholesky factor. This structure has more flexibility than regular banding, but, unlike regular Lasso applied to the entries of the Cholesky factor, results in a sparse estimator for the inverse of the covariance matrix.

Table 2 shows the ideal, rectangular shape of such data where N units (subjects, stocks, households, financial instruments, etc.) are measured repeatedly on one variable. In most longitudinal studies, the functional trajectories of the involved smooth random processes are not directly observable. Often, the observed data are noisy, sparse and irregularly spaced measurements of these trajectories. In the case that subjects don't share a common set of observation times, the notion of the discrete lag doesn't have a clear definition. In turn, it is not clear then, how one would apply smoothing to each subdiagonal of T since this relies on data observed on a regular grid. Moreover, if one believes that the data used to inform one subdiagonal could inform subdiagonals close to it, failing to smooth in both directions fails to make use of this information. In Chapter 2, we outline a proposed framework for covariance estimation based on the Cholesky decomposition, viewing T as a continuous function in both the lag direction as well as the direction orthogonal to it. Using this approach allows us to also remove any restriction on observation times being regularly spaced and the same across subject. Henceforth, we take Y_i and $\epsilon_i = \left(\epsilon_{i1}, \ldots, \epsilon_{i,m_i}\right)'$ to be continuous processes $Y(t)$, $\epsilon(t)$ observed at discrete measurement times $t_1, \ldots, t_{m_i}$. Using a likelihood-based estimation approach alongside a functional interpretation of the GARPs permits a natural way to regularize the estimator and allow any functional characterizations of the dependency structure to be entirely data driven.

Modeling $\phi_{ij} = \phi(t_i, t_j)$ as a smooth bivariate function, we cast the problem of estimating a covariance matrix as the estimation of a functional varying coefficient model. The existing body of literature surrounding these models is an extensive one; see **?**, **?**, and **?**. This class of models is both flexible and interpretable, making them a pragmatic modeling choice when understanding the underlying data generating mechanism is of as much importance as strong predictive capability.

Table 2: Ideal shape of repeated measurements.

		Occasion					
		1	2	...	t	...	m
Unit	1	y_{11}	y_{12}	...	y_{1t}	...	y_{1m}
	2	y_{21}	y_{22}	...	y_{2t}	...	y_{2m}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	i	y_{i1}	y_{i2}	...	y_{it}	...	y_{im}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	N	y_{N1}	y_{N2}	...	y_{Nt}	...	y_{Nm}

We employ two representations of the GARPs, which we refer to as the *generalized autoregressive coefficient function* within this frame. Chapter 2 presents a reproducing kernel Hilbert space framework for the estimation of both ϕ and σ^2. In Chapter 3, we discuss an alternative representation the varying coefficient function using the penalized B-splines of **?**. We properties of the P-splines that establish their connection to the usual spline penalty on the second derivative and demonstrate how their simple construction allows for extremely flexible regularization.

References

T. W. Anderson, editor. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1984.

TW Anderson. Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, pages 135–141, 1973.

Anestis Antoniadis and Jianqing Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96(455):939–967, 2001.

Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

Peter J Bickel, Elizaveta Levina, et al. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

Graciela Boente and Ricardo Fraiman. Kernel-based functional principal components. *Statistics & probability letters*, 48(4):335–345, 2000.

T Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*, volume 30. CRC Press, 1988.

Colin J Champion. Empirical bayesian estimation of normal variances and covariances. *Journal of multivariate analysis*, 87(1):60–79, 2003.

Tom YM Chiu, Tom Leonard, and Kam-Wah Tsui. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210, 1996.

Rainer Dahlhaus et al. Fitting time series models to nonstationary processes. *The annals of Statistics*, 25(1):1–37, 1997.

Michael J Daniels and Robert E Kass. Nonconjugate bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association*, 94(448): 1254–1263, 1999.

Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

Dipak K Dey, Sujit K Ghosh, and Bani K Mallick. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000.

Francis Ysidro Edgeworth. Xxii. correlated averages. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 34(207):190–204, 1892.

KR Gabriel. Ante-dependence analysis of an ordered set of variables. *The Annals of Mathematical Statistics*, pages 201–212, 1962.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

LR Haff. Empirical bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, pages 586–597, 1980.

Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, pages 85–98, 2006.

Jianhua Z Huang, Linxu Liu, and Naiping Liu. Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics*, 16(1):189–209, 2007.

Robert I Jennrich and Mark D Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, pages 805–820, 1986.

Genshiro Kitagawa and Will Gersch. A smoothness priors time-varying ar coefficient modeling of nonstationary covariance time series. *IEEE Transactions on Automatic Control*, 30(1):48–56, 1985.

Judy L Klein. *Statistical visions in time: a history of time series analysis, 1662-1938*. Cambridge University Press, 1997.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Elizaveta Levina, Adam Rothman, and Ji Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263, 2008.

Shang P Lin. A monte carlo comparison of four estimators for a covariance matrix. *Multivariate Analysis*, 6:411–429, 1985.

P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman and Hall, London, 2nd edition, 1989.

Charles E McCulloch and John M Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.

Nocolai Meinhausen and Peter Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

Jianxin Pan and Gilbert . Regression models for covariance structures in longitudinal studies. *Statistical Modelling*, 6(1):43–57, 2006.

Jianxin Pan and Gilbert Mackenzie. On modelling mean-covariance structures in longitudinal studies. *Biometrika*, 90(1):239–244, 2003.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 2012.

José C Pinheiro and Douglas M Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296, 1996.

M Pourahmadi and MJ Daniels. Dynamic conditionally linear mixed models for longitudinal data. *Biometrics*, 58(1):225–231, 2002.

Mohsen Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

Mohsen Pourahmadi. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, pages 425–435, 2000.

James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2007.

John A Rice and Bernard W Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 233–243, 1991.

Adam J Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

Charles Stein. Estimation of a covariance matrix, rietz lecture. In *39th Annual Meeting IMS, Atlanta, GA, 1975*, 1975.

Arunas Petras Verbyla. Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 493–508, 1993.

Wei Biao Wu and Mohsen Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.

Wei Biao Wu and Mohsen Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pages 1755–1768, 2009.

Ruoyong Yang and James O Berger. Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, pages 1195–1211, 1994.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

Scott L Zeger and Peter J Diggle. Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, pages 689–699, 1994.

Dale L Zimmerman and Vicente Núñez-Antón. Structured antedependence models for longitudinal data. In *Modelling longitudinal and spatially correlated data*, pages 63–76. Springer, 1997.