Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

February 19, 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. **?**, **?**, and **?** 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. **?** follow the hueristic argument presented in **?** 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

The following chapter presents a review of approaches to parsimoniously modeling covariance matrices. We focus on methods based on regularization and generalized linear modeling perspectives, which are concerned with constraining the parameter space of covariance matrices so as to

reduce its dimension, making estimation a non-Sisyphian endeavor. The generalized linear model (GLM) framework **?** 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 **?**, **?**, **?**. An analogous framework for modeling covariance matrices facilitates further developments in covariance estimation from the Bayesian, nonparametric and other paradigms.

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, $N >> M$ (**?**.) 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 **?** Lin and Perlman, 1985; Yang and Berger, 1994; Daniels and Kass, 1999; Champion, 2003; Wong, Carter and

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 $v \in \Re^M$. Together, these hurdles make parsimoniously modeling covariance matrices a great challenge in Statistics and its areas of application.

2.1 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 $p^* = \min\left(p, t - 1\right)$, 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

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

or covariance matrix of the form

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \ldots & m_{p-1} \\ m_1 & m_0 & m_1 & \ldots & m_{p-2} \\ m_2 & m_1 & m_0 & \ldots & m_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p-1} & m_{p-2} & m_{p-3} & \ldots & m_0 \end{bmatrix}, \tag{7}$$

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 $O\left(n\right)$ rather than $O\left(1\right)$.

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:

$$y_t = \mu_t + \sum_{j=1}^{t} \epsilon_j, \quad t = 1, \ldots, M, \tag{8}$$

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:

$$\sigma_{ij} = \begin{cases} \sqrt{i/j} & i \neq j \\ \\ j\sigma_\epsilon^2 & i = j, \end{cases} \tag{9}$$

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

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \quad i = 1, \ldots, M, \tag{10}$$

where the Z_i are specified matrices, the γ_i are vectors of random coefficients distributed independently as $N\left(0, G_i\right)$, 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

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

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 **?** 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.

- compound symmetry
- stationary autoregressive
- moving average models and banded structures - will be mentioned in linear model section
- heterogeneous extensions of the CS, AR, and MA models
- ARIMA models
- random coefficient models
- antedependence models - will be reviewed in linear model section

2.2 Shrinkage estimators based on the sample covariance matrix

Alternately, several have proposed applying nonparametric methods directly to elements of the sample covariance matrix or a function of the sample covariance matrix. Diggle and Verbyla (1998) introduced a nonparametric estimator obtained by kernel smoothing the sample variogram and squared residuals. Yao, Mueller, and Wang 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. [REVIEW 2009 WU AND POURAHMADI METHOD: banding the sample covariance matrix. Under the assumption of short range dependency, they show that their estimator converges to the true covariance matrix for a broad class of nonlinear processes.] The estimates yielded by these approaches, however, are not guaranteed to be positive definite.

2.2.1 Shrinking the spectrum and the correlation matrix

2.2.2 Ledoit-Wolf shrinkage estimator

2.2.3 Penalized likelihood approach

2.2.4 Elementwise shrinkage

Another way to induce parsimony is by applying a shrinkage operator elementwise to the sample covariance matrix.

2.2.5 tapering/banding estimators

2.2.6 thresholding the sample covariance matrix

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

$$s_\lambda(z) = \arg\min_\sigma \left[\frac{1}{2}(\sigma - z)^2 + J_\lambda(\sigma)\right] \tag{12}$$

For detailed discussion of the connection between penalty functions and the resulting thresholding rules, see **?**. Soft thresholding results from minimizing 12 using the lasso penalty, $J_\lambda = \lambda|\sigma|$, which corresponds to thresholding rule

$$s_\lambda(z) = \text{sign}(\sigma)(\sigma - \lambda)_+ . \tag{13}$$

2.2.7 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 **??**,

$$||\hat{\Sigma}^{\lambda} - \Sigma||^2 = \left(\sum_{i,j} \left(\hat{\sigma}_{ij}^{\lambda} - \sigma_{ij} \right)^2 \right)^{1/2} \tag{14}$$

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

$$\text{CV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(-k)} - \tilde{\Sigma}^{(k)}||_F^2, \tag{15}$$

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 **?**. This result motivates the reverse cross validation criterion, which is defined as follows:

$$\text{rCV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(k)} - \tilde{\Sigma}^{(-k)}||_F^2, \tag{16}$$

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 $\mathcal{D}_k$.

TODO: introduce bootstrap risk estimator methods here - See **?**, section 2.2.1

2.3 (Generalized) linear models for covariance estimation

2.3.1 A review of linear models and generalized linear models for covariance matrices

While systematic and data-based modeling of covariance matrices is hampered by the positive-definiteness constraint and high-dimensionality, similar though simpler obstacles in modeling the

mean vector μ (first moments) of the distribution of a random vector $Y = (y_1, \ldots, y_M)'$ has been handled quite successfully in the framework of regression analysis, leading to the powerful theory of generalized linear models (GLM). The success of GLM in handling variety of continuous and discrete data is mainly due to relying on a link function $g(\cdot)$ and a linear predictor $g(\cdot) = X\beta$ to induce unconstrained parameterization and reduce 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.

The areas of time series analysis (Klein, 1997) and variance components (Searle, Casella and McCulloch, 1992, Chap. 2) are among the oldest in dealing with modeling covariance matrices using covariates implicitly and explicitly, respectively. In a sense, they provide the much needed core methods and ideas. In fact, time series techniques based on spectral and Cholesky decompositions provide suitable tools for handling the awkward positivedefiniteness constraint on a stationary covariance matrix (function). However, unlike modeling the mean vector where a link function acts component-wise on the vector μ, link functions for covariance matrices cannot act componentwise since positive-definiteness is a simultaneous constraint on all entries of a matrix. Not surprisingly, because of the complicated structure of a general covariance matrix, the most successful modeling approaches need to rely on decomposing a covariance matrix into its "variance" and "dependence" components. The idea of regression and its diagnostic techniques work well for the logarithm of the variances, but their analogues need to be developed for the more complicated dependence components. The three major methods for producing such pairs, i.e. the variance-correlation, spectral (eigenvalue) and the Cholesky decompositions of several covariances are reviewed in Section 2. However, the latter being less familiar is described next for a single covariance matrix

2.3.2 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:

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

Soon thereafter, the simple structures of time series and variance components models motivated **?** to define the class of linear covariance models:

$$\Sigma = \sum_{i=1}^{q} \alpha_q U_q \tag{17}$$

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

$$\Sigma = \sum_{i=1}^{M} \sum_{j=1}^{M} \sigma_{ij} U_{ij}, \tag{18}$$

where U_{ij} is an $M \times M$ matrix with a 1 in the (i, j) position, and zeros everywhere else. 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 17 and the covariance parameterization given in 18 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

$$\sigma_{ij} = 0 \quad \text{for } |i - j| > q, \tag{19}$$

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

$$\sigma^{ij} = 0 \quad \text{for } |i - j| > p. \tag{20}$$

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 $AD(p)$ **?** 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 20. Closely connected are the classes of variable order AD models and varying order, varying coefficient autoregressive models **?** in which the coefficients and order of antedependence depend on time.

2.3.3 Log-linear covariance models

The constraint on the α_is in 17 was eliminated with the introduction of log-linear covariance models; see **?**. For a general covariance matrix having spectral decomposition

$$\Sigma = P \Lambda P', \tag{21}$$

its matrix logarithm, denoted $\log \Sigma$, and defined by $log\Sigma = P \log \Lambda P'$ is a symmetric matrix with unconstrained entries taking values on $\Re$. A log-linear model for Σ may be written as

$$\log \Sigma = \sum_{i=1}^{q} \alpha_i U_i, \tag{22}$$

where the U_is are as before in 17 and the α_is are now unconstrained. The α_is, however, now lack statistical interpretation since $\log \Sigma$ is a highly nonlinear operation. But for diagonal Σ, $\log \Sigma = \text{diag}\,(\sigma_{11}, \ldots, \sigma_M M)$, and model 22 reduces to modeling of heterogeneous variances, which has been extensively studied. Detailed presentation is given in **?**, **?**m and in references therein.

2.3.4 GLMs

To satisfy the positive-definiteness constraint, methods have been developed and applied to certain reparameterizations of the covariance structure. Chiu, Leonard, and Tsui modeled the matrix logarithm of the covariance matrix. Early nonparametric work using the spectral decomposition of the covariance matrix included that of Rice and Silverman (1991) which discussed smoothing and smoothing parameter choice for eigenfunction estimation for regularly-spaced data. Staniswalis and Lee (1998) extended kernel-based smoothing of eigenfunctions to functional data observed on irregular grids. However, when the data are sparse in the sense that there are few repeated within-subject measurements and measurement times are quite different from subject-to-subject, approximation of the functional principal component scores defined by the Karhunen-Loeve expansion of the stochastic process by usual integration is unsatisfactory and requires numerical quadrature. Many have explored regression-based approaches using the Spectral decomposition, framing principal components analysis as a least-squares optimization problem. Among many others, Zou, Hastie and Tibshirani (2006) imposed penalties on regression coefficients to induce sparse loadings. [REVIEW THE METHODS OF HUANG, KAUFMAN, YAO HERE]

We adopt the approach based on the Cholesky decomposition. The modified Cholesky decomposition (MCD) has received much attention in the covariance estimation literature, as it ensures positive-definite covariance estimates, and, unlike the spectral decomposition whose parameters follow an orthogonality constraint, the Cholesky decomposition are unconstrained and have an attractive statistical interpretation as particular regression coefficients and variances. The Cholesky decomposition is similar to the spectral decomposition in that ? is diagonalized by a lower triangular matrix T:

$$T\Sigma T' = D,$$

where the nonredundant entries of T are unconstrained and more meaningful statistically than those of the orthogonal matrix of the spectral decomposition. The matrix T is constructed from the regression coefficients when yt is regressed on its predecessors:

$$y_t = \sum_{j=1}^{t-1} \phi_{t,j} y_j + \epsilon_t, \tag{23}$$

where the (t, j) entry of T is ϕ_{tj} , the negatives of the regression coefficients and the (t, t) entry of D is $\sigma_t^2 = var(\epsilon_t)$, the innovation variance. A schematic view of the components of a covariance matrix obtained through successive regressions (Gram-Schmidt orthogonalization procedure) is given in Table 2. Since the ϕ_{ij}s are regression coefficients, it is evident that for any unstructured covariance matrix these and the log innovation variances are unconstrained, in the sequel they are referred to as the generalized autoregressive parameters (GARP) and innovation variances (IV) of Y or ? (Pourahmadi, 1999, 2000). Interestingly, this regression approach reveals the equivalence of modeling a covariance matrix to that of dealing with a sequence of $p-1$ varying-coefficient and varying-order regression models. Consequently, one can bring the entire regression machinery to the service of the unintuitive task of modeling covariance matrices. Stated differently, the framework above is similar to that of using increasing order autoregressive models in approximating the covariance matrix or the spectrum of a stationary time series.

The covariance matrix Σ of a zero-mean random vector $Y = (y_1, \ldots, y_m)'$ has the following unique modified Cholesky decomposition (Newton, 1988)

$$T\Sigma T' = D, \tag{24}$$

where T is a lower triangular matrix with 1?s as its diagonal entries and $D = \text{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ is a diagonal matrix. An attractive feature of this decomposition is that unlike the entries of Σ, the subdiagonal entries of T and the log of the diagonal elements of D, $\log(\sigma_m^2)$, $t = 1, \ldots, m$, are not constrained. This permits one to impose structures on the unconstrained parameters without worrying about the resulting estimator not satisfying the positive-definiteness constraint. Denote estimators of T and D in 27 by $\hat{T}$ and $\hat{D}$, which may be obtained by fitting linear models or some other structural models; then an estimator of Σ given by $\Sigma = \hat{T}^{-T}\hat{D}\hat{T}^{-T}$ is guaranteed to be positive-definite. From this perspective, covariance modeling can be considered an extension of generalized linear models **?**. Factoring Σ as in 24 provides a link function $g(\Sigma) = (T, \log(D))$ where $\log(D) = \text{diag}(\log(\sigma_1^2), \ldots, \log(\sigma_m^2))$. Parametric, nonparametric, or Bayesian models may then be applied to the unconstrained entries of T and $\log(D)$. Whereas other decompositions are permutation-invariant, the interpretation of the regression model induced by the MCD assumes a natural (time) ordering among the variables in Y.

immediately leads to the modified Cholesky decomposition 24. It also can be used to clarify the close relation between the decomposition (2) and the time series ARMA models in that the latter is means to diagonalize a Toeplitz covariance matrix, for details see Pourahmadi (2001, Sec. 4.2.5).

In sharp contrast, the fact that the lower triangular matrix T in the Cholesky decomposition of a covariance matrix Σ is unconstrained makes it ideal for nonparametric estimation. Wu and Pourahmadi (2003) have used local polynomial estimators to smooth the subdiagonals of T. For the moment, denoting such estimators of T and D in (2) by T and D, an estimator of Σ given by $\Sigma = \hat{T}^{-1}D\hat{T}^{-1'}$ is guaranteed to be positive-definite. Although one could smooth rows and columns of T, the idea of smoothing along its subdiagonals is motivated by the similarity of the regressions in (3) to the varying-coefficients autoregressions (Kitagawa and Gersch, 1985, 1996; Dahlhaus, 1997): Xm

Xm j=0

$$f_{j,p}\left(t/p\right) y_{t_j} = \sigma_p\left(t/p\right)\epsilon_t, \quad t = 0, 1, 2, \ldots, M, \tag{25}$$

where $f_{0,p}\left(\right) = 1$, $f_{j,p}\left(\right)$, 1 ? j ? m, and ?p() are continuous functions on $[0, 1]$ and ?t 30 is a sequence of independent random variables each with mean zero and variance one. This analogy and comparison with the matrix T for stationary autoregressions having constant entries along subdiagonals suggest taking the subdiagonals of T to be realizations of some smooth univariate functions:

$$\phi_{t,t-j} = f_{j,p}\left(t/p\right), \quad \sigma_t = \sigma_p\left(t/p\right).$$

The details of smoothing and selection of the order m of the autoregression and a simulation study comparing performance of the sample covariance matrix to smoothed estimators are given in Wu and Pourahmadi (2003). Due to the closer connection between entries of T and the family of regression (3), it is conceivable that some of the entries of T could be zero or close to it. Smith and Kohn (2002) have used a prior 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, Liu and Pourahmadi (2004) using penalized likelihood with L_1-penalty to estimate T for Gaussian data. A commonly utilized approach in previous work is to model $\phi_{ijk} = z_{ijk}^T\gamma$ where z_{ijk} is a vector of powers of time differences and γ is a vector of unknown "dependence" parameters to be estimated from the data. **?**, **?**, **?**, and **?** define

$$z_{ijk}^T = \left(1, t_{ij} - t_{ik}, \left(t_{ij} - t_{ik}\right)^2, \ldots, \left(t_{ij} - t_{ik}\right)^{q-1}\right) \tag{26}$$

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 $\sigma^2\left(t\right)$ 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. Modeling ϕ in such a way is equivalent to specifying a Toeplitz structure for Σ. A $p \times p$ Toeplitz matrix M is a matrix with elements m_{ij} such that $m_{ij} = m_{|i-j|}$ i.e. a matrix of the form

The estimated covariance matrix may be considerably biased when the specified parametric model is far from the truth. To avoid model misspecification that potentially accompanies parametric analysis, many have alternatively proposed nonparametric and semiparametric techniques approaches to estimation. While these estimators can be very flexible and thus exhibit low bias, this advantage can be offset with high variance. To balance the tradeoff between bias and variance, shrinkage or regularization may be applied to estimates to improve stability of estimators. **?** proposed nonparametric estimation of the covariance matrix of longitudinal data by smoothing raw sample variogram ordinates and squared residuals. [DISCUSS THE NONPARAMETRIC SMOOTHER OF HANS GEORG MULLER HERE] However, neither of these methods ensure that the resulting estimates are positive-definite.

Several others have proposed methods for covariance estimation within the same paradigm of a smooth, continuous function underlying a discretized covariance matrix associated with the observed data. **?** employ the Cholesky decomposition to guarantee positive-definiteness and imposed structure on the elements of the Cholesky decomposition and heuristically argue that $\phi_{t,t-l}$ should be monotonically decreasing in l. That is, the effect of y_{t-l} on y_t through the autoregressive parameterization should decrease as the distance in time between the two measurements increases. In similar spirit, others including **?** and **?** enforce such structure by setting $\phi_{t,t-l}$ equal to zero for l large enough, or equivalently, setting all subdiagonals of T to zero beyond the K^{th} off-diagonal. The tuning parameter K is chosen using a model selection criterion such as Akaike information criterion, Bayesian information criterion, or cross validation or a variant thereof. In terms of the autoregressive model corresponding to the Cholesky decomposition, this form of regularization, known as "banding" the Cholesky factor T, is equivalent to regressing y_t on only its K immediate predecessors, setting $\phi_{tj} = 0$ for $t - j > K$.

From this perspective, it is apparent that the presentation of covariance estimation as a least squares regression problem suggests that the familiar ideas of model regularization for least-squares regression can be used for estimating covariances. Wu and Pourahmadi **?** proposed a two-step estimation procedure using nonparametric smoothing for regularized estimation of large covariance matrices. 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. The use of the Cholesky parameterization guarantees that their estimate is guaranteed to be positive-definite, however, their procedure is not capable of handling missing data. **?**

however, their two-step method did not utilize the information that many of the subdiagonals of T are essentially zeros at the first step. Inefficient estimation may result because of ignoring regularization structure in constructing the raw estimator.

Several have applied these approaches to covariance estimation; **?** jointly model the mean and covariance matrix of longitudinal data using basis function expansions. They employ the Cholesky decomposition of the covariance matrix and treat the subdiagonals of T as smooth functions, approximated by B-splines. Estimation is carried out by maximizing the 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.

Alternatively, one can view T as a bivariate function,

Several others have considered this approach to covariance estimation; **?** assume a stationary process, restricting covariance estimates to a specific class of functions. They as well as Huang, Liu, and Liu **?** follow the hueristic argument presented by **?** that $\phi_{t,t-l}$ is monotone decreasing in l and set off-diagonal elements of either the covariance matrix or the Cholesky factor corresponding to large lags to zero. As in **?**, **?**, and **?**, we treat covariance estimation as a function estimation problem where the covariance matrix is viewed as the evaluation of a smooth function at particular design points.

including **?** and **?** have proposed nonparametric estimators of a specific covariance matrix (or its inverse) rather than the parameters of a covariance function.

? do not utilize the Cholesky parameterization, and their estimates are not guaranteed to be positive definite. We combine the advantages of bivariate smoothing as in **?** with the added utility of the Cholesky parameterization in **?**; in doing so, we present a flexible and coherent approach to covariance estimation, while simultaneously we ensuring positive definiteness of estimates.Rather than shrinking element of the Cholesky factor to zero after a particular value of l, we choose to softly enforce monotonicity in l by using a hinge penalty as in the work of **?**.

3 The Cholesky Decomposition and the MLE for Σ

Let $Y = (y_1, y_2, \ldots, y_m)'$ denote a mean zero random vector with variance-covariance matrix Σ, which we can think of as the time-ordered measurements on one subject in a longitudinal study. To present a comprehensive overview our estimation procedure, we begin with the representation of the covariance matrix, Σ, in terms of its Cholesky decomposition. Decomposing Σ in such a way allows for both an unconstrained parameterization and statistically meaningful interpretation of covariance parameters. For any positive definite matrix Σ, there exists a unique lower triangular matrix T with diagonal entries equal to 1 which diagonalizes Σ:

$$T\Sigma T^T = D \tag{27}$$

The convenient statistical interpretation of the parameters of the covariance matrix then comes if we consider, for $t = 2, \ldots, m$, regressing y_t on its predecessors $y_1, \ldots, y_{t-1}$, letting

$$y_i = \sum_{j=1}^{i-1} \phi_{ij} y_j + \sigma_i \epsilon_i, \tag{28}$$

where var $(\epsilon_i) = \sigma_i^2$. If we take the i-j^{th} element T to be $-\phi_{ij}$ for $j < i$, and take the i^{th} diagonal entry of D to be var $(\epsilon_i) = \sigma_i^2$, a vectorized expression for Model 28 is given by

$$\boldsymbol{\epsilon} = TY. \tag{29}$$

and taking covariances on both sides of (**??**), we see that T and D satisfy 27. Immediately, we have that $\Sigma^{-1} = T'D^{-1}T$. The regression coefficients $\{\phi_{ij}\}$ are referred to as the *generalized autoregressive parameters* (GARPs), and the $\{\sigma_{ij}\}$ are referred to as the *innovation variances* (IVs.) Assuming that Y follows a multivariate normal distribution, the loglikelihood function $\ell(Y, \Sigma)$ satisfies

$$-2\ell(Y, \Sigma) = \log|\Sigma| + Y'\Sigma Y \tag{30}$$

From 27, we have that

$$|\Sigma| = |D| = \prod_{i=1}^{m} \sigma_i^2$$

Table 1: 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}

and

$$\Sigma^{-1} = T'D^{-1}T.$$

Thus, 30 can be written in terms of the prediction errors and their variances of the non-redundant entries of (T, D):

$$\begin{aligned} -2\ell\left(Y, \Sigma\right) &= \log|D| + Y'T'D^{-1}TY \\ &= \sum_{i=1}^{m} \log \sigma_i^2 + \sum_{i=1}^{m} \frac{\epsilon_i^2}{\sigma_i^2}, \end{aligned} \tag{31}$$

where $\epsilon_1 = y_1$ and $\epsilon_i = y_i - \sum_{j=1}^{i-1} \phi_{ij} y_j$. Maximum likelihood estimation or any of its penalized variants may then be employed to obtain estimates of T and D.

Unlike many of those before who have used the Cholesky decomposition as a means of modeling Σ, we allow observed time points to be individual-specific and not necessarily regularly spaced. Let $Y_1, \ldots, Y_N$ denote a random sample of mean zero vectors of longitudinal measurements taken on N subjects having common covariance structure Σ. We allow subject i to have observation vector $y_i = \left(y_{i1}, \ldots, y_{i,m_i}\right)'$ with corresponding vector of observation times $\left(t_{i1}, \ldots, t_{i,m_i}\right)'$. Accommodating the subject-specific sample sizes and measurement times requires merely adding a subscript, and Model 28 becomes

$$y_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} y_{ik} + \sigma_{ij} \epsilon_{ij}, \tag{32}$$

where ϕ_{ijk} is the autoregressive coefficient corresponding to the pair of measurements observed at time t_{ij} and t_{ik}. A vectorized representation of Model 32 can be obtained as before by adding the necessary parameters to T and D.

Table 2: 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

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{m1} & -\phi_{m2} & \cdots & -\phi_{m,m-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \ddots \\ y_m \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \ddots \\ \epsilon_m \end{bmatrix} \tag{33}$$

TODO: here, conclude with our view of the GARPs and IVs as functions, but allude to how this differs from the stationary approach of Pourahmadi and successive regressions by relaxing the stationarity assumption and viewing T as a continuous bivariate function. Move all the remaining details after this to Chapter 2.

In many experiments, for example, most longitudinal studies, the functional trajectories of the involved smooth random processes are not directly observable. In these cases, the observed data are noisy, sparse and irregularly spaced measurements of these trajectories. Removing the restriction that subjects having common covariance structure also share common, equally-spaced observation times encourages the interpretation of vectors Y_i and $\epsilon_i = \left(\epsilon_{i1}, \ldots, \epsilon_{i,m_i}\right)'$ as continuous time 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 while making minimal assumption about the form of the dependency structure itself. Other approaches have proposed using functional representation of the ϕ_{ij} and σ_j, but use multiple one-dimensional functions $\phi_l(t)$ to approximate the subdiagonals of T. See Huang et al, Pourahmadi? When IVs are modeled as constant in T, this equates to modeling y_t as a stationary process.

We prefer to take a full data-driven approach to the parameterization of the regression parameters in 32, modeling $\phi_{ij} = \phi(t_i, t_j)$ as a smooth bivariate function, casting covariance estimation as estimation of a functional varying coefficient model, which have already been extensively studied. While flexible enough to closely approximate nearly any functional relationship between a set

of predictors and a response, they can also be incredibly interpretable, making them a pragmatic modeling choice when understanding the underlying data generating mechanism is of as much importance as strong predictive capability. By leveraging a likelihood-based estimation procedure, we can naturally relax the restriction that the data on individual curves $Y_1, \ldots, Y_N$ be observed on a grid of time points which are equally spaced and common to all subjects. Simultaneously, by employing a functional view of the autoregressive coefficients $\phi_{t,j}$, we can seamlessly incorporate regularization framework that accompanies the usual nonparametric function estimation problem. We make only mild assumptions about the GARPs in terms of the direction perpendicular to the subdiagonals of T, but instead utilize penalties that shrink the coefficient function toward forms in concordance with previously proposed stationary parameterizations. In Chapter 2, we present a two functional representations of Model 32. One such representation elicits a reproducing kernel Hilbert space framework for estimation of the varying coefficient function, blah blah blah