Regularized Nonparametric Covariance Estimation via Cholesky Decompostion

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Tayler A. Blake, C.S., D.S.

Graduate Program in Department of Statistics

The Ohio State University

2016

Dissertation Committee:

Dr. Yoonkyung Lee, Advisor

Dr. Kate Calder

Sebastian Kurtek

Abstract

In the era of "Big Data", the ability to collect and store massive amounts of information has made access to functional and high dimensional data much more prevalent. With this prevalence comes a strong need for methods of estimating large covariance matrices. Sample covariance matrices become extremely unstable estimates of covariance structure as dimension increases, and the desire to impose a positive-definite constraint on estimates further adds further complexity to the problem. 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 un-equally 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. In our first approach, we assume that the solution to the optimization problem belongs to the reproducing kernel Hilbert space spanned by the direct sum of the subspace spanned by the first m Bernoulli polynomials and the corresponding orthogonal penalized subspace.

Later, we consider an alternative construction of the function space using a truncated power basis, reflected about the design points. This formulation allows for intuitive specification of the penalty functional to achieve shrinkage toward commonly used models, such as those assuming stationarity, short term dependence, or decaying dependence as the difference between time points increases. We present numerical results and data analysis to illustrate the utility of the proposed method.

This is dedicated to the one I love . . . la la la . . .

Acknowledgments

I thank everyone who has ever had a cow. . . .

In reality, this is the only page of the dissertation which the author has full control of. You can write anything you want here, and no one can tell you it's wrong (except if the margins don't line up!!!!).

Vita

January 0, 1800	Born - Cowtown, USA
1900	B.S. Cow Science
1950	M.S. Cow-Dairy Science
1985-present	Graduate Teaching Associate, Holstein University.

Publications

Research Publications

B. Simpson "Milking a Cow". *Journal of Dairy Science*, 00(2):277–287, Feb. 1900.

Fields of Study

Major Field: Department of Statistics

Table of Contents

Page

List of Tables

Table **Page**

List of Figures

Figure **Page**

Chapter 1: Introduction

Nearly all statistical procedures in classical multivariate data analysis, time series analysis, spatial statistics and, more recently, the growing field of statistical learning require an estimate of the covariance matrix or its inverse. Specifically, 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 rely heavily on a covariance estimate, and such an estimate plays a critical role in the performance of the technique. Covariance estimation is an open problem and has been explored extensively in previously work; these hurdles are generally recognized for the difficulty in covariance estimation: the instability of sample covariance matrices in high dimensions, irregularly or sparsely sampled data, and enforcing that covariance estimates are positive definite.

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 and biomedicine, public health, biology, biomechanics and environmental science with specific applications including fMRI, spectroscopic imaging, genetics and gene microarrays among many others, presenting a need for effective covariance estimation in the challenging situation when 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 classical functional data or times series data, where each individual corresponds to a curved sampled densely at a fine grid of longitudinal times, where it is typical that the number of observed time points on any individual is larger than the number of individuals. Alternatively, we may consider sparse longitudinal data where measurement times may be almost completely unique 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. 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.

To overcome the hurdle of enforcing covariance estimates to be positive definite, several have considered several 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. The autoregressive coefficients are often referred to in the literature as the *generalized autoregressive parameters*, or *GARPs*, and the error variances are often called the *innovation variances*, or *IVs*.

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.

This differs from many previous works including that of **?**] and **?**] in that they are concern themselves with estimating a specific covariance matrix rather than the parameters of a covariance function. **?**], **?**], and **?**] have viewed the covariance matrix as the evaluation of a smooth function at particular design points. **?**] do not utilize the Cholesky parameterization, and their estimates are not guaranteed to be positive definite. **?**] 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. 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 **?**].

The rest of the paper is organized as follows: Section 2 summarizes the general penalized estimation approach in general, and introduces the transformed coordinates and the penalties for

stationarity and non-monotonicity. Section 3 presents a detailed discussion of optimization and computational issues. Section 4 presents a simulation study and a real example to examine the performance of our methods as well as others. Section 5 concludes with discussion and future work.

Chapter 2: The Cholesky Decomposition: an unconstrained parameterization

For any positive-definite matrix covariance matrix, Σ, we have the following unique decomposition:

$$T'\Sigma T = D \tag{2.1}$$

where T is a lower triangular matrix with diagonal entries all equal to 1, and D is a diagonal matrix with positive diagonal entries. The lower triangular entries of T are unconstrained, and when the data have a natural ordering (as in time or space), they have a meaningful statistical interpretation.

Blah blah blah

Blah Blah Blah

2.1 An autoregressive model for elements of the covariance

Pourahmadi (1999) proposed modeling

$$y_t = \sum_{j=1}^{t-1} \phi_{tj} y_j + \sigma_t \epsilon_t \quad t = 1, \ldots, p \tag{2.2}$$

where $\hat{y}_t$ is the least squares predictor of y_t based on its predecessors, and $\{\epsilon_t = y_t - \hat{y}_t\}$ are uncorrelated with $Cov\,(\epsilon) = D = diag\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\right)$, $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_p)^T$. Then we can write

$$\epsilon = TY \qquad T = \begin{cases} -\phi_{ij} & i > j \\ 1 & i = j \\ 0 & i < j \end{cases}$$

and it follows that

$$Cov\left(\epsilon\right) = TCov\left(Y\right)T^T = T\Sigma T^T = D$$

We refer to the $\{\phi_{tj}\}$ and the $\{\sigma_t^2\}$ as the generalized autoregressive parameters (GARPs) and the generalized innovation variances (GIVs), respectively.

2.2 Covariance estimation as bivariate function estimation

2.2.1 Likelihood specification for ϕ and σ^2

Blah

Blah

Blah

2.2.2 Expressing the GARPs and IVs as functions

Rather than a vector of longitudinal data points, view $Y = \left(y_{t_1}, \ldots, y_{t_n}\right)^T$ and $\epsilon = \left(\epsilon_{t_1}, \ldots, \epsilon_{t_n}\right)^T$, $t_1 \leq \cdots \leq t_n$, as a discretizations of the stochastic processes $y\left(t\right)$ and $e\left(s\right) \sim \mathcal{WN}\left(0, 1\right)$. We view ϕ_{ij} and σ_i^2 as values of the smooth function $\phi\left(s, t\right)$, $0 \leq s < t \leq 1$ and $S\left(t\right) = \log\sigma^2\left(t\right)$, $0 \leq t \leq 1$ evaluated at the design points. Consequently, we view the entries of Σ as values of a smooth covariance function, $\gamma\left(s, t\right)$, evaluated at the distinct pairs of design points. We define y as a stochastic process as follows:

$$y\left(s\right) = \int_0^s \phi\left(s, t\right) y\left(t\right) dt + \sigma\left(s\right) e\left(s\right) \tag{2.3}$$

Re-expressing (2.2) in terms of $\phi\left(s, t\right)$ and $\sigma^2\left(t\right)$, we have

$$\hat{y}(t_i) = \sum_{j=1}^{i-1} \phi(t_i, t_j)\, y(t_j) + \sigma(t_i)\, \epsilon(t_i) \quad i = 1, \ldots, n \tag{2.4}$$

We first turn our focus to the estimation of $\phi(s,t)$ and assume that $\sigma^2(t)$ is known, and WLOG, let $\sigma^2(t) = 1$ for all $t \in [0,1]$. It is well known that the sample covariance matrix is unstable when the dimensionality of the data is high, so the attraction of regularization is obvious.

2.2.3 Reparameterizing the Generalized Autoregressive Coefficient Function

Rather than imposing structure on the unconstrained values of $\phi(s,t)$, we instead consider a rotation of the axes of the input space and re-express ϕ in terms of the transformed coordinates $\begin{bmatrix} l \\ m \end{bmatrix} = \begin{bmatrix} s-t \\ \frac{(s+t)}{2} \end{bmatrix}$

$$\begin{aligned} \phi(s,t) &= \phi^*\left(s-t, \frac{s+t}{2}\right) \\ &= \phi^*(l,m) \end{aligned}$$

Imposing structure on ϕ^* naturally leads to null models presented in earlier work on covariance estimation with appropriate choice of penalty in the smoothing spline problem formulation.

Several approaches to estimating the covariance parameters have been suggested, the simplest of which is the usual maximum likelihood estimation assuming a Normal distribution for ϵ. Garcia, Kohli, and Pourahmadi (2001) make the heuristic argument that, for fixed t, $\phi_{t,t-l}$ should be small for large values of l, assuming that the linear relationship between y_t and y_{t-l} diminishes as lag increases. They enforce $\phi_{t,t-l}$ to be monotonically decreasing in l. To model ϕ, they examine empirical regressograms: plots of $\hat{\phi}_{t,t-j}$ against lag, j and use these to choose a parametric form

for $\hat{\phi}(t, t-j)$. Similarly, they produce innovariograms, plots of the log estimated innovation variances $\hat{\sigma}_t^2$ versus t, and select an appropriate functional form. They choose to model the innovation variance function as polynomial function of t, so that model selection is equivalent to choosing the appropriate degree of polynomial, using ordinary least squares to fit model coefficients. Note that this approach

Chen et. al. presented a semiparametric approach for simultaneously estimating a mean function and covariance structure, modeling $\phi(s, t)$ as a polynomial in $s - t$. By specifying the generalized autoregressive coefficient function to be a function of lag alone, they make the implicit assumption that the underlying stochastic process is stationary. This dramatically aids in dimension reduction, as the number of distinct entries in the covariance matrix decreases from $p(p-1)/2$ to p as a result.

Huang et. al (2006) estimate the elements of T via normal maximum likelihood with a Lasso penalty, which introduces sparsity in T with zeros but in arbitrarily placed entries. Using the Cholesky decomposition to ensure positive-definiteness, Wu and Pourahmadi (2003) presented a $k - diagonal$ estimator which bands the inverse of the covariance matrix by smoothing down the first k diagonals of T and setting the rest to zero, choosing the number of nonzero diagonals by AIC using a normal likelihood. Huang et. al. use spline functions to smooth $\sigma_t^2\}_{t=1}^{p}$ and $\{\phi_{t,t-j}\}_{j=1}^{t-1}$, the sub-diagonals of T which hold the lag-J regression coefficients and are closely related to time-varying autoregressive models. They use penalized maximum likelihood to estimate the spline coefficients, treating the $p-1$ sub-diagonals as $p-1$ separate smoothing spline problems. While computationally does not make use of the potential information about the dependence structure lying in the direction orthogonal to the diagonal; we propose using bivariate smoothing to utilize information in both directions.

2.2.4 The Multicenter AIDS Cohort Study and a problem notation illustration

CD4 cells counts are a commonly used marker in the progression of AIDS, as they are an integral assessment of immune system status. Data from the Multicenter AIDS Cohort Study provides data collected on 283 homosexual males who were infected between 1984 and 1991. Repeated measurements included CD4 cell percentages (CD4 cell counts divided by the total number of lymphocytes, a certain type of white blood cells.) All subjects were scheduled to have measurements taken at semi-annual visits, but there are different numbers of observations per individual and different observation times t_{ij} for each subject due to many subjects missing appointments and the fact that infections arose at different times within the study period. Denote the vectors of measurements on individual i by $Y_i = (y(t_{i1}), \ldots, y(t_{in_i}))^T$ with corresponding measurement times $T_i = (t_{i1}, \ldots, t_{in_i})^T$, $0 \leq t_{ij} < t_{ik} \leq 1$ for $j < k$. A total of $N_Y = \sum_{i=1}^{N} n_i = 2376$ measurements were taken on $N = 283$ subjects, with the number of within-subject measurements, $\{n_i\}$, ranging from 1 to 14. Denote the set of unique observations times by $\mathcal{T} = \bigcup_{i=1}^{N} \bigcup_{j=1}^{n_i} \{t_{ij}\}$ and the number of unique observation times by $|\mathcal{T}| = N_T$. Relabel the elements so that $\mathcal{T} = \{t_1, \ldots, t_{N_T}\}$ and define

$$Y = (Y(t_1), \ldots, Y(t_{N_T}))' \qquad Cov(Y) = \Sigma_{N_T \times N_T}$$

where the ijth entry of Σ is defined to be $\Sigma_{ij} = Cov(y(t_i), y(t_j)) = \gamma(t_i, t_j)$. In other words, we view the entries of Σ as the value of a smooth bivariate function γ at $\{(t_i, t_j)\}$. Writing (2.4) in terms of the transformed coordinates, (l_{ij}, m_{ij}), the autoregressive model arising from the Cholesky decomposition of Σ becomes

$$\begin{aligned}\hat{y}(t_i) &= \sum_{j=1}^{i-1} \phi(t_i, t_j)\, y(t_j) + \epsilon(t_i) \\ &= \sum_{j=1}^{i-1} \phi^*\left(t_i - t_j, \frac{1}{2}(t_i - t_j)\right) y(t_j) + \epsilon(t_i) \\ &= \sum_{j=1}^{i-1} \phi^*(l_{ij}, m_{ij})\, y(t_j) + \epsilon(t_i) \qquad (2.5)\end{aligned}$$

where $\phi^* : (0,1) \times (0,1) \to \mathrm{R}^+$ is a smooth bivariate function, and the transformed design points have been scaled so that $l_{ij}, m_{ij} \in (0,1)$.

2.3 A Structured Family of Nonparametric Models for the Covariance Structure

Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ be the reproducing kernel Hilbert space (r.k.h.s) corresponding to the tensor product of the first-order and second-order Sobolev spaces:

$$\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m, \quad \mathcal{H}_l = W_2(0,1), \quad \mathcal{H}_m = W_1(0,1) \text{ where}$$

$$W_m(0,1) \equiv \{f : \quad f', \ldots, f^{(m-1)} \text{absolutely continuous}, \int_0^1 \left(f^{(m)}\right)^2 dt < \infty\}$$

We seek $\phi^*(\cdot,\cdot) \in \mathcal{H}$ which minimizes

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=2}^{n_i} \sigma_{ij}^{-2}\left(y(t_{ij}) - \sum_{k=1}^{n_i-1} \phi^*\left(l_{jk}^i, m_{jk}^i\right) y(t_{ik})\right)^2 + \lambda J(\phi^*) \qquad (2.6)$$

where $P_1\phi^*$ is the projection of ϕ^* onto $\mathcal{H}_1$, $J(\phi^*) = ||P_1\phi^*||^2$. Define the differential operator $M_\nu f = \int_0^1 f^{(m)}(x)\, dx\ , \quad \nu = 1, \ldots, m$ and endow $W_m(0,1)$ with inner product

$$\langle f, g\rangle = \langle f, g\rangle_0 + \langle f, g\rangle_1 = \sum_{\nu=0}^{m-1} M_\nu f M_\nu g + \int_0^1 f^{(m)}(x)\, g^{(m)}(x)\, dx \qquad (2.7)$$

which induces norm

$$||f||^2 = \langle f, f\rangle = \langle f, f\rangle_0 + \langle f, f\rangle_1 = ||P_0 f||^2 + ||P_1 f||^2$$

Let $k_j(x) = B_j(x)/j!$ for $x \in [0,1]$, where $B_j(x)$ is the j^{th} Bernoulli polynomial which can be defined according to the recursive relationship:

$$B_0(x) = 1, \qquad \frac{d}{dx} B_r(x) = r B_{r-1}(x)$$

Noting that $M_\nu B_r = \delta_{\nu - r}$, W_m can be written as a direct sum of the m orthogonal subspaces: $\{k_r\}_{r=0}^{m-1}$ and W_m^1. Here, $\{k_r\}$ is the subspace spanned by k_r and W_m^1 is the space orthogonal to $W_m^0 \equiv \{1\} \oplus \{k_1\} \oplus \cdots \oplus \{k_{m-1}\}$ which satisfies

$$W_m^1 = \{f : M_\nu f = 0, \;\; \nu = 0, 1, \ldots, m-1\}$$

Writing $\mathcal{H}$ as the tensor product of the two decomposed Sobolev spaces, we have

$$\begin{aligned}
\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m &= W_2 \otimes W_1 \\
&= \left[W_2^0 \oplus W_2^1\right] \otimes \left[W_1^0 \oplus W_1^1\right] \\
&= \left[[\{1\} \oplus \{k_1\}] \oplus W_2^1\right] \otimes \left[\{1\} \oplus W_1^1\right] \\
&= [\{1\} \oplus \{k_1\}] \oplus W_2^1 \oplus W_1^1 \oplus \left[\{k_1\} \otimes W_1^1\right] \oplus \left[W_2^1 \otimes W_1^1\right] \\
&\equiv \left[\mathcal{H}_{\mu^*} \oplus \mathcal{H}_l^0\right] \oplus \left[\mathcal{H}_l^1 \oplus \mathcal{H}_m^1 \oplus \mathcal{H}_{lm}^{01} \oplus \mathcal{H}_{lm}^{11}\right] \\
&= \mathcal{H}_0 \oplus \mathcal{H}_1
\end{aligned} \tag{2.8}$$

where the functional components corresponding to $\mathcal{H}_\mu$, $\mathcal{H}_l^0$, $\mathcal{H}_l^1$, $\mathcal{H}_m^1$, and $[\mathcal{H}_{lm}^{01} \oplus \mathcal{H}_{lm}^{11}]$ are the overall mean, the nonparametric main effect of l, the parametric main effect of l, the parametric

main effect of m, the nonparametric-parametric interaction, and the parametric-parametric interaction (between l and m). Given this decomposition of the function space, any $\phi^* \in \mathcal{H}$ may be written as a sum of components from each of the

$$\phi^*(l,m) = \mu^* + \phi_l^*(l) + \phi_m^*(m) + \phi_{lm}^*(l,m) \tag{2.9}$$

where $\int_0^1 \phi_l^*(l)\,dl = \int_0^1 \phi_m^*(m)\,dm = 0$, $\int_0^1 \phi_{lm}^*(l,m)\,dl = \int_0^1 \phi_{lm}^*(l,m)\,dm = 0$. The reproducing kernel (r.k.) for $\{k_r\}$ is $k_r(x)\,k_r(x')$. It can be verified that the r.k. for W_m^1 (Craven and Wahba 1979) is given by $R^1(x,x') = k_m(x)\,k_m(x') + (-1)^{m-1}\,k_{2m}([x-x'])$ where $[\alpha]$ is the fractional part of α. The r.k. for W_m is given by

$$\begin{aligned} R(x,x') &= R^0(x,x') + R^1(x,x') \\ &= \left[\sum_{\nu=1}^{m-1} k_\nu(x)\,k_\nu(x')\right] + \left[k_m(x)\,k_m(x') + (-1)^{m-1}\,k_{2m}([x-x'])\right] \end{aligned}$$

Using the fact that the r.k. for a tensor product space is the product of the corresponding reproducing kernels, the r.k. for $\mathcal{H}$ is given by

$$\begin{aligned} R((l,m),(l',m')) &= R_l(l,l') \times R_m(m,m') \\ &= \left[R_l^0(l,l') + R_l^1(l,l')\right] \times \left[R_m^0(l,l') + R_m^1(l,l')\right] \\ &= R_l^0(l,l')\,R_m^0(m,m') + R_l^0(l,l')\,R_m^1(m,m') \\ &\qquad + R_l^1(l,l')\,R_m^0(m,m') + R_l^1(l,l')\,R_m^1(m,m') \\ &= \left[k_1(l)\,k_1(l')\right] + \left[R_l^1(l,l') + k_1(l,l')\,R_m^1(m,m') + R_l^1(l,l')\,R_m^1(m,m')\right] \\ &= R^0((l,m),(l',m')) + R^1((l,m),(l',m')) \end{aligned} \tag{2.10}$$

We must introduce some notation to simplify the following expression of the form of the elements in $\mathcal{H}$. Denote the set of unique pairs of observed within-subject time points and the corresponding set of unique transformed coordinates by $\mathcal{W}$ and $\mathcal{W}^*$, respectively:

$$\begin{aligned}\mathcal{W} &= \bigcup_{i=1}^{N}\bigcup_{j>k}(t_{ij}, t_{ik}) \\ \mathcal{W}^* &= \bigcup_{i=1}^{N}\bigcup_{j>k}\left(t_{ij} - t_{ik}, \frac{1}{2}(t_{ij} + t_{ik})\right) = \bigcup_{i=1}^{N}\bigcup_{j>k}\left(l_{jk}^i, m_{jk}^i\right)\end{aligned}$$

with $|\mathcal{W}| = |\mathcal{W}^*| = N_{\phi^*}$. For simplicity of presentation, relabel the elements of $\mathcal{W}^*$ so that

$$\mathcal{W}^* = \{(l_1, m_1), (l_2, m_2), \ldots, \left(l_{N_{\phi^*}}, m_{N_{\phi^*}}\right)\}$$

Then we may verify that any $\phi^* \in \mathcal{H}$ can be written

$$\phi^*(l, m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{n} c_i R_1((l, m), (l_i, m_i)) + \rho(l, m)$$

where $\rho \perp \mathcal{H}_0 = \{1\} \oplus \{k_1\},\ span\{R_1((l_i, m_i), \cdot)\}$. We do so by demonstrating that ρ does not improve the first term in (2.6) (the data fit functional) and only adds to the penalty term, $J(\phi^*)$. Consequently, if $\hat{\phi}^*$ is the minimizer of (2.6), then $\rho = 0$. Using the properties of reproducing kernels, we can rewrite ϕ^* as an inner product of itself with R:

$$
\begin{aligned}
\phi^*(l_j, m_j) &= \langle R((l_j, m_j), (\cdot,\cdot)), \phi^*(\cdot,\cdot)\rangle \\
&= \langle R_0((l_j, m_j), (\cdot,\cdot)) + R_1((l_j, m_j), (\cdot,\cdot)), d_0 + d_1 k_1(\cdot) \\
&\qquad + \sum_{i=1}^{N_{\phi^*}} c_i R_1((l_i, m_i), (\cdot,\cdot)) + \rho((\cdot,\cdot))\Big\rangle \\
&= \langle R_0((l_j, m_j), (\cdot,\cdot)), d_0 + d_1 k_1(\cdot)\rangle + \left\langle R_0((l_j, m_j), (\cdot,\cdot)), \sum_{i=1}^{N_{\phi^*}} c_i R_1((l_i, m_i), (\cdot,\cdot))\right\rangle \\
&\quad + \langle R_0((l_j, m_j), (\cdot,\cdot)), \rho((\cdot,\cdot))\rangle + \langle R_1((l_j, m_j), (\cdot,\cdot)), d_0 + d_1 k_1(\cdot)\rangle \\
&\quad + \left\langle R_1((l_j, m_j), (\cdot,\cdot)), \sum_{i=1}^{N_{\phi^*}} c_i R_1((l_i, m_i), (\cdot,\cdot))\right\rangle + \langle R_1((l_j, m_j), (\cdot,\cdot)), \rho((\cdot,\cdot))\rangle \\
&= \langle R_0((l_j, m_j), (\cdot,\cdot)), d_0 + d_1 k_1(\cdot)\rangle + \left\langle R_1((l_j, m_j), (\cdot,\cdot)), \sum_{i=1}^{N_{\phi^*}} c_i R_1((l_i, m_i), (\cdot,\cdot))\right\rangle \\
&\quad + \underbrace{\langle R_0((l_j, m_j), (\cdot,\cdot)), \rho(\cdot,\cdot)\rangle}_{0} + \underbrace{\langle R_1((l_j, m_j), (\cdot,\cdot)), \rho(\cdot,\cdot)\rangle}_{0} \\
&= d_0 + d_1 k_1(\cdot) + \sum_{i=1}^{N_{\phi^*}} c_i R_1((l_i, m_i), (l_j, m_j))
\end{aligned}
$$

Rewriting the data fit functional, we have that

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y(t_{ij}) - \sum_{k=1}^{j-1}\phi^*(t_{ij}, t_{ik})\, y(t_{ik})\right)^2 \\
&= \sum_{i=1}^{N}\sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y(t_{ij}) - \sum_{k=1}^{j-1}\left\langle R\left(\left(l_{jk}^i, m_{jk}^i\right), (\cdot,\cdot)\right), \phi^*(\cdot,\cdot)\right\rangle y(t_{ik})\right)^2
\end{aligned}
$$

which is free of ρ. Consider the contribution of any nonzero ρ to $J(\phi^*)$:

$$
\begin{aligned}
J(\phi^*) &= ||P_1\phi^*||^2 \\
&= \left\langle \sum_{i=1}^{N_{\phi^*}} c_i R_1\left((l_i, m_i), (\cdot,\cdot)\right) + \rho(\cdot,\cdot), \sum_{j=1}^{N_{\phi^*}} c_j R_1\left((l_j, m_j), (\cdot,\cdot)\right) + \rho(\cdot,\cdot) \right\rangle \\
&= || \sum_{i=1}^{N_{\phi^*}} c_i R_1\left((l_i, m_i), (\cdot,\cdot)\right) ||^2 + ||\rho||^2
\end{aligned}
$$

Thus, including ρ in ϕ^* only increases the penalty without improving (decreasing) the data fit functional, so we indeed have that the minimizer of (2.6) has the form

$$\phi^*(l,m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{N_{\phi^*}} c_i R_1\left((l,m),(l_i,m_i)\right) \tag{2.11}$$

Consider moving along the entries of Σ in the l direction toward the diagonal. T was explicitly constructed with unit diagonal so that $\phi_{tt} = \phi^*_{0t} = 1$. Considering this constraint when viewing ϕ^* as a continuous function,

$$\lim_{l \to 0} \phi^*(l,m) = 1 \quad for\ all\ m \in (0,1) \tag{2.12}$$

To obtain solutions satisfying (2.12), we isolate the seach the minimizer of (2.6) to those belonging to the convex subspace of $\mathcal{H}$ given by

$$\mathcal{C} = \{\phi^* : \phi^*(0,m), \quad m \in (0,1)\}$$

We approximate the set of functions ϕ^* satisfying the infinite set of linear constraints, $\mathcal{C}$, with functions satisfying the finite family of constraints, $\mathcal{C}_{N_m}$:

$$\mathcal{C}_{N_m} = \{\phi^* : \phi^*(0,m_j), \quad j = 1,\ldots,N_m\}$$

where N_m denotes the total number of unique observed values of m: $N_m = |\mathcal{M}|$, $\mathcal{M} = \bigcup_{i=1}^{N_{\phi^*}} m_i$. Using a similar argument to that presented above, Villalobos and Wahba have shown that if

$$L_1, \ldots, L_{N_{\phi^*}}, C_1, \ldots, C_{N_m}$$

are linearly independent functionals with L_i being the usual evaluation functional and C_j defines the jth linear constraint in that $C_j\phi^* = \phi^*(0, m_j)$, then the unique minimizer of (2.6) has the form

$$\phi^*(l,m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{N_{\phi^*}} c_i R_1\left((l,m),(l_i,m_i)\right) + \sum_{j=1}^{N_m} b_j R_1\left((l,m),(0,m_j)\right) \tag{2.13}$$

Expressing the components in the ANOVA decomposition of ϕ^* given by (2.9) as expansions of $\{1\}, \{k_1\}$, and $\{R_1\left((l_i,m_i),(\cdot,\cdot)\right) = R_1^l(l_i,\cdot) + R_1^m(m_i,\cdot) + R_1^{lm}\left((l_i,m_i),(\cdot,\cdot)\right)\}$, we can write

$$\begin{aligned}
\phi^*(l,m) &= \mu^* + \phi_l^*(l) + \phi_m^*(m) + \phi_{lm}^*(l,m) \\
&= d_0 + \left[d_1 k_1(l) + \sum_{i=1}^{N_{\phi^*}} c_i R_1^l(l,l_i) + \sum_{j=1}^{N_m} b_j R_1^l(l,0) \right] \\
&\quad + \left[\sum_{i=1}^{N_{\phi^*}} c_i R_1^m(m,m_i) + \sum_{j=1}^{N_m} b_j R_1^m(m,m_j) \right] \\
&\quad + \left[\sum_{i=1}^{N_{\phi^*}} c_i R_1^{lm}\left((l,l_i),(m,m_i)\right) + \sum_{i=1}^{N_m} b_j R_1^{lm}\left((l,0),(m,m_j)\right) \right]
\end{aligned}$$

With this decomposition of ϕ^*, it is easy to see that any non-stationarity of $y(t)$ is captured by the main effect of m, ϕ_m^* and the $l - m$ interaction, ϕ_{lm}^*. The function spaces corresponding to ϕ_m^* and ϕ_{lm}^* belong entirely to $\mathcal{H}_1$, so that any functional component representing non-stationarity is penalized, including those belonging to the span of linear functions.

2.4 Model Fitting via Penalized Maximum Likelihood

Let Φ^* be the $N_{\phi^*} \times 1$ vector of regression coefficients given by (2.5) corresponding to ϕ^* evaluated at the elements of $\mathcal{W}^*$, $\Phi^* = \left(\phi_1^*, \phi_2^*, \ldots, \phi_{N_{\phi^*}}^*\right)^T$. Let $d = (d_0, d_1)^T$, $c = \left(c_1, \ldots, c_{N_{\phi^*}}\right)^T$, and $b = (b_1, \ldots, b_{N_m})^T$. Define K_{11}, K_{12}, K_{22}, B_1, and B_2 as follows:

$$
\begin{array}{ll}
K_{11}[i,j] = R_1((l_i, m_i),(l_j, m_j)) & i,j = 1,\ldots,N_{\phi^*} \\
K_{12}[i,j] = R_1((l_i, m_i),(0, m_j)) & i = 1,\ldots,N_{\phi^*},\ j = 1,\ldots,N_m \\
K_{22}[i,j] = R_1((0, m_i),(0, m_j)) & i,j = 1,\ldots,N_m \\
B_1[i,j] = k_j(l_i) & i = 1,\ldots,N_{\phi^*},\ j = 1,2 \\
B_2[i,j] = k_j(0) & i = 1,\ldots,N_m,\ j = 1,2
\end{array}
$$

$$
K = \begin{bmatrix} K_{11} & K_{12} \\ K_{12}^T & K_{22} \end{bmatrix} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix}; \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}
$$

In matrix notation:

$$
\Phi^* = B_1 d + K_1 a
$$

where a is the $(N_{\phi^*} + N_m) \times 1$ vector $a = \left(c^T\ b^T\right)^T$. Let $Y_{i,(-1)}$ denote the vector of the last $n_i - 1$ responses for individual i:

$$
Y_{i,(-1)} = \left(y(t_{i2}), y(t_{i3}), \ldots, y(t_{i,n_i})\right)^T
$$

and concatenate these to construct the $(n - N) \times 1$ vector $Y_{(-1)} = \left(Y_{1,(-1)}^T, Y_{2,(-1)}^T, \ldots, Y_{N,(-1)}\right)^T$ where $n = \sum_{i=1}^N n_i$. Let $D = diag\left(\sigma_{12}^2, \ldots, \sigma_{1,n_1}^2, \ldots, \sigma_{N2}^2, \ldots, \sigma_{N,n_N}^2\right)$. For appropriate specification of $(n - N) \times N_{\phi^*}$ design matrix, Z_Y, our goal is to minimize:

$$
Q(a,d) = \frac{1}{2}\left(Y_{(-1)} - Z_Y(B_1 d + Ka)\right)^T D^{-1}\left(Y_{(-1)} - Z(Bd + Ka)\right) + \lambda a^T K c \tag{2.14}
$$

subject to $B_2 d + K_2 a - \mathbf{1} = 0$, where $Q(a,d)$ is the log-likelihood of $e = Y_{(-1)} - \hat{Y}_{(-1)} \sim \mathcal{N}(0, D)$ Augmenting the $(n - N) \times 1$ residual vector $Y_{(-1)} - Z_Y(B_1 d + Rc)$ with the $N_m \times 1$ vector of zeros corresponding to these constraints leaves the data fit functional in (2.14) unchanged. Let

$$
Z_{(n-N)\times\left(N_{\phi^*}+N_m\right)} = [Z_Y\ \mathrm{I}_{N_m}]
$$

$$
Y_{aug} = \left(Y_{(-1)}^T\ \ \mathbf{1}^T\right)^T
$$

and augment the diagonal entries of D with $\mathbf{1}_{N_m}$. Define

$$
Q^*(a,d) = [Y_{aug} - Z(Bd + Ka)]^T D^{-1}[Y_{aug} - Z(Bd + Ka)] + \lambda a^T K a \tag{2.15}
$$

We obtain the following normal equations by differentiation with respect to a and d:

$$\frac{\partial Q^*}{\partial a} = -(ZK)^T D^{-1} [Y_{aug} - Z(Bd + Ka)] + \lambda Ka = 0 \quad (2.16)$$

$$\frac{\partial Q^*}{\partial d} = -(ZB)^T D^{-1} [Y_{aug} - Z(Bd + Ka)] = 0 \quad (2.17)$$

Using the QR decomposition of B, we may write $B = [Q_1\ Q_2] \begin{bmatrix} R \\ 0 \end{bmatrix} = Q_1 R$, Q othogonal. Premultiplying (2.17) by K^{-1}, we have

$$\lambda a = Z^T D^{-1} [Y_{aug} - Z(Bd + Ka)] \quad (2.18)$$

$$\implies 0 = (2.17) = -B^T \left[Z^T D^{-1} (Y_{aug} - Z(Bd + Ka)) \right]$$

$$= -\lambda B^T a$$

$$\implies a = Q_2 e \ for\ e \in \mathcal{R}^{N_{\phi^*} + N_m - 2}$$

Multiplying (2.18) by $\left(Z^T D^{-1} Z\right)^{-1}$, (which we note is full rank as long as $N_Y \geq N_{\phi^*}$) we have

$$\left(Z^T D^{-1} Z\right)^{-1} Z^T D^{-1} Y_{aug} = Bd + \left[K + \lambda \left(Z^T D^{-1} Z\right)^{-1} \right] a \quad (2.19)$$

$$\left(Z^T D^{-1} Z\right)^{-1} Z^T D^{-1} Y_{aug} = Bd + \left[K + \lambda \left(Z^T D^{-1} Z\right)^{-1} \right] Q_2 e$$

$$\implies Q_2^T \left(Z^T D^{-1} Z\right)^{-1} Z^T D^{-1} Y_{aug} = Q_2^T \left[K + \lambda \left(Z^T D^{-1} Z\right)^{-1} \right] Q_2 e$$

$$\implies e = \left[Q_2^T \left[K + \lambda \left(Z^T D^{-1} Z\right)^{-1} \right] Q_2 \right]^{-1} Q_2^T \left(Z^T D^{-1} Z\right)^{-1} Z^T D^{-1} Y_{aug}$$

$$\implies a = Q_2 \left[Q_2^T \left[K + \lambda \left(Z^T D^{-1} Z\right)^{-1} \right] Q_2 \right]^{-1} Q_2^T \left(Z^T D^{-1} Z\right)^{-1} Z^T D^{-1} Y_{aug}$$

Solving for d using (2.19) and the QR decomposition of B, we obtain

$$Bd = Q_1 R = \left(Z^T D^{-1} Z\right)^{-1} Z^T D^{-1} Y_{aug} - \left[K + \lambda \left(Z^T D^{-1} Z\right)^{-1} \right] a$$

$$\implies d = R^{-1} Q_1^T \left[\left(Z^T D^{-1} Z\right)^{-1} Z^T D^{-1} Y_{aug} - \left[K + \lambda \left(Z^T D^{-1} Z\right)^{-1} \right] a \right]$$

By the construction of $\mathcal{H}$ as defined in (2.8), $J(\phi^*) = ||\phi_l^{*\prime\prime}||^2 + ||\phi_m^*||^2 + ||\phi_{lm}^*||^2$. The null space of $J(\cdot)$, $\mathcal{H}_0$, is the set of functions $\{\phi \in \mathcal{H} : J(\phi) = 0\}$:

$$\mathcal{H}_0 = \mathcal{H}_\mu \oplus \mathcal{H}_l^0 = \{1\} \oplus \{k_1\}$$

The null space is comprised of functions of the form $\phi_0^*(l, m) = d_0 + d_1 k_1(l)$ which obey stationarity in the autoregressive process defined by (2.4), belonging to the class of linear functions of the continuous time lag only. While the solution defined as the minimizer of

Appendix A: The Data on Cows

This is the data that was used to produce the table in Table **??**. In 1990, 57 people had cows. In 1991, 80 people had cows. In 1992, 199 people had cows.

Appendix B: Important commands defined in `osudissert96`

The following is a list of all commands available in `osudissert96`.

```
\author{First Middle Last}
\title{The title of the thesis}
\authordegrees{degree1, degree2}
\unit{Department of Whatever The Name Is}
\degree{Doctor of Philosophy}
\committee{Dissertation}
\advisorname{Name of advisor}  % Possible usage "Prof. Big Dude"
\member{Name of committee member}

\thesis  % this makes it a thesis rather than a dissertation,
         %  similar to the documentclass option [masters] or [ms].

\maketitle

\disscopyright      % or \blankpage

\begin{abstract}
\end{abstract}

\begin{externalabstract}
\end{externalabstract}


\dedication{This is dedicated to \ldots}

\begin{acknowledgements}
\end{acknowledgements}

\begin{vita}
```

```
\dateitem{Important Date}{ Why its important}

\begin{publist}
\researchpubs
\pubitem{Bibliography item (from BibTeX?)}

\instructpubs
\pubitem{Bibliography item (from BibTeX?)}
\end{publist}

\begin{fieldsstudy}
\majorfield*        % which uses \unit above
% \majorfield{Your Major Field}
\begin{studieslist}
\studyitem{Topic 1}{Prof.\ 1}
\studyitem{Topic 2}{Prof.\ 2}
\studyitem{Topic 3}{Prof.\ 3}
\end{studieslist}
%% Note:  If there were only one field of study, the following list
%%        would best be done using the following command:
%%  \onestudy{Only Topic}{Only Professor}
\end{fieldsstudy}

\end{vita}
```