Estimating Variance Components in Functional Linear Models with Applications to Genetic Heritability

Matthew Reimherr *
Pennsylvania State University

Dan Nicolae
University of Chicago

Abstract

Quantifying heritability is the first step in understanding the contribution of genetic variation to the risk architecture of complex human diseases and traits. Heritability can be estimated for univariate phenotypes from non-family data using linear mixed effects models. There is, however, no fully developed methodology for defining or estimating heritability from longitudinal studies. By examining longitudinal studies, researchers have the opportunity to better understand the genetic influence on the temporal development of diseases, which can be vital for populations with rapidly changing phenotypes such as children or the elderly. To define and estimate heritability for longitudinally measured phenotypes, we present a framework based on functional data analysis, FDA. While our procedures have important genetic consequences, they also represent a substantial development for FDA. In particular, we present a very general methodology for constructing optimal, unbiased estimates of variance components in functional linear models. Such a problem is challenging as likelihoods and densities do not readily generalize to infinite dimensional settings. Our procedure can be viewed as a functional generalization of the minimum norm quadratic unbiased estimation procedure, MINQUE, presented by C. R. Rao, and is equivalent to residual maximum likelihood, REML, in univariate settings. We apply our methodology to the Childhood Asthma Management Program, CAMP, a four year longitudinal study examining the long term effects of daily asthma medications on children.

1 Introduction

Heritability is the proportion of phenotypic variation that is attributable to genetic factors. Estimating heritability is important for assessing the efficiency of disease/trait gene discovery studies, for evaluating the potential to predict genetic risk for disease, and for comparing traits across groups and populations. It can be also be utilized to quantify the

* *Address for correspondence*: Matthew Reimherr, Department of Statistics, Pennsylvania State University, University Park, PA, 16802, USA. E-mail: mreimherr@psu.edu

so called *heritability gap*; the difference between what phenotypic variation we think should be explainable by genetics, and what we can currently explain via known disease variants identified from genome wide association studies. Recently, methods for partitioning heritability into classes of genetic variation [41, 18] have been used to advance our knowledge on what drives the genetic architecture of human traits: rare versus common variation [40], expression quantitative trait loci (eQTLs) and other classes of functional variation [8].

Traditionally, heritability has been estimated by comparing trait similarity between close relatives such as twins, siblings or parent-offspring pairs. Recently, there have been a number of methods introduced for estimating the heritability from panels of genome-wide single nucleotide polymorphisms, SNPs, using non–family based data in scalar settings [40, 15, 38, 44, 43]. These studies may consist of hundreds of thousands or even a million SNPs, making this an ultrahigh dimensional problem. To handle such data, these approaches employ a methodology similar to that of classical heritability estimation, but genetic similarity is estimated from SNP data and not inferred from familial relationships. Geneticists often perceive heritability as some fixed quantity that is independent of time. However, this clearly need not be true, especially for populations that are changing rapidly with age (young, older, sick populations etc.). For example, cholesterol levels and other blood biomarkers have a much larger genetic component in the first few years of life than in later years when environmental factors (such as diet, climate etc.) start impacting these traits. Ignoring longitudinal trends when estimating the heritability of such biomarkers in an age-diverse or rapidly changing group of subjects will lead to estimates that could be biased in unexpected ways; there may be complex temporal patterns which are heritable, but cannot be seen by examining a single time point.

Extending scalar based methods to longitudinal settings is challenging as there is no unique way of defining *variation* in higher dimensions. Furthermore, it may be desirable to have a measure which is not highly sensitive to the temporal spacing of the data since this could cause different studies to obtain different results even under otherwise similar

conditions. A framework built on functional data analysis, FDA, allows us to develop meaningful measures which address theses concerns. FDA is a rapidly growing branch of statistics which is concerned with developing statistical procedures for data which come from smooth stochastic processes. Such methods have been developed and applied very successfully in areas such as human growth patterns [27, 5, 39], genetics [35, 22, 30], finance [34, 16, 17], geomagnetic activity patterns [10, 9], and neuroimaging [7, 31, 45], to name only a few. Here each subject would contribute a curve representing the evolution of the interested phenotype over time, though that curve may only be observed at a small number to time points and corrupted with noise. A common approach for applying FDA methods is to use the large body of methods for nonparametric smoothing to first estimate the subject by subject curves, and then apply statistical methods for Hilbert space models. Such an approach can be very effective when working with a relatively balanced temporal design and data which is believed to arise from smooth stochastic processes. For data which is observed at a very sparse and irregular number of time points, an approach based on pooled nonparametric smoothing would be more beneficial. Throughout the paper we will work with objects of the form

$$Y_n(t) = \text{value of phenotype for subject } n \text{ at time } t,$$

where nonparametric smoothers (in our case penalized B–Splines) are used to construct the entire $Y_n(t)$ curve. For N subjects, we then analyze the $Y_1, \ldots, Y_N$ curves as a random sample from $L^2(\mathcal{T})$, where $\mathcal{T}$ is the time interval of the study. An illustration of such objects can be seen in Figure 1 which plots the smoothed phenotypic curves for a subset of 100 subjects from the Childhood Asthma Management Program (CAMP), which we discuss in detail in Section 6. More details on this data and the application of B–Splines to it can be found in [30]; they also include a plot of the raw data to compare with the smoothed trajectories we work with here.

Current methods for estimating the heritability of a univariate phenotype with non–family based data utilize mixed effects models [40]. A random effect is introduced for each

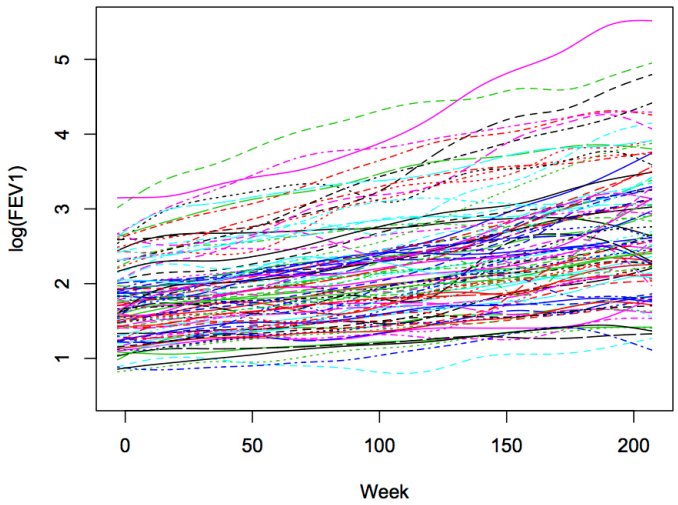

Figure 1: Plots of the log($FEV1$) curves for the first one hundred children from CAMP. Each trajectory is estimated using penalized B–Splines.

SNP, so that collectively, the variation of the phenotype can be factored into variation coming from the SNPs and variation coming from a random error. Individuals which have more similar genetic make ups will have more correlated phenotypes depending on the heritability of the trait. Generalizing this idea to the functional setting, leads to the *functional mixed effects model*, FME,:

$$Y_n(t) = \sum_{j=1}^{J} X_{nj}\beta_j(t) + \sum_{k=1}^{K} G_{nk}\alpha_k(t) + \varepsilon_n(t), \tag{1}$$

where X_{nj} are non–SNP predictors and G_{nk} are the genotypes. We treat the β_j as fixed effects and α_k as random, which provides an analogous factorization of the variability of Y_n that we can use to estimate heritability. A number of researchers have explored various versions of the functional mixed effects model. Some key results include the following: an FME with within subject dependence only is examined in [11], a Bayesian wavelet approach for a very general FME is explored in [24], a wavelet approach (non-Bayesian) is given in [1], a methodology based on functional principle components is discussed in [2], and a more general penalized regression procedure is presented in [32] (as well as a wider view of the

literature which may be useful to readers). However, none of these methods are designed to handle the case where K is large. Indeed, maybe the best (currently) developed method is implemented in the R function *pffr()* [14] in the refund package which can include random effects which are expressed as B–Spline expansions. While pffr does not include inferential tools for the covariance estimates, it can fit a functional mixed model. However, applying it to a subset of our data with only $K = 5$ ($N = 540$), the function takes around 3.5 hours on a Intel quad-core i7 desktop with 16GB of ram; with $K = 10$ it takes around 8 hours. Thus implementing it in our data application is currently not an option. We tested several other methods, and we always ran into a problem with computation. None of the methods were designed around having very large K or included any sort of inferential tools for the covariance estimates. Our approach is quite different from previous procedures in that we first carry out subject by subject smoothing and then use Hilbert space theory to construct optimal quadratic forms. While our methods are presented in the context of L^2, very similar arguments would work for more general Hilbert space valued observations. Additionally, the number of random effects, K, in genetic settings, may be in the millions, meaning that $K \gg N$, making this an ultrahigh dimensional problem and very different from more typical random effects settings. We refer readers to [12] for an extensive introduction to Hilbert space methods for functional data.

Statistical theory for mixed effects models is well developed, with a multitude of texts on the subject. We mention [33] which is particularly useful for a background of the methods presented here. While the functional analogue of the mixed effects model is conceptually straightforward, the additional estimation techniques and theory therein is not. Generally, it is difficult to define densities or likelihoods for infinite dimensional objects, and thus maximum likelihood or residual maximum likelihood techniques do not readily transfer. Thus we propose a methodology we call *Functional MINQUE* or *F-MINQUE*, which extends the MINQUE (minimum norm, quadratic unbiased estimate) procedure developed in [28] and [29] (see also the MINVAR procedure in [19] and [20]). MINQUE is concerned with finding

optimal quadratic forms for estimating variance components. The I–MINQUE procedure [4], on which our procedure is based, in the scalar setting is equivalent to residual maximum likelihood (REML) with normally distributed random effects and errors [33, Pg. 398]. However, I–MINQUE is not directly based on likelihoods which gives a clearer generalization to the functional setting. The idea behind MINQUE is as follows. Let $\mathbf{Y}$ be an N dimensional random vector representing the response variable of interest. Assume that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta},$$

where $\mathbf{X}$ is a matrix of predictors, $\boldsymbol{\beta}$ is a vector of fixed effects, and $\boldsymbol{\delta}$ is the error of the model. Assume that the error satisfies the following moment assumptions:

$$\mathrm{E}[\boldsymbol{\delta}] = 0 \qquad \text{and} \qquad \mathrm{E}[\boldsymbol{\delta}\boldsymbol{\delta}^T] = \sum_{i=1}^{I} \mathbf{H}_i \tau_i,$$

where $\mathbf{H}_i$ are known matrices and τ_i are unknown positive scalar parameters that need to be estimated. Then the goal of MINQUE is to find optimal unbiased estimators of the form $\hat{\tau}_i = \mathbf{Y}^T \mathbf{A}_i \mathbf{Y}$, where $\mathbf{A}_i$ is an $n \times n$ matrix. In F–MINQUE we have the same goal, but now the response variables are functions.

The remainder of the paper is organized as follows. In Section 2 we present the functional mixed effect model and define heritability for functional phenotypes. In Section 3 we outline our estimation procedure for a general covariance structure which includes the functional mixed effects model as a special case. In Section 4 we provide some asymptotic guarantees which are demonstrated via simulations in Section 5, and in Section 6 we apply our methodology to data coming from the Childhood Asthma Management Program, CAMP.

2 Functional Mixed Effects Model

Denote by $Y_n(t)$ the value of a particular quantitative phenotype measured at time t on subject n. Our present goal is to determine what proportion of the variability in $Y_n(t)$ can be attributed to genetic factors, commonly known as *heritability.* However, our aim is not simply to determine heritability at a fixed t, but globally for the entire stochastic process. To that end, we use a functional mixed effects model to allow for fixed covariates (such as gender, race, etc), while also including random effects which, in the context of genetics, induce dependence between subjects driven by their relatedness, i.e. how similar they are genetically.

Definition 1 (Functional Mixed Effects Model)**.** *Assume that $\{Y_n\}$ are observed random functions in $L^2(\mathcal{T})$, where $\mathcal{T}$ is a closed interval. Let $\{\mathbf{X}_n\}$ and $\{\mathbf{G}_n\}$ be observed sequences of J and K dimensional covariate vectors respectively. Assume the following relationship for $n = 1, \ldots, N$*

$$\begin{aligned} Y_n(t) &= \sum_{j=1}^{J} X_{nj}\beta_j(t) + \sum_{k=1}^{K} G_{nk}\alpha_k(t) + \varepsilon_n(t) \\ &= \mathbf{X}_n^T\boldsymbol{\beta}(t) + \mathbf{G}_n^T\boldsymbol{\alpha}(t) + \varepsilon_n(t), \end{aligned}$$

where $\boldsymbol{\beta}(t)$ are the fixed effects whose coordinates are in $L^2(\mathcal{T})$. The sequences $\{\alpha_k\}$ and $\{\varepsilon_n\}$ are each iid random elements of $L^2(\mathcal{T})$, independent of each other, with zero means and finite covariances.

In the above definition we do not include a Gaussian assumption, though it will be used later on. Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_N(t))^T$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T$, $\mathbf{G} = (\mathbf{G}_1, \ldots, \mathbf{G}_N)^T$, and $\boldsymbol{\varepsilon}(t) = (\varepsilon_1(t), \ldots, \varepsilon_N(t))^T$. Then we have that

$$\mathbf{Y}(t) = \mathbf{X}\boldsymbol{\beta}(t) + \mathbf{G}\boldsymbol{\alpha}(t) + \boldsymbol{\varepsilon}(t).$$

The major difference between the above and a typical random effects model is that the fixed effects, random effects, and error terms take values in a function space. In our genetics context the $\mathbf{G}_n$ would be vectors consisting of hundereds of thousands of standardized genotypes.

Genetic Heritability

To quantify genetic heritability, we exploit the mixed effects factorization and the underlying function space, $L^2(\mathcal{T})$. If we assume that $\mathrm{E}\,\|\alpha_k\|^2 < \infty$ and $\mathrm{E}\,\|\varepsilon_n\|^2 < \infty$, then we can express the total variability of the function Y_n as

$$\mathrm{E}\,\|Y_n - \mathbf{X}_n^T\boldsymbol{\beta}\|^2 = \int_{\mathcal{T}} \mathrm{E}[(Y_n(t) - \mathbf{X}_n^T\boldsymbol{\beta}(t))(Y_n(t) - \mathbf{X}_n^T\boldsymbol{\beta}(t))]\ dt.$$

Recall that the $L^2(\mathcal{T})$ norm is given by

$$\|x\|^2 = \int_{\mathcal{T}} x(t)^2\ dt.$$

Throughout, whenever we write the norm of a function, we will mean the $L^2(\mathcal{T})$ norm. From here on, we omit the domain $\mathcal{T}$ as it will be understood that all integrals are over $\mathcal{T}$ unless otherwise stated. We can express the integrand as

$$\mathrm{E}[(Y_n(t) - \mathbf{X}_n^T\boldsymbol{\beta}(t))(Y_n(t) - \mathbf{X}_n^T\boldsymbol{\beta}(t))] = \mathbf{G}_n^T\mathbf{G}_n C_\alpha(t,t) + C_\varepsilon(t,t),$$

where C_α and C_ε are the covariance functions of α_k and ε_n respectively. The variability of Y_n can therefore be factored into one piece driven by genetic factors and a second piece given by an independent error term:

$$\mathrm{E}\,\|Y_n - \mathbf{X}_n^T\boldsymbol{\beta}\|^2 = \mathbf{G}_n^T\mathbf{G}_n \int C_\alpha(t,t)\ dt + \int C_\varepsilon(t,t)\ dt.$$

Typically the matrix $\mathbf{G}$ is standardized such that $N^{-1}\sum \mathbf{G}_n^T\mathbf{G}_n = 1$. Though such a standardization is by no means necessary, it does lead to a relatively simple way of defining heritability. In particular, the heritability of a functional phenotype can be defined as as the proportion of L^2–variation explained by genetics:

$$H := \frac{\int C_\alpha(t,t)\ dt}{\int C_\alpha(t,t)\ dt + \int C_\varepsilon(t,t)\ dt} = \frac{\mathrm{E}\,\|\alpha_1\|^2}{\mathrm{E}\,\|\alpha_1\|^2 + \mathrm{E}\,\|\varepsilon_1\|^2}.$$

The heritability, H, may be interpreted as the proportion of explained variability averaged over time. We emphasize that the standardization which lead to the above expression is not necessary, but simply a matter of notational convenience. The primary idea behind the above measure is that one can define heritability as the ratio of the magnitude of the random effects term and the overall variability of the process. Notice that once the terms C_α and C_ε can be estimated, a local time varying measure of heritability can also be utilized. In the application section, we present both. Furthermore, it might be useful to consider the first few eigenfunctions of C_α, which can be interpreted as the *most heritable components* or shapes of the phenotype, i.e., the temporal patterns in the data which are driven most by genetics.

3 Functional MINQUE

We present a very general form of the functional MINQUE procedure, F-MINQUE, which includes the discussed genetic scenarios, but also allows for structures found in more classical linear mixed effects settings. Assume that the response function $\mathbf{Y}(t)$ can be expressed as

$$\mathbf{Y}(t) = \mathbf{X}\boldsymbol{\beta}(t) + \boldsymbol{\delta}(t)$$

where $\boldsymbol{\delta}$ is Gaussian with

$$\mathrm{E}[\boldsymbol{\delta}(t)] = 0 \qquad \text{and} \qquad \mathrm{E}[\boldsymbol{\delta}^T(t)\boldsymbol{\delta}(s)] = \sum_{i=1}^{I} \mathbf{H}_i C_i(t,s).$$

Here $\{\mathbf{H}_i\}$ are known $N \times N$ linearly independent (for identifiability) matrices and $C_i(t,s)$ are unknown covariance functions. Typically one assumes that $\mathbf{H}_I$ is the identity so that the last term corresponds to the random error term in the model, though this is not required. Working with the $\{\mathbf{H}_i\}$ allows for a more general framework, as well as being more convenient notationally. Furthermore, it is the $\{\mathbf{H}_i\}$ which uniquely determine the model as any invertible rotation may be applied to $\mathbf{G}$ from the right (i.e. on the rows) without changing the resulting distribution. For quantifying genetic heritability we would take $I = 2$, $\mathbf{H}_1 = \mathbf{G}\mathbf{G}^T$, and $\mathbf{H}_2 = \mathbf{I}$. In our application we also explored how the heritability is distributed amongst the 22 autosomal chromosomes (not shown here). In that case, $I = 23$ for the 22 autosomal chromosomes and one error term. The $\mathbf{H}_i$ would be obtained by dividing the $\mathbf{G}$ up by chromosome. However, there are a wide range of other models that fit within this framework. This includes block/random intercept models where, for example, $\mathbf{H}_1$ would be a block diagonal matrix representing block membership, and $\mathbf{H}_2$ would be the identity. Another especially useful application of our methods would be to longitudinal studies (not necessarily genetic) consisting of related individuals. In that case one could include a random effect representing familial membership. One would have $\mathbf{H}_1$ being a block diagonal matrix consisting of the various families and $\mathbf{H}_2$ would be the error in the model. Additionally, one could have that the α_k in the mixed effects model are actually dependent, as long as the structure of that dependence is known enough to be absorbed by a $\mathbf{H}_i$ matrix (with a possible adjustment to $C_i(t,s)$ function).

The goal of the F-MINQUE procedure is then to construct estimates $\hat{C}_i$ which are quadratic forms,

$$\hat{C}_i(t,s) = \mathbf{Y}(t)^T \mathbf{A}_i \mathbf{Y}(s),$$

with $\mathbf{A}_i$ chosen such that $\mathrm{E}[\hat{C}_i] = C_i$ and $\mathrm{E}\,\|\hat{C}_i - C_i\|^2$ is minimized. Taking expected values of the quadratic forms, we can express

$$\mathrm{E}[\mathbf{Y}(t)^T \mathbf{A}_i \mathbf{Y}(s)] = \mathrm{trace}(\boldsymbol{\beta}^T(t)\mathbf{X}^T\mathbf{A}_i\mathbf{X}\boldsymbol{\beta}(s)) + \sum_{i_1=1}^{I} \mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1})C_{i_1}(t,s).$$

So the unbiasedness condition implies that we should choose $\mathbf{A}_i$ such that

$$\mathbf{X}^T\mathbf{A}_i\mathbf{X} = 0 \qquad \text{and} \qquad \mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}) = 1_{i=i_1}. \tag{2}$$

Since we want unbiasedness for all parameter values, this assumption is necessary. Assuming that $\mathbf{A}_i$ is of such a form, we then have

$$\mathrm{E}\,\|\hat{C}_i - C_i\|^2 = \int\int \mathrm{Var}(\hat{C}_i(t,s))\; dt\; ds.$$

However, for each fixed t and s the integrand is a multivariate quadratic form based on normal random vectors. Therefore one can show that (see for example Lemma 3)

$$\begin{aligned}\mathrm{Var}(\hat{C}_i(t,s)) &= 2\,\mathrm{trace}\left[\left(\mathbf{A}_i\sum_{i_1=1}^{I}\mathbf{H}_{i_1}C_{i_1}(t,s)\right)^2\right] \\ &= 2\sum_{i_1,i_2}\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2})C_{i_1}(t,s)C_{i_2}(t,s).\end{aligned}$$

So the target function we are attempting to minimize can be expressed as

$$2\sum_{i_1,i_2}\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2})c_{i_1,i_2}, \quad \text{with} \quad c_{i_1,i_2} = \int\int C_{i_1}(t,s)C_{i_2}(t,s)\; dt\; ds,$$

and subject to the constraints given in (2). Since the data can always be transformed such that $\mathbf{X}^T\mathbf{A}_i\mathbf{X} = 0$, we can assume, without loss of generality, that $\mathbf{X} = 0$. Such an approach is common in REML style methods. One would apply a linear transformation which is orthogonal to the fixed effects, estimate the variance components off of the transformed

data, and finally estimate the fixed effects on the original data. If $\mathbf{X}$ is $N \times J$ then the transforming will reduce the dimension from N to $N - J$. For example, $\mathbf{Y}(t)$ is originally N dimensional, but will be $N - J$ dimensional after the transformation. The optimal choice of $\mathbf{A}_i$ will depend on the underlying functions C_i, thus we propose the following iterative procedure and include its derivation in Appendix A.

1. **Initialize:** Let $c^0_{i_2,i_3} \equiv 1$. Compute the matrices, for $i_1 = 1, \ldots, I$,

$$\mathbf{B}^0_{i_1} = \text{vec}^{-1}\left[\left(\sum_{i_2,i_3} c^0_{i_2,i_3}(\mathbf{H}_{i_2} \otimes \mathbf{H}_{i_3})\right)^{-1} \text{vec}(\mathbf{H}_{i_1})\right],$$

$$\mathbf{D}^0(i_1, i_2) = \text{trace}(\mathbf{H}_{i_1}\mathbf{B}_{i_2}).$$

   The initial matrix for $\mathbf{A}_i$, for $i = 1, \ldots, I$, is then

$$\mathbf{A}^0_i = \left[((\mathbf{D}^{-1}\mathbf{1}_i)^T \otimes \mathbf{I}_{N\times N}) \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_I \end{pmatrix}\right],$$

   where $\mathbf{1}_i$ is a vector of length I which is zero everywhere except the i^{th} coordinate which is 1. The $\mathbf{B}$ and $\mathbf{D}$ matrices are just notationally convenient objects for solving the Lagrangian.

2. **Iterate:** Form the estimates

$$\hat{C}^0_i(t, s) = \mathbf{Y}(t)^T \mathbf{A}^{(0)}_i \mathbf{Y}(s).$$

   Compute the scalars

$$c^1_{i,i_1} = \int\int C^0_i(t, s) C^0_{i_1}(t, s)\ dt\ ds.$$

   Iterate, replacing c^{k-1}_{i,i_1} with c^k_{i,i_1} and labeling the new estimates $\mathbf{A}^{(k)}$, etc, for $k = 1, \ldots$. One can check for convergence by plotting the c^k_{i,i_1} across k.

3. **Truncate:** If the final estimates, $\hat{C}_i$, have any negative eigenvalues, then set those eigenvalues to 0 to ensure a nonnegative definite estimate.

Using starting values of $c^0_{i_2,i_3} \equiv 1$ is natural as such weights, in some sense, treat all the C_i as equal. In fact, even without the iteration, one still has a consistent estimator, though iterating achieves a lower mean squared error. The final truncation step is optional, but if one does not truncate the negative eigenvalues then the estimates need not be positive–semidefinite. This also occurs in the scalar setting as it is sometimes possible to obtain negative variance estimates.

Standard errors for the $\hat{C}_i$ can be found in Theorem 1 when $I = 2$ and for a general I in Lemma 3. Standard errors for $\hat{H}_i$ can be found by combining the delta method with Theorem 1 or Lemma 3, the arguments of which can be found in Corollary 1.

Estimating Fixed Effects

The F-MINQUE procedure was carried out assuming that we had "transformed out" the fixed effects. On the transformed data we can estimate the variance components, and then on the original data we can then estimate the fixed effects using generalized least squares. Notice that for $J < N$, we can find a $(N - J) \times N$ matrix $\mathbf{A}_X$ such that

$$\mathbf{A}_X\mathbf{X} = 0 \quad \text{and} \quad \mathbf{A}_X\mathbf{A}_X^T = \mathbf{I}_{(N-J)\times(N-J)}.$$

Such a matrix can be constructed using the singular value decomposition of $\mathbf{X}$. Define $\tilde{\mathbf{Y}}(t) = \mathbf{A}_X\mathbf{Y}(t)$. Then we have

$$\mathrm{E}[\tilde{\mathbf{Y}}(t)] = 0 \qquad \text{and} \qquad \mathrm{E}[\tilde{\mathbf{Y}}(t)\tilde{\mathbf{Y}}^T(s)] = \sum_{i=1}^{I} A_X\mathbf{H}_iA_X^TC_i(t,s) = \sum_{i=1}^{I} \tilde{\mathbf{H}}_iC_i(t,s).$$

So we can then estimate the $\{C_i\}$ using $\tilde{\mathbf{Y}}$ and $\{\tilde{\mathbf{H}}_i\}$. With those estimates in hand, we can now construct the generalized least squares estimator for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}(t) = (\mathbf{X}^T\mathbf{W}^{-1}(t)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}(t)\mathbf{Y}(t),$$

where

$$\mathbf{W}(t) = \sum_{i=1}^{I} \hat{C}_i(t,t)\mathbf{H}_i.$$

Implementation and Computational Considerations

While the F-MINQUE algorithm, as defined, is theoretically precise, it should not be applied "as is" to data sets with large sample sizes. The primary computational difficulty is obtaining the $\mathbf{B}_{i_1}$ matrices. Their explicit solution involves inverting an $N^2 \times N^2$ matrix, which is computationally prohibitive when N is even moderately sized. It is, therefore, much more efficient to work directly with the equation

$$\sum_{i_2,i_3} c_{i_2,i_3}\mathbf{H}_{i_2}\mathbf{B}_{i_1}\mathbf{H}_{i_3} = \mathbf{H}_{i_1}.$$

One can then use any number of algorithms to numerically solve for $\mathbf{B}_{i_1}$. Since all of the involved objects are symmetric, we have had excellent results in using the *symmlq* routine in MATLAB [26] which is based on conjugate gradient descent. Solving for $\mathbf{B}_{i_1}$ will require evaluating the left hand side of the above equation a number of times. Each iteration as stated involves multiplying 3 large matrices I^2 times. We can speed things up if we can reduce the number of matrix multiplications. To do this, we propose the following. Using the spectral theorem for symmetric matrices we can express

$$c_{i_2,i_3} = \sum_{j=1}^{I} \lambda_j v_j(i_2)v_j(i_3),$$

where λ_j and v_j are the eigenvalues and eigenvectors respectively. Since I is typically small, this can be done very quickly. We can then express

$$\sum_{i_2,i_3} c_{i_2,i_3}\mathbf{H}_{i_2}\mathbf{B}_{i_1}\mathbf{H}_{i_3} = \sum_{j=1}^{I}\lambda_j\left(\sum_{i_2} v_j(i_2)\mathbf{H}_{i_2}\right)\mathbf{B}_{i_1}\left(\sum_{i_3} v_j(i_3)\mathbf{H}_{i_3}\right) = \sum_{j=1}^{I}\lambda_j\mathbf{V}_j\mathbf{B}_{i_1}\mathbf{V}_j.$$

Working with the $\mathbf{V}_j$ we can reduce the number of summands from I^2 to I. Finally, we can omit those summands with extremely small λ_j. For example, in the iteration step, only one of the λ_j will be nonzero. By including a relatively small threshold on the λ_j we can further speed up the procedure by omitting those summands that contribute a negligible amount to the sum. The final equation will look like

$$\sum_{j=1}^{I}\lambda_j 1\{\lambda_j > h\}\mathbf{V}_j\mathbf{B}_{i_1}\mathbf{V}_j = \mathbf{H}_{i_1},$$

where h is some very small number. We have had good success with $h = 0.0001 \times \sum|\lambda_j|/I$, but any number that is small enough to catch the essentially zero eigenvalues will work. Carrying out these alterations we have found that full day calculations are reduced to the order of minutes. In the CAMP example discussed in Section 6, it took less than an hour (on an Apple laptop with 2.6 GHz i7 processor) to run through 10 iterations (convergence occurred in 5) of the F-MINQUE procedure with $N = 540$ subjects and $I = 23$ corresponding to 22 autosomal chromosomes and 1 error term. When combining all autosomal chromosomes and taking $I = 2$, the procedure takes less that 30 seconds.

4 Some Asymptotic Theory

Here we present some asymptotic guarantees of the procedure in the simpler setting when

$$\mathbf{Y}(t) = \mathbf{G}\boldsymbol{\alpha}(t) + \boldsymbol{\varepsilon}(t).$$

Asymptotic arguments for the more general case are a bit cumbersome and difficult to interpret since we do not assume a simple block structure. We make the following assumption.

Assumption 1. *Assume that*

$$\mathbf{Y}(t) = \mathbf{G}\boldsymbol{\alpha}(t) + \boldsymbol{\varepsilon}(t),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ *and* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$ *are independent* K *and* N *dimensional vectors, respectively, of iid Gaussian processes in* $L^2(\mathcal{T})$ *with mean zero and covariance* C_α *and* C_ε*. We further assume that* $\mathrm{E}\,\|\alpha_k\|^2 < \infty$ *and* $\mathrm{E}\,\|\varepsilon_n\|^2 < \infty$ *for* $1 \leq k \leq K$ *and* $1 \leq n \leq N$*.*

This implies that

$$\mathrm{E}[Y_{n_1}(t)Y_{n_2}(s)] = C_\alpha(t,s)\mathbf{G}_{n_1}^T\mathbf{G}_{n_2} + C_\varepsilon(t,s)1_{n_1=n_2},$$

or in matrix form

$$\mathrm{E}[\mathbf{Y}(t)\mathbf{Y}^T(s)] = C_\alpha(t,s)\mathbf{G}\mathbf{G}^T + C_\varepsilon(t,s)\mathbf{I}.$$

By the spectral theorem we can write

$$\mathbf{G}\mathbf{G}^T = \mathbf{U}\mathbf{d}\mathbf{U}^T,$$

where $\mathbf{d} = \mathrm{diag}(d_1, d_2, \ldots, d_N)$ is a diagonal matrix of eigenvalues of $\mathbf{G}^T\mathbf{G}$, and $\mathbf{U}$ is a matrix of eigenvectors. As we will see later on, the asymptotic properties for our estimates depend directly on the behavior of the $\{d_n\}$. Since $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are Gaussian, we can write

$$\mathbf{Y}(t) = \mathbf{U}\mathbf{d}^{1/2}\mathbf{U}^T\boldsymbol{\alpha}'(t) + \boldsymbol{\varepsilon}(t),$$

where $\boldsymbol{\alpha}'$ is now an N dimensional vector of iid mean zero gaussian processes with covariance function C_α. If we multiply both sides by $\mathbf{U}^T$ we obtain

$$\mathbf{U}^T\mathbf{Y}(t) = \mathbf{d}^{1/2}\mathbf{U}^T\boldsymbol{\alpha}'(t) + \mathbf{U}^T\boldsymbol{\varepsilon}(t).$$

Now notice that the covariance of $\mathbf{U}^T\boldsymbol{\varepsilon}$ is $\mathbf{U}^T\mathbf{I}\mathbf{U}C_\varepsilon(t,s) = \mathbf{I}C_\varepsilon(t,s)$, and similarly the covariance of $\mathbf{U}^T\boldsymbol{\alpha}'(t)$ is $\mathbf{I}C_\alpha(t,s)$. Therefore, we can write

$$\mathbf{U}^T\mathbf{Y}(t) = \mathbf{Y}'(t) = \mathbf{d}^{1/2}\mathbf{U}^T\boldsymbol{\alpha}'(t) + \mathbf{U}^T\boldsymbol{\varepsilon}(t) = \mathbf{d}^{1/2}\boldsymbol{\alpha}'(t) + \boldsymbol{\varepsilon}'(t).$$

Thus we have (linearly) transformed the data so that the coordinates are independent (since we are assuming normality), but not identically distributed. For notational simplicity, we drop the primes and assume that

$$\mathbf{Y}(t) = \mathbf{d}^{1/2}\boldsymbol{\alpha}(t) + \boldsymbol{\varepsilon}(t), \tag{3}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are two independent N–dimensional vectors whose coordinates are iid Gaussian processes with covariance functions C_α and C_ε respectively.

Since we have diagonalized the problem, the optimal quadratic form will not include off–diagonal terms in the sum:

$$\sum_{n=1}^{N} w_n Y_n(t) Y_n(s).$$

Theorem 1. *Let Assumption 1 hold and let $\{Y_n\}$ be defined as in (3). Define the estimates*

$$\hat{C}_\alpha(t,s) = \sum_{n=1}^{N} w_n Y_n(t) Y_n(s) \qquad \textit{and} \qquad \hat{C}_\varepsilon(t,s) = \sum_{n=1}^{N} u_n Y_n(t) Y_n(s).$$

We then have the following:

a) ***Unbiased:*** *If $\sum w_n d_n = 1$ and $\sum w_n = 0$, then $\hat{C}_\alpha$ is unbiased, i.e., $\mathrm{E}[\hat{C}_\alpha] = C_\alpha$. If $\sum u_n d_n = 0$ and $\sum u_n = 1$, then $\hat{C}_\varepsilon$ is unbiased.*

b) ***Consistency:*** *If in addition, for $j = 0, 1, 2$*

$$\sum d_n^j w_n^2 \to 0 \quad \textit{and} \quad \sum d_n^j u_n^2 \to 0, \quad \textit{as} \quad N \to \infty$$

then $\hat{C}_\alpha$ *and* $\hat{C}_\varepsilon$ *are consistent in the sense*

$$\mathrm{E}\,\|\hat{C}_\alpha - C_\alpha\|^2 \to 0 \quad \textit{and} \quad \mathrm{E}\,\|\hat{C}_\varepsilon - C_\varepsilon\|^2 \to 0.$$

c) ***Joint Normality:*** *If in addition, there exist positive numbers* γ_1, γ_2, τ_1,τ_2, ζ_1, *and* ζ_2 *such that*

$$\frac{\sum w_n^2 d_n^j}{\sum w_n^2} \to \gamma_j, \quad \frac{\sum u_n^2 d_n^j}{\sum u_n^2} \to \tau_j, \quad \frac{\sum w_n u_n d_n^j}{\sqrt{\sum w_n^2}\sqrt{\sum u_n^2}} \to \zeta_j \quad \textit{for } j = 1, 2,$$

and

$$\frac{\sum w_n^4 (d_n+1)^4}{(\sum w_n^2)^2} \to 0, \quad \textit{and} \quad \frac{\sum u_n^4 (d_n+1)^4}{(\sum u_n^2)^2} \to 0,$$

then

$$\frac{\hat{C}_\alpha - C_\alpha}{\sqrt{\sum w_n^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_\alpha) \quad \textit{and} \quad \frac{\hat{C}_\varepsilon - C_\varepsilon}{\sqrt{\sum u_n^2}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Gamma_\varepsilon)$$

where

$$\begin{aligned}\Gamma_\alpha(t_1,t_2,t_3,t_4) &= \gamma_2(C_\alpha(t_1,t_3)C_\alpha(t_2,t_4) + C_\alpha(t_1,t_4)C_\alpha(t_2,t_3)) \\ &+ \gamma_1(C_\alpha(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\alpha(t_1,t_4)C_\varepsilon(t_2,t_3)) \\ &+ (C_\varepsilon(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\varepsilon(t_1,t_4)C_\varepsilon(t_2,t_3))\end{aligned}$$

and

$$\begin{aligned}\Gamma_\varepsilon(t_1,t_2,t_3,t_4) &= \tau_2(C_\alpha(t_1,t_3)C_\alpha(t_2,t_4) + C_\alpha(t_1,t_4)C_\alpha(t_2,t_3)) \\ &+ \tau_1(C_\alpha(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\alpha(t_1,t_4)C_\varepsilon(t_2,t_3)) \\ &+ (C_\varepsilon(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\varepsilon(t_1,t_4)C_\varepsilon(t_2,t_3)).\end{aligned}$$

Additionally, the asymptotic normality occurs jointly with cross covariance operator

$$\begin{aligned}\Gamma_{\alpha,\varepsilon}(t_1,t_2,t_3,t_4) = \zeta_2(C_\alpha(t_1,t_3)C_\alpha(t_2,t_4) + C_\alpha(t_1,t_4)C_\alpha(t_2,t_3)) \\ + \zeta_1(C_\alpha(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\alpha(t_1,t_4)C_\varepsilon(t_2,t_3)) \\ + (C_\varepsilon(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\varepsilon(t_1,t_4)C_\varepsilon(t_2,t_3)).\end{aligned}$$

The Proof of Theorem 1 can be found in Appendix B. The somewhat complicated looking expressions for Γ_α and Γ_ε are a well known consequence of estimating the variance of a multidimensional object. In the vector case, the variability of a mean estimate is described via covariance matrices, while the variability of a covariance matrix estimate is given in the form of higher order tensors. An analogous problem occurs with the estimation of covariance functions. For further details regarding the use of tensors in multivariate problems see [23].

The weights for the original (untransformed) problem can still be obtained. If, as before, we express $\mathbf{Y}' = \mathbf{U}\mathbf{Y}$, then we have

$$\begin{aligned}\sum w_n Y_n'(t)Y_n'(s) &= \sum w_n U_n^T \mathbf{Y}(t) U_n^T \mathbf{Y}(s) = \mathbf{Y}(t)^T \left(\sum w_n U_n U_n^T\right) \mathbf{Y}(s) \\ &= \mathbf{Y}(t)^T \left(\mathbf{U}\mathbf{W}\mathbf{U}^T\right) \mathbf{Y}(s).\end{aligned}$$

Thus, we can use the $\mathbf{U}$ matrix to obtain the optimal weights on the original scale.

Combining the delta method with Theorem 1, one can also obtain standard errors for the heritability estimates.

Corollary 1. *Under the Assumptions of Theorem 1–c) $\int \hat{C}_\alpha(t,t)\, dt$ is approximately normally distributed with mean $\int C_\alpha(t,t)\, dt$ and variance*

$$\sigma_{11} := \sum w_n^2 \int\int \Gamma_\alpha(t,t,s,s)\, dt\, ds = 2\sum w_n^2(\gamma_2\langle C_\alpha, C_\alpha\rangle + \gamma_1\langle C_\alpha, C_\varepsilon\rangle + \langle C_\varepsilon, C_\varepsilon\rangle).$$

Similarly $\int \hat{C}_\varepsilon(t,t)\, dt$ is approximately normally distributed with mean $\int C_\varepsilon(t,t)\, dt$ and vari-

ance

$$\sigma_{22} := \sum u_n^2 \int\int \Gamma_\varepsilon(t,t,s,s)\ dt\ ds = 2\sum u_n^2(\tau_2\langle C_\alpha, C_\alpha\rangle + \tau_1\langle C_\alpha, C_\varepsilon\rangle + \langle C_\varepsilon, C_\varepsilon\rangle).$$

The asymptotic covariance between $\int \hat{C}_\alpha(t,t)\ dt$ *and* $\int \hat{C}_\varepsilon(t,t)\ dt$ *is given by*

$$\begin{aligned}\sigma_{12} :=& \sqrt{\sum w_n^2}\sqrt{\sum u_n^2}\int\int \Gamma_{\alpha,\varepsilon}(t,t,s,s)\ dt\ ds \\ =& 2\sqrt{\sum w_n^2}\sqrt{\sum u_n^2}(\zeta_2\langle C_\alpha, C_\alpha\rangle + \zeta_1\langle C_\alpha, C_\varepsilon\rangle + \langle C_\varepsilon, C_\varepsilon\rangle).\end{aligned}$$

Let

$$h^T := \left(\frac{\int C_\varepsilon(t,t)\ dt}{\left(\int C_\alpha(t,t)\ dt + \int C_\varepsilon(t,t)\ dt\right)^2}, -\frac{\int C_\alpha(t,t)\ dt}{\left(\int C_\alpha(t,t)\ dt + \int C_\varepsilon(t,t)\ dt\right)^2}\right)$$

and define the 2×2 *matrix* $\Sigma = \{\sigma_{i,j}\}$. *By the multivariate delta method we can conlude*

$$\hat{H} - H \overset{\mathcal{D}}{\approx} N(0, h^T\Sigma h).$$

We conclude the section by examining what the assumptions imply for $\{d_n\}$. Practically, the important take away is that the eigenvalues can't go to zero too fast. The d_n determine how much information is available for estimating C_α, if they go to zero too quickly, then one does not have enough information to consistently estimate the variance components. We give two examples to illustrate this point. For simplicity assume that $C_\alpha = C_\varepsilon$. The optimal weights (arguments given in Proof of Theorem 1 part (d)) are then given by

$$w_n = \frac{a_1 - d_n a_0}{(a_1^2 - a_0 a_2)(d_n+1)^2} \quad \text{and} \quad u_n = \frac{d_n a_1 - a_2}{(a_1^2 - a_0 a_2)(d_n+1)^2}, \tag{4}$$

where

$$a_j = \sum_{n=1}^{N} \frac{d_n^j}{(d_n+1)^2},$$

and we assume that $a_1^2 - a_0 a_2 \neq 0$. That the weights under the scenarios below satisfy our assumptions can be verified directly and thus we omit the arguments for brevity (available upon request).

EXAMPLE 1: Assume that the d_n are obtained by ordering an iid sequence of nonnegative random variables with finite fourth moments. Furthermore, assume that

$$\mathrm{Var}\,(d_n) \neq 0.$$

EXAMPLE 2: Assume that $d_n = n^{-\delta}$ for $0 < \delta < 1/2$.

The point of the above two examples is that the d_n can be a random sequence, even taking the value zero, or it can be a sequence that converges to zero as long as it doesn't converge too quickly, and one will still have consistency of $\hat{C}_\alpha$ and $\hat{C}_\varepsilon$.

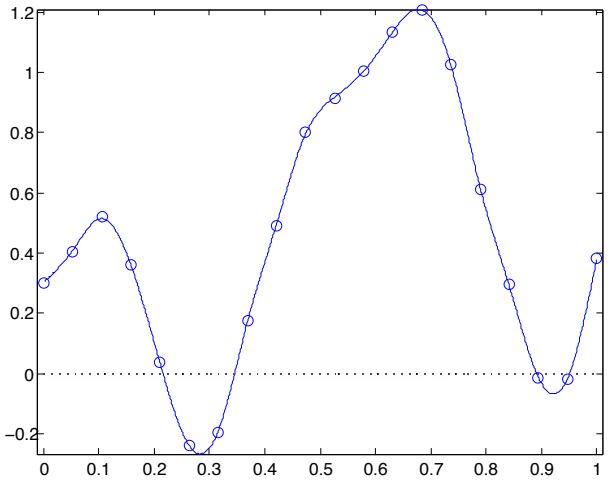

Figure 2: A randomly selected curve for $K_0 = 10/10000$, $H = 0.5$, and $M = 10$. Circles represent the observed value while the curve is the resulting functional object.

5 Simulation Study

We carry out a simulation study to evaluate the behavior of our procedure for varying levels of heritability and numbers of causal SNPs. We also aim to compare our method to a more

traditional FDA approach. In particular, we use the PACE package in Matlab to carry out sparse/longitudinal FPCA. We then estimate the heritability using the resulting scores and the same quadratic form approach as F–MINQUE and MINQUE. The number of PCs is chosen using an explained variance threshold of 95%. This approach reflects what is now a common practice in FDA, namely, to project the data using FPCA and do the analyses on the projections.

In practice, complex phenotypes tend to have a large number of causal SNPs, but we don't expect all or nearly all SNPs to be associated with a trait. Thus, it is important simulate data which has different numbers of causal SNPs among many null SNPs. Intuitively, one expects our method to perform better when there are many causal SNPs as this is more inline with the random effects assumption of our model. To that end, we take the number of causal SNPs, K_0, to be one of four different scenarios: $K_0 = 10, 1000, 10000$ and for the last we take 10 causal SNPs which explain half the heritability and 10000 that explain the other half (for a total of 10010 causal SNPs). In the tables these scenarios are labeled $1, 2, 3$, and 4 respectively. We consider three heritability scenarios, $H = 0.1, 0.25, 0.5$. We perform 1000 repetitions throughout.

To ensure the simulated data has a realistic structure we use the genotypes from the CAMP study in Section 6. We include all subjects so that $N = 540$ and we randomly select the causal SNPs from $K = 186,026$ SNPs which consists of all SNPs genotyped on all subjects with minor allele frequencies greater than 1%. Therefore, in every scenario, the total number of SNPs being included in the mixed effects model is far greater than the sample size ($K \gg N$) or the number of causal SNPs ($K \gg K_0$). We generate the processes according to the model

$$Y_n(t) = \sum_{k=1}^{K_0} G_{nk}\alpha_k(t) + \varepsilon_n(t),$$

where the G_{nk} are the standardized genotypes of the randomly selected causal SNPs (assumed unknown when applying our procedure).

		$M = 10$							
		FS-MINQUE				Sparse FPCA + MINQUE			
		Raw		Truncated		Raw		Truncated	
K_0	H	E[$\hat{H}$]	MSE[$\hat{H}$]	E[$\hat{H}$]	MSE[$\hat{H}$]	E[$\hat{H}$]	MSE[$\hat{H}$]	E[$\hat{H}$]	MSE[$\hat{H}$]
1	0.1	0.102	0.036	0.234	0.031	0.171	0.06	0.266	0.055
2	0.1	0.096	0.034	0.229	0.029	0.173	0.059	0.266	0.052
3	0.1	0.105	0.033	0.234	0.03	0.161	0.058	0.255	0.048
4	0.1	0.105	0.035	0.236	0.031	0.178	0.059	0.269	0.054
1	0.25	0.239	0.05	0.326	0.029	0.309	0.072	0.368	0.055
2	0.25	0.253	0.031	0.324	0.021	0.315	0.06	0.364	0.047
3	0.25	0.247	0.032	0.32	0.021	0.323	0.056	0.37	0.046
4	0.25	0.25	0.038	0.327	0.024	0.325	0.065	0.375	0.052
1	0.5	0.454	0.055	0.483	0.032	0.545	0.085	0.563	0.068
2	0.5	0.495	0.03	0.506	0.02	0.586	0.057	0.592	0.05
3	0.5	0.498	0.028	0.508	0.02	0.585	0.055	0.591	0.049
4	0.5	0.479	0.037	0.495	0.024	0.569	0.058	0.578	0.049
		$M = 20$							
		FS-MINQUE				Sparse FPCA + MINQUE			
		Raw		Truncated		Raw		Truncated	
K_0	H	E[$\hat{H}$]	MSE[$\hat{H}$]	E[$\hat{H}$]	MSE[$\hat{H}$]	E[$\hat{H}$]	MSE[$\hat{H}$]	E[$\hat{H}$]	MSE[$\hat{H}$]
1	0.1	0.101	0.039	0.235	0.032	0.152	0.059	0.262	0.05
2	0.1	0.093	0.036	0.229	0.029	0.167	0.052	0.267	0.048
3	0.1	0.102	0.033	0.234	0.03	0.173	0.052	0.269	0.05
4	0.1	0.1	0.032	0.23	0.028	0.149	0.051	0.256	0.044
1	0.25	0.234	0.049	0.324	0.028	0.314	0.067	0.375	0.053
2	0.25	0.25	0.031	0.324	0.021	0.309	0.05	0.362	0.04
3	0.25	0.247	0.035	0.323	0.022	0.324	0.049	0.372	0.042
4	0.25	0.244	0.036	0.322	0.023	0.317	0.057	0.373	0.045
1	0.5	0.462	0.058	0.489	0.034	0.535	0.077	0.558	0.06
2	0.5	0.502	0.029	0.511	0.02	0.577	0.05	0.583	0.043
3	0.5	0.498	0.027	0.507	0.019	0.596	0.053	0.6	0.047
4	0.5	0.488	0.036	0.503	0.023	0.565	0.059	0.576	0.05

Table 1: Estimated means and mean squared errors for the FS-MINQUE procedure (left two columns) and the more traditional FDA approach consisting of first applying sparse FPCA and then a multivariate version of MINQUE (right two columns, equivalent to REML). We consider three heritability levels, four causal SNP levels, and two levels for the number of sampled time points, M. "Truncated" means setting any negative eigenvalues in the covariance function estimates to zero, while "Raw" leaves them as is. The value K_0 refers to scenarios for the different number of SNPs: 10, 1000, 10000, and 10/10000 respectively. Each scenario is repeated 1000 times.

We generate the $\varepsilon_n(t)$ as Gaussian processes according to the Matérn covariance with parameters $(0, 1, 0, 1/4, 5/2)$ corresponding to the mean, variance, nugget, scale, and smoothness. For the number of time points per curve, we consider $M = 10$ and $M = 20$ equally spaced time points on $[0, 1]$. This ensures the sampling is close the data application we consider later on. The data are converted to functional objects using 100 B–splines and a very small penalty on the second derivative ($\lambda = 10^{-6}$). An example of one such curve is plotted in Figure 2 for $K_0 = 10/10000$, $H = 0.5$, and $M = 20$. The functional object is the solid line while the circles are the observed points.

The α_k are generated in the same way, but scaled by a factor of $H^{1/2}(1-H)^{-1/2}K_0^{-1/2}$, so that the heritability comes out to be H. In the case where we have 10 large signals and 10000 small signals, the two are scaled separately and a $\sqrt{2}$ is also included in the denominator so that the heritability is evenly split between the two types.

Simulation results are summarized in Table 1. We include "Raw" versions in which we do not truncate the negative eigenvalues and a "Truncated" version where we do. The F–MINQUE results seem very robust against the number of causal SNPs. Even when there are only 10 causal SNPs, F–MINQUE seems to work very well. The only exception concerns the raw values with a small number of large effects and a mid range heritability level. There we see that we have a bit more estimation error than in the other settings. We also see a very interesting dynamic occurring when truncating the negative eigenvalues. When truncating, one no longer has an unbiased estimate of the heritability (though this effect is diminished at higher heritability levels). However, the decrease in variance is able to offset the difference, resulting in a lower mean squared error. As the heritability moves closer to 0.5 and the heritability spreads over a large number of SNPs, one has to do less truncating and the two procedures almost converge. We also note that F–MINQUE does not appear to be sensitive to M, performing similarly in both settings.

The sparse method seems to be less accurate. In all cases it overshoots the true heritability level, resulting in a larger bias and MSE as compared to the functional approach.

		$M = 10$				$M = 20$			
		FS-MINQUE		S-FPCA + MINQUE		Functional		S-FPCA + MINQUE	
K_0	H	Raw	Trunc	Raw	Trunc	Raw	Trunc	Raw	Trunc
1	0.1	0.934	0.936	0.879	0.877	0.924	0.933	0.873	0.886
2	0.1	0.949	0.947	0.881	0.887	0.925	0.944	0.884	0.886
3	0.1	0.94	0.943	0.877	0.894	0.945	0.939	0.881	0.883
4	0.1	0.949	0.934	0.873	0.881	0.942	0.948	0.886	0.904
1	0.25	0.889	0.931	0.835	0.865	0.877	0.919	0.828	0.859
2	0.25	0.948	0.957	0.859	0.883	0.946	0.959	0.872	0.885
3	0.25	0.932	0.954	0.867	0.873	0.926	0.952	0.873	0.875
4	0.25	0.931	0.949	0.851	0.869	0.932	0.952	0.864	0.878
1	0.5	0.843	0.924	0.699	0.745	0.827	0.911	0.728	0.781
2	0.5	0.925	0.961	0.786	0.801	0.929	0.96	0.81	0.832
3	0.5	0.942	0.963	0.785	0.796	0.933	0.963	0.785	0.799
4	0.5	0.906	0.956	0.796	0.821	0.91	0.947	0.772	0.804

Table 2: Coverage rates for 95% confidence intervals for the Heritability estimates using the FS-MINQUE procedure (left two columns) and the more traditional FDA approach consisting of first applying sparse FPCA and then a multivariate version of MINQUE (right two columns, equivalent to REML). We consider three heritability levels, four causal SNP levels, and two levels for the number of sampled time points, M. "Truncated" means setting any negative eigenvalues in the covariance function estimates to zero, while "Raw" leaves them as is. The value K_0 refers to scenarios for the different number of SNPs: 10, 1000, 10000, and 10/10000 respectively. Each scenario is repeated 1000 times.

Interestingly, increasing M does not seem to help improve its performance. At first we thought this might be due to only having an explained variance threshold of 95%, but upping this to 99% made no difference. What seems to be happening is that the sparse method smooths out too much variability, resulting in inflated heritability levels. Likely this could be remedied by a more careful approach, but the method described here likely happens a great deal in practice: one uses FPCA to reduce the dimension and then carries out all analyses on the projections. This simulation illustrates how such an approach can go wrong.

As a final comparison, we present coverage rates for our confidence intervals in Table 2. The results suggest that while our coverage rates are good, they do a bit worse when one has a small number of large effects ($K_0 = 1$ or 4). However, this problem is alleviated by working with the truncated versions. This is maybe surprising given the increase in bias

for the truncated values. A possible explanation is that one gets better standard errors by truncating the negative eigenvalues since the estimated covariance function is also used to estimate standard errors. The sparse method has poor coverage due to its inherent bias. In light of the results summarized in Tables 1 and 2 we generally recommend working with the truncated estimates.

6 Results

The Childhood Asthma Management Project, CAMP, was a longitudinal study examining the long term effects of two daily asthma medications, Budesonide and Necrodmil, on children. The study consisted of 813 children making 16 clinical visits over the course of four years. The first three visits were only a week or so apart, but subsequent visits were more spread, extending up to a little over four months. Pulmonary function tests are routinely used for assessing asthma status, and quantifying the genetic contribution to them could lead to a better understanding of the genetic architecture of asthma. Our goal is to estimate the heritability of one of these measures of lung function. To that end, we take the functional response variable in our linear model to be log(FEV1), the volume of air one can expel out of their lungs in one second (on the log scale). We examine a subset of that study consisting of 540 Caucasian children ages 5–12 (at the beginning of the study) at 16 time points, common amongst all subjects, spread over 4 years. We include fixed effects for age (at beginning of study), gender, and treatment. Further details are available in [36, 37], and the data is available on dbGaP study accession number phs000166.v2.p1. The efficacy of the FDA approach to this data was first explored in [30], who found many nonlinear patterns using functional regression. We refer readers to that work for more details on analyzing this data using FDA tools.

Instead of using the matrix of genotypes $\mathbf{G}$, we use a slight modification [40] and work

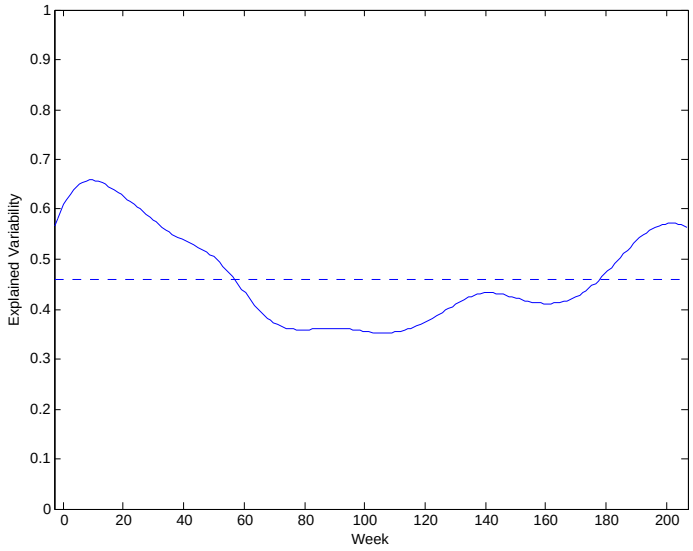

Figure 3: Heritability of $\log(FEV1)$ plotted over time. Overall heritability is 46% and plotted as a dashed line.

directly with the resulting $\mathbf{H}$ matrix. In particular we take

$$\mathbf{H}(n,n') = \begin{cases} \frac{1}{K}\sum_{k=1}^{K} \frac{(G_{nk}-2\hat{p}_k)(G_{n'k}-2\hat{p}_k)}{2\hat{p}_k(1-\hat{p}_k)} & n \neq n' \\ 1 + \frac{1}{K}\sum_{k=1}^{K} \frac{G_{nk}^2-(1+2\hat{p}_k)G_{nk}+2\hat{p}_k^2}{2\hat{p}_k(1-\hat{p}_k)} & n = n' \end{cases},$$

where $\hat{p}_k = \frac{1}{2N}\sum_{n=1}^{N} G_{nk}$ is the allele frequency for SNP k. Here $\mathbf{H}$ is nearly the "covariance" between individuals when using the standardized genotypes, but the diagonal has been modified to get an unbiased estimate of the inbreeding coefficient [40].

To estimate the heritability we only include those SNPs which are measured on all subjects and whose minor allele frequency is above 1%, resulting in 186,026 SNPs. The eigenvalues of the resulting $\mathbf{H}$ matrix decreased very slowly, implying that the results of Theorem 1 hold reasonably well. We also use the truncated version of F–MINQUE. We find that about 46% of the variation averaged over time in the log of FEV1 is attributable to common SNPs. Unfortunately, the standard error of the estimate is 28%, meaning that the

heritability estimates are still very noisy, and additional data is needed for more accurate estimation. Standard errors of similar orders were observed in our simulations (not shown here). In Figure 3 we plot the estimated heritability

$$\frac{\hat{C}_\alpha(t,t)}{\hat{C}_\alpha(t,t)+\hat{C}_\varepsilon(t,t)},$$

over time with a solid line and a dashed line indicating our overall heritability estimate. As we can see, the heritability seems to dip down in the middle of the study. It is not entirely clear why this would happen, but one possibility could be growth spurts. Given the age ranges of the children, it is possible that more of the children are growing rapidly in the middle of the study which might dampen the heritability estimate as the children would look "more independent." However, we emphasize that given the size of the standard error for the Heritability estimate, all interpretations should be taken loosely.

7 Conclusions

We have presented a framework based on functional data analysis for defining and estimating the heritability of a longitudinally measured human trait. Longitudinal data presents an important but challenging problem for genome wide association studies due to the added complexity of the data. A framework based on FDA allows for very flexible tools that exploit the smooth development of the underlying phenotypes. Our methods work well even for sparse longitudinal settings as long as the design is relatively balanced, and in fact, we outperform sparse FPCA approaches which tend to throw out too much variability. We have used our procedure to analyze data coming from CAMP, which results in a heritability estimate of around 46%. Due to the nature of the phenotype (log(FEV1)) and the relatively small sample (for GWAS), the estimate is still fairly noisy.

From an FDA perspective, we have provided important tools for estimating variance components in functional linear models. While our motivation was the mixed effects scheme used

for estimating heritability, our methods apply much more generally to any functional linear model whose covariance can be expressed as a sum of known matrices times unknown covariance functions. Another major application of our tools would be to family based studies (not necessarily genetic). There one would include random effects based on familial relationships. In the context of our model, the $\mathbf{H}_1$ would be a block diagonal matrix representing families and $\mathbf{H}_2$ would be the error of the model. Our methods produce optimal, unbiased estimates, and don't require working with densities or likelihoods. In the univariate case, our approach becomes equivalent to REML. Our estimation scheme, F-MINQUE, extends the classical MINQUE procedure for estimating variance components to functional data. In the case of two components, one of which is the iid error, we demonstrate how consistency and asymptotic normality depend on the behavior of the eigenvalues of the first variance component. It would be useful to establish these properties for the more general multi-component settings, though the conditions would likely be less intuitive.

There are a few areas where our methods could be improved/extended. Firstly, our methods build upon curve by curve smoothing, which works well for balanced data applications with a relatively large number of temporal observations like CAMP. However, for applications with very badly balanced designs and with small numbers of observations, our methods would need to be adjusted. In particular, subject level smoothing is often counter productive as it can often involve smoothing/interpolating temporal regions with no observations. Thus our methodology would benefit from generalizing to the *sparse functional data* case [42], where subjects are often pooled before applying any smoother. Secondly, our asymptotic results are based on quadratic forms of functional objects. However, we don't take into account the random nature of the matrix which arises from using an iterative procedure. Our methods guarantee that the estimators are consistent and asymptotically normal (since the estimates have this property with no iterating), but the standard errors maybe off. Our simulations haven't suggested that this to be a noticeable problem, but it still an area for improvement.

Lastly, an important extension of our procedure would be to more general functional linear models with integral operators and functional/time–varying covariates. Another important development would be for generalized functional linear models [25, 21] which would allow for non–continuous phenotypes which are commonly found in clinical trials.

Acknowledgements

We are grateful to the two anonymous referees and associate editor for their careful comments which has lead to a much improved paper. We are also grateful to Sonja Greven, Jefferey Morris, and Fabian Scheipl for their input on a number of issues concerning this work.

References

[1] A. Antoniadis and T. Sapatinas. Estimation and inference in functional mixed–effects models. *Computational Statistics and Data Analysis*, 51:4793–4813, 2007.

[2] J. A. D. Aston, J.-M. Chiou, and J. P. Evans. Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society, Series C*, 68:837–857, 2006.

[3] D. Bosq. *Linear Processes in Function Spaces*. Springer, New York, 2000.

[4] K. Brown. Asymptotic behavior of MINQUE–type estimators of variance components. *The Annals of Statistics*, 4:746–754, 1976.

[5] K. Chen and H. Müller. Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society, Series B*, 74:67–89, 2012.

[6] X. Chen and H. White. Central limit and functional central limit theorems for Hilbert–valued dependent heterogeneous arrays with applications. *Econometric Theory*, 14:260–284, 1998.

[7] S. Costafreda, G. Barker, and M. Brammer. Bayesian wavelet-based analysis of functional magnetic resonance time series. *Magnetic Resonance Imaging*, 27(4):460–469, 2009.

[8] E. Gamazon, H. Im, C. Liu, D. Nicolae, and N. Cox. The convergence of eQTL mapping, heritability estimation and polygenic modeling: Emerging spectrum of risk variation in bipolar disorder. *arXiv*, arXiv:1303.6227, 2013.

[9] O. Gromenko and P. Kokoszka. Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. *Computational Statistics and Data Analysis*, 59:82–94, 2013.

[10] O. Gromenko, P. Kokoszka, L. Zhu, and J. Sojka. Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends. *The Annals of Applied Statistics*, 6:669–696, 2012.

[11] W. Guo. Functional mixed effects models. *Biometrics*, 58:121–128, 2002.

[12] L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

[13] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

[14] A. Ivanescu, A. Staicu, F. Scheipl, and S. Greven. Penalized function-on-function regression. *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 254*, 2013.

[15] J. Y. J, S. Lee, M. Goddard, and P. Visscher. GCTA: a tool for genome–wide complex trait analysis. *American Journal of Human Genetics*, 88:76–82, 2011.

[16] P. Kokoszka and M. Reimherr. Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, 34:116–129, 2012.

[17] P. Kokoszka and M. Reimherr. Predictability of shapes of intraday price curves. *The Econometrics Journal*, Forthcoming, 2013.

[18] E. Kostem and E. Eskin. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *American Journal of Human Genetics*, 92:558–564, 2013.

[19] L. LaMotte. On non–negative quadratic unbiased estimation of variance components. *Journal of the American Statistical Association*, 68:728–730, 1973.

[20] L. LaMotte. Quadratic estimation of variance components. *Biometrics*, 29:311–330, 1973.

[21] Y. Li, N. Wang, and R. J. Carroll. Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, 105(490):621–633, 2010.

[22] L. Luo, Y. Zhu, and M. Xiong. Quantitative trait locus analysis for next–generation sequencing with the functional linear models. *Journal of Medical Genetics*, 49(8):513–524, 2012.

[23] P. McCullagh. *Tensor Methods in Statistics*. Chapman & Hall, 1987.

[24] J. Morris and R. J. Carroll. Wavelet–based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68:179–199, 2006.

[25] H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *The Annals of Statistics*, 33:774–805, 2005.

[26] C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. *SINUM*, 12:617–629, 1975.

[27] J. O. Ramsay, N. Altman, and R. D. Bock. Variation in height acceleration in the fels growth data. *The Canadian Journal of Statistics*, 22(1):89–102, 1994.

[28] C. Rao. Estimation of variance and covariance components–MINQUE theory. *Journal of Multivariate Analysis*, 1:257–275, 1971.

[29] C. Rao. Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis*, 1:445–456, 1971.

[30] M. Reimherr and D. Nicolae. A functional data analysis approach for genetic association studies. *Annals of Applied Statistics*, 8:406–429, 2014.

[31] P. Reiss, M. Mennes, E. Petkova, L. Huang, M. Hoptman, B. Biswal, S. Colcombe, X.-N. Zuo, and M. Milham. Extracting information from functional connectivity maps via function-on-scalar regression. *NeuroImage*, 56:140–148, 2011.

[32] F. Scheipl, A.-M. Staicu, and S. Greven. Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 2014.

[33] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. Wiley, New York, 1992.

[34] M. V. Stefanescu, F. Serban, and S. Dedu. Portfolio optimization using data analysis techniques. In *Proceedings of the 12th WSEAS international conference on Mathematical methods, computational techniques and intelligent systems*, pages 146–151, Stevens Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS).

[35] R. Tang and H. Müller. Time-synchronized clustering of gene expression trajectories. *Biostatistics*, 10:32–45, 2009.

[36] The Childhood Asthma Management Program Research Group. The childhood asthma management program (CAMP): design, rationale, and methods. *Controlled Clinical Trials*, 20:91–120, 1999.

[37] The Childhood Asthma Management Program Research Group. Long–term effects of budesonide or nedocromil in children with asthma. *New England Journal of Medicine*, 343:1054–1063, 2000.

[38] S. Vattikuti, J. Guo, and C. C. Chow. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genetics*, 8(3):e1002637, 2012.

[39] N. Verzelen, W. Tao, and H.-G. Müller. Inferring stochastic dynamics from functional data. *Biometrika*, 99(3):533–550, 2012.

[40] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42:565–569, 2010.

[41] J. Yang, T. Manolio, L. Pasquale, E. B. E, N. Caporaso, J. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. Hayes, W. Hill, M. Landi, A. Alonso, G. Lettre, P. Lin, H. L. H, W. Lowe, R. Mathias, M. Melbye, E. Pugh, M. Cornelis, B. Weir, M. Goddard, and P. Visscher. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics*, 43:519–525, 2011.

[42] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590, 2005.

[43] X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genetics*, 9(2):e1003264, 2013.

[44] X. Zhou and M. Stephens. Genome-wide efficient mixed–model analysis for association studies. *Nature Genetics*, 44:821–826, 2012.

[45] V. Zipunnikov, B. Caffo, D. Yousem, C. Davatzikos, B. Schwartz, and C. Crainiceanu. Functional principal component model for high-dimensional brain imaging. *NeuroImage*, 58:772–784, 2011.

Appendix A: F-MINQUE Derivation

We begin by taking the expected value of the $\hat{C}_i(t,s)$ estimate given in Section 3:

$$\begin{aligned}\mathrm{E}[\mathbf{Y}(t)^T\mathbf{A}_i\mathbf{Y}(s)] &= \mathrm{E}[\mathrm{trace}(\mathbf{Y}(t)^T\mathbf{A}_i\mathbf{Y}(s))] = \mathrm{trace}(\mathbf{A}_i\,\mathrm{E}[\mathbf{Y}(s)\mathbf{Y}(t)^T]) \\ &= \mathrm{trace}\left(\mathbf{A}_i\left(\sum_{i_1=1}^{I}\mathbf{H}_{i_1}C_{i_1}(t,s)\right)\right) \\ &= \sum_{i_1=1}^{I}\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1})C_{i_1}(t,s).\end{aligned}$$

So, to obtain an unbiased estimate, we require

$$\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}) = 1_{i=i_1}.$$

To compute the variance, notice that we can create the larger quadratic form

$$\mathrm{Var}[\mathbf{Y}(t)^T\mathbf{A}_i\mathbf{Y}(s)] = \mathrm{Var}\left[\begin{pmatrix}\mathbf{Y}(t)\\ \mathbf{Y}(s)\end{pmatrix}^T\begin{pmatrix}\mathbf{0} & \mathbf{A}_i\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{Y}(t)\\ \mathbf{Y}(s)\end{pmatrix}\right].$$

Since $\mathbf{Y}$ is Gaussian, we have that

$$\mathrm{Var}[\mathbf{Y}(t)^T\mathbf{A}_i\mathbf{Y}(t)] = 2\,\mathrm{trace}\left[\left(\begin{pmatrix}\mathbf{0} & \mathbf{A}_i\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathrm{E}[\mathbf{Y}(t)\mathbf{Y}(t)^T] & \mathrm{E}[\mathbf{Y}(t)\mathbf{Y}(s)^T]\\ \mathrm{E}[\mathbf{Y}(s)\mathbf{Y}(t)^T] & \mathrm{E}[\mathbf{Y}(s)\mathbf{Y}(s)^T]\end{pmatrix}\right)^2\right].$$

Carrying out the matrix multiplication one arrives at

$$\begin{aligned}\mathrm{Var}[\mathbf{Y}(t)^T\mathbf{A}_i\mathbf{Y}(s)] &= 2\,\mathrm{trace}[(\mathbf{A}_i\,\mathrm{E}[\mathbf{Y}(t)\mathbf{Y}(s)^T])^2]\\ &= 2\,\mathrm{trace}\left[\left(\mathbf{A}_i\sum_{i_1=1}^{I}\mathbf{H}_{i_1}C_{i_1}(t,s)\right)^2\right]\\ &= 2\sum_{i_1,i_2}\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2})C_{i_1}(t,s)C_{i_2}(t,s)\end{aligned}$$

So the target function we are attempting to minimize can be expressed as

$$2\sum_{i_1,i_2}\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2})c_{i_1,i_2},$$

where $c_{i_1,i_2} = \int\int C_{i_1}(t,s)C_{i_2}(t,s)\,dt\,ds$, and subject to the constraints

$$\mathbf{X}^T\mathbf{A}_i\mathbf{X} = 0 \qquad \text{and} \qquad \mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}) = 1_{i=i_1}.$$

Since the data can always be transformed so that $\mathbf{X}^T\mathbf{A}_i\mathbf{X} = 0$, we will assume, without loss of generality, that $\mathbf{X} = 0$. Properties of matrix calculus give that

$$\begin{aligned}\frac{\partial}{\partial\mathbf{A}_i}\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2}) &= 2\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2}\\ \frac{\partial}{\partial\mathbf{A}_i}\mathrm{trace}[\mathbf{A}_i\mathbf{H}_{i_1}] &= \mathbf{H}_{i_1}.\end{aligned}$$

The Lagrangian is given by

$$2\sum_{i_1,i_2}\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2})c_{i_1,i_2} + \sum_{i_1=1}^{I}\lambda_{i_1}(\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1}) - 1_{i=i_1})$$

Taking partials with respect to $\mathbf{A}_i$ and setting equal to zero we obtain

$$4\sum_{i_1,i_2}c_{i_1,i_2}\mathbf{H}_{i_1}\mathbf{A}_i\mathbf{H}_{i_2} + \sum_{i_1=1}^{I}\lambda_{i_1}\mathbf{H}_{i_1} = 0.$$

Vectorizing the problem (stacking the columns) we have that

$$4\sum_{i_1,i_2} c_{i_1,i_2}(\mathbf{H}_{i_1}\otimes\mathbf{H}_{i_2})\,\mathrm{vec}(\mathbf{A}_i)+\sum_{i_1=1}^{I}\lambda_{i_1}\,\mathrm{vec}(\mathbf{H}_{i_1})=0.$$

Solving for $\mathbf{A}_i$ we obtain

$$\mathrm{vec}(\mathbf{A}_i)=-\frac{1}{4}\left(\sum_{i_1,i_2} c_{i_1,i_2}(\mathbf{H}_{i_1}\otimes\mathbf{H}_{i_2})\right)^{-1}\left(\sum_{i_1=1}^{I}\lambda_{i_1}\,\mathrm{vec}(\mathbf{H}_{i_1})\right).$$

Along with the constraints

$$\mathrm{trace}(\mathbf{A}_i\mathbf{H}_{i_1})=1_{i=i_1}.$$

Simplifying notation we can express

$$\mathbf{A}_i=-\frac{1}{4}\sum_{i_1=1}^{I}\lambda_{i_1}\mathbf{B}_{i_1}=-\frac{1}{4}(\boldsymbol{\lambda}^T\otimes\mathbf{I})\begin{pmatrix}\mathbf{B}_1\\ \vdots\\ \mathbf{B}_I\end{pmatrix},$$

where

$$\mathbf{B}_{i_1}=\mathrm{vec}^{-1}\left[\left(\sum_{i_1,i_2} c_{i_1,i_2}(\mathbf{H}_{i_1}\otimes\mathbf{H}_{i_2})\right)^{-1}\mathrm{vec}[\mathbf{H}_{i_1}]\right].$$

Turning to the constraints, we obtain the linear equations, for $i_1=1,\ldots,I$,

$$-\frac{1}{4}\sum_{i_2=1}^{I}\lambda_{i_2}\,\mathrm{trace}(\mathbf{B}_{i_2}\mathbf{H}_{i_1})=\delta_{i,i_1}.$$

Define the matrix

$$\mathbf{D}(i_1,i_2)=\mathrm{trace}(\mathbf{H}_{i_1}\mathbf{B}_{i_2}),$$

define the vector $\mathbf{1}_i$ whose coordinates are all zero except the i^{th} which is 1, and let $\boldsymbol{\lambda}$ be the vector of multipliers. Then the system of linear equations can be expressed as

$$-\frac{1}{4}\mathbf{D}\boldsymbol{\lambda}=\mathbf{1}_i.$$

So we can write

$$\boldsymbol{\lambda}=-4\mathbf{D}^{-1}\mathbf{1}_i$$

and we have that

$$\mathbf{A}_i=((\mathbf{D}^{-1}\mathbf{1}_i)^T\otimes\mathbf{I})\begin{pmatrix}\mathbf{B}_1\\ \vdots\\ \mathbf{B}_I\end{pmatrix}.$$

Which can easily be calculated using $\mathbf{H}_i$ and the equations

$$\mathbf{B}_{i_1} = \text{vec}^{-1}\left[\left(\sum_{i_2,i_3} c_{i_2,i_3}(\mathbf{H}_{i_2} \otimes \mathbf{H}_{i_3})\right)^{-1} \text{vec}(\mathbf{H}_{i_1})\right]$$
$$\mathbf{D}(i_1, i_2) = \text{trace}(\mathbf{H}_{i_1}\mathbf{B}_{i_2}).$$

Appendix B: Proof of Theorem 1

A good introduction to the notation and methods used throughout this section can be found in [3]. Note that we define $x \otimes y$ between two elements of a Hilbert space $\mathbb{H}$ to be an operator acting on $\mathbb{H}$, defined as

$$(x \otimes y)(h) = \langle y, h\rangle x,$$

for $h \in \mathbb{H}$. When writing the norm ($\|\cdot\|$) of an operator we will mean this to be the Hilbert–Schmidt norm unless otherwise stated.

Lemma 1 (A Lyapunov condition for Hilbert spaces)**.** *For each $1 \leq N < \infty$, let $X_{1,N}, \ldots, X_{N,N}$ be independent elements of a separable Hilbert space $\mathbb{H}$ with mean zero and $\mathrm{E}\,\|X_{n,N}\|^{2+\delta} < \infty$ for $1 \leq n \leq N$. If*

$$C_N := \sum_{n=1}^{N} \mathrm{E}[X_{n,N} \otimes X_{n,N}] \to C, \tag{5}$$

in the the space of Nuclear operators and

$$\sum_{n=1}^{N} \mathrm{E}\,\|X_{n,N}\|^{2+\delta} \to 0, \tag{6}$$

as $N \to \infty$, then

$$Z_N := \sum_{n=1}^{N} X_{n,N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, C),$$

in $\mathbb{H}$.

Proof. Combining Lemma 3.2 and Remark 3.3(a) from [6], Z_N will be asymptotically Gaussian in $\mathbb{H}$ if, for each $h \in \mathbb{H}$, $\langle Z_N, h\rangle$ is asymptotically Gaussian in $\mathbb{R}$ and

$$C_N \to C,$$

with respect to the Nuclear norm. The first condition simply implies that the finite dimensional projections converge in distribution, while the second condition guarantees tightness. Since the second requirement is assumed, we need only establish convergence of the projections.

Let $h \in \mathbb{H}$ be such that $\langle h, Ch\rangle \neq 0$. Then the Lyapunov condition applied to $\langle h, Z_N\rangle$ requires

$$\frac{\sum_{n=1}^{N} \mathrm{E}[\langle h, X_{n,N}\rangle^{2+\delta}]}{\left(\mathrm{E}\langle h, \sum_{n=1}^{N}(X_{n,N} \otimes X_{n,N})h\rangle\right)^{(2+\delta)/2}} \to 0.$$

Combining our assumptions with the continuous mapping theorem indicates that the denominator converges:

$$\left\langle h, \sum_{n=1}^{N}(X_{n,N} \otimes X_{n,N})h \right\rangle \to \langle h, Ch\rangle \neq 0.$$

We can then bound the numerator using the Cauchy–Schwarz inequality

$$\sum_{n=1}^{N} \mathrm{E}[\langle h, X_{n,N}\rangle^{2+\delta}] \leq \|h\|^{2+\delta} \sum_{n=1}^{N} \mathrm{E}[\|X_{n,N}\|^{2+\delta}],$$

which tends to zero by assumption. Therefore

$$\langle Z_N, h\rangle \to \mathcal{N}(0, \langle h, Ch\rangle),$$

and the claim holds (note that when $\langle h, CH\rangle = 0$ we have that $\langle Z_N, h\rangle \xrightarrow{P} 0$). □

Lemma 2. *Let X be a mean zero Gaussian process in a separable Hilbert space $\mathbb{H}$. If $\mathrm{E}\,\|X\|^2 < \infty$, then*

$$\mathrm{E}\,\|X\|^p \leq \left(\mathrm{E}\,\|X\|^2\right)^{p/2} 2^{p/2} \frac{\Gamma(\frac{p+1}{2})}{\sqrt{\pi}} < \infty$$

for any $0 \leq p < \infty$.

Proof. Let $\{z_k\}$ be iid standard normal random variables and $\{\lambda_k\}$ be the eigenvalues of the covariance operator of X. Then

$$\|X\|^2 \overset{\mathcal{D}}{=} \sum_{k=1}^{\infty} \lambda_k z_k^2.$$

By Jensen's inequality, for $p > 2$,

$$\begin{aligned}
\mathrm{E}\,\|X\|^p = \mathrm{E}[(\|X\|^2)^{p/2}] &= \mathrm{E}\left[\left(\sum_{k=1}^{\infty} \lambda_k z_k^2\right)^{p/2}\right] \\
&= \left(\sum \lambda_k\right)^{p/2} \mathrm{E}\left[\left(\sum_{k=1}^{\infty} \frac{\lambda_k}{\sum \lambda_k} z_k^2\right)^{p/2}\right] \\
&\leq \left(\sum \lambda_k\right)^{p/2} \mathrm{E}\left[\sum_{k=1}^{\infty} \frac{\lambda_k}{\sum \lambda_k} |z_k|^p\right]
\end{aligned}$$

Applying Fubini's theorem we can conclude

$$\mathrm{E}\,\|X\|^p \leq \left(\sum \lambda_k\right)^{p/2} \mathrm{E}[|z_k|^p] = \left(\mathrm{E}\,\|X\|^2\right)^{p/2} 2^{p/2} \frac{\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}.$$

Which completes the proof. □

Lemma 3. *Let* $\mathbf{Y}$ *be a mean zero random element of the* N *fold product space* $L^2(\mathcal{T}) \times \cdots \times L^2(\mathcal{T})$*. Assume that each component of* $\mathbf{Y}$ *(which is a random element of* $L^2(\mathcal{T})$*) has a finite second normed moment. Assume that* $\mathbf{A}$ *and* $\mathbf{B}$ *are two* $N \times N$ *symmetric matrices. Then for elements* $t_1, t_2, t_3,$ *and* t_4 *of* $\mathcal{T}$ *the following holds*

$$\mathrm{E}[\mathbf{Y}(t_1)^T \mathbf{A} \mathbf{Y}(t_2)] = \mathrm{trace}(\mathbf{A}\,\mathrm{E}[\mathbf{Y}(t_1)^T \mathbf{Y}(t_2)])$$

and if $\mathbf{Y}$ *is Gaussian*

$$\begin{aligned}
&\mathrm{Cov}(\mathbf{Y}(t_1)^T \mathbf{A} \mathbf{Y}(t_2), \mathbf{Y}(t_3)^T \mathbf{B} \mathbf{Y}(t_4)) \\
&= \mathrm{trace}(\mathbf{A}\,\mathrm{E}[\mathbf{Y}(t_1)^T \mathbf{Y}(t_3)] \mathbf{B}\,\mathrm{E}[\mathbf{Y}(t_2)^T \mathbf{Y}(t_4)]) \\
&+ \mathrm{trace}(\mathbf{A}\,\mathrm{E}[\mathbf{Y}(t_1)^T \mathbf{Y}(t_4)] \mathbf{B}\,\mathrm{E}[\mathbf{Y}(t_2)^T \mathbf{Y}(t_3)]).
\end{aligned}$$

Proof. It will be convenient to introduce the notation

$$\mathrm{E}[\mathbf{Y}(t_1)^T \mathbf{Y}(t_2)] = \mathbf{\Sigma}(t_1, t_2).$$

The first statement is follows using properties of the trace

$$\begin{aligned}
\mathrm{E}[\mathbf{Y}(t_1)^T \mathbf{A} \mathbf{Y}(t_2)] &= \mathrm{E}[\mathrm{trace}(\mathbf{Y}(t_1)^T \mathbf{A} \mathbf{Y}(t_2))] \\
&= \mathrm{E}[\mathrm{trace}(\mathbf{A} \mathbf{Y}(t_2) \mathbf{Y}(t_1)^T)] \\
&= \mathrm{trace}(\mathbf{A}\,\mathrm{E}[\mathbf{Y}(t_2) \mathbf{Y}(t_1)^T]) = \mathrm{trace}(\mathbf{A} \mathbf{\Sigma}(t_2, t_1)).
\end{aligned}$$

Since covariances are symmetric the first claim holds. Turning to the second claim, it will be useful to fix t_1, t_2, t_3 and t_4. Define the larger vector

$$\mathbf{Z}^T = (\mathbf{Y}(t_1)^T, \mathbf{Y}(t_2)^T, \mathbf{Y}(t_3)^T, \mathbf{Y}(t_4)^T).$$

Then we have

$$\mathbf{Y}(t_1)^T \mathbf{A} \mathbf{Y}(t_2) = \mathbf{Z}^T \begin{pmatrix} \mathbf{0} & 0.5\mathbf{A} & \mathbf{0} & \mathbf{0} \\ 0.5\mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \mathbf{Z},$$

and we can do the same thing with the second quadratic form $\mathbf{Y}(t_3)^T \mathbf{B} \mathbf{Y}(t_4)$. By our

normality assumption the covariance between the two quadratic forms can be computed as

$$\begin{aligned}
&\mathrm{Cov}(\mathbf{Y}(t_1)^T\mathbf{A}\mathbf{Y}(t_2), \mathbf{Y}(t_3)^T\mathbf{B}\mathbf{Y}(t_4)) \\
&= 2\,\mathrm{trace}\left(\begin{pmatrix} \mathbf{0} & 0.5\mathbf{A} & \mathbf{0} & \mathbf{0} \\ 0.5\mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \mathrm{E}[\mathbf{Z}\mathbf{Z}^T] \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.5\mathbf{B} \\ \mathbf{0} & \mathbf{0} & 0.5\mathbf{B} & \mathbf{0} \end{pmatrix} \mathrm{E}[\mathbf{Z}\mathbf{Z}^T]\right).
\end{aligned}$$

And we have

$$\mathrm{E}[\mathbf{Z}\mathbf{Z}^T] = \begin{pmatrix} \boldsymbol{\Sigma}(t_1,t_1) & \dots & \boldsymbol{\Sigma}(t_1,t_4) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}(t_4,t_1) & \dots & \boldsymbol{\Sigma}(t_4,t_4) \end{pmatrix}.$$

Therefore

$$\begin{aligned}
&\begin{pmatrix} \mathbf{0} & 0.5\mathbf{A} & \mathbf{0} & \mathbf{0} \\ 0.5\mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \mathrm{E}[\mathbf{Z}\mathbf{Z}^T] \\
&= \begin{pmatrix} 0.5\mathbf{A}\boldsymbol{\Sigma}(t_1,t_2) & 0.5\mathbf{A}\boldsymbol{\Sigma}(t_2,t_2) & 0.5\mathbf{A}\boldsymbol{\Sigma}(t_2,t_3) & 0.5\mathbf{A}\boldsymbol{\Sigma}(t_2,t_4) \\ 0.5\mathbf{A}\boldsymbol{\Sigma}(t_1,t_1) & 0.5\mathbf{A}\boldsymbol{\Sigma}(t_1,t_2) & 0.5\mathbf{A}\boldsymbol{\Sigma}(t_1,t_3) & 0.5\mathbf{A}\boldsymbol{\Sigma}(t_1,t_4) \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}
\end{aligned}$$

and

$$\begin{aligned}
&\begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0.5\mathbf{B} \\ \mathbf{0} & \mathbf{0} & 0.5\mathbf{B} & \mathbf{0} \end{pmatrix} \mathrm{E}[\mathbf{Z}\mathbf{Z}^T] \\
&= \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0.5\mathbf{B}\boldsymbol{\Sigma}(t_1,t_4) & 0.5\mathbf{B}\boldsymbol{\Sigma}(t_2,t_4) & 0.5\mathbf{B}\boldsymbol{\Sigma}(t_3,t_4) & 0.5\mathbf{B}\boldsymbol{\Sigma}(t_4,t_4) \\ 0.5\mathbf{B}\boldsymbol{\Sigma}(t_1,t_3) & 0.5\mathbf{B}\boldsymbol{\Sigma}(t_2,t_3) & 0.5\mathbf{B}\boldsymbol{\Sigma}(t_3,t_3) & 0.5\mathbf{B}\boldsymbol{\Sigma}(t_3,t_4) \end{pmatrix}
\end{aligned}$$

We then have that

$$\begin{aligned}
&\mathrm{Cov}(\mathbf{Y}(t_1)^T\mathbf{A}\mathbf{Y}(t_2), \mathbf{Y}(t_3)^T\mathbf{B}\mathbf{Y}(t_4)) \\
&= 2[0.25\,\mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_2,t_3)\mathbf{B}\boldsymbol{\Sigma}(t_1,t_4)) + 0.25\,\mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_2,t_4)\mathbf{B}\boldsymbol{\Sigma}(t_1,t_3)) \\
&+ 0.25\,\mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_1,t_3)\mathbf{B}\boldsymbol{\Sigma}(t_2,t_4)) + 0.25\,\mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_1,t_4)\mathbf{B}\boldsymbol{\Sigma}(t_2,t_3))].
\end{aligned}$$

However, since $\mathrm{trace}(\mathbf{X}) = \mathrm{trace}(\mathbf{X}^T)$ and all matrices involved are symmetric we have that

$$\mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_2,t_3)\mathbf{B}\boldsymbol{\Sigma}(t_1,t_4)) = \mathrm{trace}(\boldsymbol{\Sigma}(t_1,t_4)\mathbf{B}\boldsymbol{\Sigma}(t_2,t_3)\mathbf{A}) = \mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_1,t_4)\mathbf{B}\boldsymbol{\Sigma}(t_2,t_3)).$$

So we can conclude that

$$\begin{aligned}&\mathrm{Cov}(\mathbf{Y}(t_1)^T\mathbf{A}\mathbf{Y}(t_2), \mathbf{Y}(t_3)^T\mathbf{B}\mathbf{Y}(t_4))\\&= \mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_1,t_4)\mathbf{B}\boldsymbol{\Sigma}(t_2,t_3)) + \mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma}(t_1,t_3)\mathbf{B}\boldsymbol{\Sigma}(t_2,t_4)).\end{aligned}$$

which proves the second claim. □

Proof of Theorem 1: Since the proof $\hat{C}_\varepsilon$ is nearly the same as Theorem $\hat{C}_\alpha$ we present only the latter.

a) Since $\mathrm{E}[Y_n \otimes Y_n] = d_n C_\alpha + C_\varepsilon$, ensuring unbiasedness means that we need

$$C_\alpha = \mathrm{E}\left[\sum w_n Y_n \otimes Y_n\right] = C_\alpha \sum w_n d_n + C_\varepsilon \sum w_n,$$

or equivalently

$$\sum w_n d_n = 1 \qquad \text{and} \qquad \sum w_n = 0.$$

Examining the given weights in (4) we have that

$$\begin{aligned}\sum w_n d_n &= \sum d_n \frac{a_1 - d_n a_0}{(a_1^2 - a_0 a_2)(d_n+1)^2}\\&= \frac{a_1}{a_1^2 - a_0 a_2}\sum \frac{d_n}{(d_n+1)^2} - \frac{a_0}{a_1^2 - a_0 a_2}\sum \frac{d_n^2}{(d_n+1)^2}\\&= \frac{a_1}{a_1^2 - a_0 a_2} a_1 - \frac{a_0}{a_1^2 - a_0 a_2} a_2 = 1,\end{aligned}$$

and a similar argument shows that $\sum w_n = 0$. Therefore $\hat{C}_\alpha$ using the weights (4) is indeed an unbiased estimate.

b) Assuming we have an unbiased estimate, we can express

$$\begin{aligned}\mathrm{E}\,\|\hat{C}_\alpha - C_\alpha\|^2 &= \mathrm{E}[\langle \hat{C}_\alpha - C_\alpha, \hat{C}_\alpha - C_\alpha\rangle]\\&= \sum_{n,m}^{N} w_n w_m\, \mathrm{E}[\langle Y_n \otimes Y_n - d_n C_\alpha - C_\varepsilon, Y_m \otimes Y_m - d_n C_\alpha - C_\varepsilon\rangle]\end{aligned}$$

Since $\mathrm{E}[Y_n \otimes Y_n] = d_n C_\alpha + C_\varepsilon$ we can expand the inner product to obtain

$$\mathrm{E}\,\|Y_n \otimes Y_n - (d_n C_\alpha + C_\varepsilon)\|^2 = \mathrm{E}\,\|Y_n \otimes Y_n\|^2 + \|d_n C_\alpha + C_\varepsilon\|^2.$$

By definition

$$\begin{aligned}\|Y_n \otimes Y_n\| &= \|(d_n^{1/2}\alpha_n + \varepsilon_n) \otimes (d_n^{1/2}\alpha_n + \varepsilon_n)\|^2\\&= \|d_n \alpha_n \otimes \alpha_n + d_n^{1/2}(\alpha_n \otimes \varepsilon_n + \varepsilon_n \otimes \alpha_n) + \varepsilon_n \otimes \varepsilon_n\|^2.\end{aligned}$$

Relating the norm to the inner product and using its symmetry we have

$$\begin{aligned}\|Y_n \otimes Y_n\|^2 = d_n^2\|\alpha_n \otimes \alpha_n\|^2 + d_n\|\alpha_n \otimes \varepsilon_n + \varepsilon_n \otimes \alpha_n\|^2 + \|\varepsilon_n \otimes \varepsilon_n\|^2 \\ + 2d_n\langle \alpha_n \otimes \alpha_n, \alpha_n \otimes \varepsilon_n + \varepsilon_n \otimes \alpha_n\rangle + 2d_n\langle \alpha_n \otimes \alpha_n, \varepsilon_n \otimes \varepsilon_n\rangle \\ + 2d_n^{1/2}\langle \alpha_n \otimes \varepsilon_n + \varepsilon_n \otimes \alpha_n, \varepsilon \otimes \varepsilon\rangle.\end{aligned}$$

Both $2d_n\langle \alpha_n \otimes \alpha_n, \alpha_n \otimes \varepsilon_n + \varepsilon_n \otimes \alpha_n\rangle$ and $2d_n^{1/2}\langle \alpha_n \otimes \varepsilon_n + \varepsilon_n \otimes \alpha_n, \varepsilon \otimes \varepsilon\rangle$ have mean zero. Expanding the inner products, taking expected values, and gathering like terms yields

$$\mathrm{E}\,\|Y_n \otimes Y_n\|^2 = d_n^2\,\mathrm{E}\,\|\alpha_n\|^4 + 2d_n\,\mathrm{E}\,\|\alpha_n\|^2\,\mathrm{E}\,\|\varepsilon_n\|^2 + 4d_n\langle C_\alpha, C_\varepsilon\rangle + \mathrm{E}\,\|\varepsilon_n\|^4.$$

Similarly we have

$$\|d_nC_\alpha + C_\varepsilon\|^2 = d_n^2\|C_\alpha\|^2 + 2d_n\langle C_\alpha, C_\varepsilon\rangle + \|C_\varepsilon\|^2.$$

So we can express

$$\begin{aligned}\mathrm{E}\,\|\hat{C}_\alpha - C_\alpha\|^2 = (\mathrm{E}\,\|\alpha_1\|^4 - \|C_\alpha\|^2)\sum w_n^2 d_n^2 \\ + 2(\mathrm{E}\,\|\alpha_1\|^2\,\mathrm{E}\,\|\varepsilon_1\|^2 + \langle C_\alpha, C_\varepsilon\rangle)\sum w_n^2 d_n \\ + (\mathrm{E}\,\|\varepsilon_1\|^4 - \|C_\varepsilon\|^2)\sum w_n^2.\end{aligned}$$

Which tends to zero as $N \to \infty$.

c) Since we wish to establish the joint normality of $\hat{C}_\alpha$ and $\hat{C}_\varepsilon$, for this section we need to consider their joint behavior. We can apply Lemma 1 with

$$X_{n,N}^T = (X_{n,N,\alpha}, X_{n,N,\varepsilon}) = \left(\frac{w_n}{\sqrt{\sum w_n^2}}, \frac{u_n}{\sqrt{\sum u_n^2}}\right)(Y_n \otimes Y_n - d_nC_\alpha - C_\varepsilon).$$

The above is a product of an element of $\mathbb{R}^2$ and an element of $L^2(\mathcal{T} \times \mathcal{T})$. Therefore $X_{n,N}$ is an element of the seperable Hilbert space $L^2(\mathcal{T} \times \mathcal{T}) \times L^2(\mathcal{T} \times \mathcal{T})$ equipped with the norm

$$\|(x, y)^T\|^2 = \|x\|^2 + \|y\|^2,$$

for $x, y \in L^2(\mathcal{T} \times \mathcal{T})$.

The covariance of $X_{n,N,\alpha}$ is given by

$$\mathrm{E}[X_{n,N,\alpha} \otimes X_{n,N,\alpha}] = \frac{w_n^2\,\mathrm{E}[(Y_n \otimes Y_n - d_nC_\alpha - C_\varepsilon) \otimes (Y_n \otimes Y_n - d_nC_\alpha - C_\varepsilon)]}{\sum w_n^2}.$$

The numerator can be expressed as

$$\begin{aligned}&\mathrm{E}[(Y_n \otimes Y_n - d_nC_\alpha - C_\varepsilon) \otimes (Y_n \otimes Y_n - d_nC_\alpha - C_\varepsilon)] \\ =&\,\mathrm{E}[Y_n \otimes Y_n \otimes Y_n \otimes Y_n] - (d_nC_\alpha + C_\varepsilon) \otimes (d_nC_\alpha + C_\varepsilon).\end{aligned}$$

Using Isserlis' theorem [13] we have

$$\begin{aligned}
&\mathrm{E}[Y_n(t_1)Y_n(t_2)Y_n(t_3)Y_n(t_4)] \\
=&\,\mathrm{E}[Y_n(t_1)Y_n(t_2)]\,\mathrm{E}[Y_n(t_3)Y_n(t_4)] \\
+&\,\mathrm{E}[Y_n(t_1)Y_n(t_3)]\,\mathrm{E}[Y_n(t_2)Y_n(t_4)] \\
+&\,\mathrm{E}[Y_n(t_1)Y_n(t_4)]\,\mathrm{E}[Y_n(t_2)Y_n(t_3)].
\end{aligned}$$

So we can express

$$\begin{aligned}
&\mathrm{E}[Y_n(t_1)Y_n(t_2)Y_n(t_3)Y_n(t_4)] - (d_nC_\alpha(t_1,t_2) + C_\varepsilon(t_1,t_2))(d_nC_\alpha(t_3,t_4) + C_\varepsilon(t_3,t_4)) \\
=&\,\mathrm{E}[Y_n(t_1)Y_n(t_3)]\,\mathrm{E}[Y_n(t_2)Y_n(t_4)] + \mathrm{E}[Y_n(t_1)Y_n(t_4)]\,\mathrm{E}[Y_n(t_2)Y_n(t_3)] \\
=&(d_nC_\alpha(t_1,t_3) + C_\varepsilon(t_1,t_3))(d_nC_\alpha(t_2,t_4) + C_\varepsilon(t_2,t_4)) \\
&+ (d_nC_\alpha(t_1,t_4) + C_\varepsilon(t_1,t_4))(d_nC_\alpha(t_2,t_3) + C_\varepsilon(t_2,t_3)) \\
=&d_n^2(C_\alpha(t_1,t_3)C_\alpha(t_2,t_4) + C_\alpha(t_1,t_4)C_\alpha(t_2,t_3)) \\
&+ d_n(C_\alpha(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\alpha(t_1,t_4)C_\varepsilon(t_2,t_3)) \\
&+ (C_\varepsilon(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\varepsilon(t_1,t_4)C_\varepsilon(t_2,t_3)).
\end{aligned}$$

So we obtain

$$\begin{aligned}
\sum \mathrm{E}[X_{n,N,\alpha}(t_1,t_2)X_{n,N,\alpha}(t_3,t_4)] \to&\, \gamma_2(C_\alpha(t_1,t_3)C_\alpha(t_2,t_4) + C_\alpha(t_1,t_4)C_\alpha(t_2,t_3)) \\
&+ \gamma_1(C_\alpha(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\alpha(t_1,t_4)C_\varepsilon(t_2,t_3)) \\
&+ (C_\varepsilon(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\varepsilon(t_1,t_4)C_\varepsilon(t_2,t_3)) \\
&:= \Gamma(t_1,t_2,t_3,t_4).
\end{aligned}$$

Note that the above convergence will also hold in the space of Nuclear operators (but writing the final form is a bit more awkward). An identitical argument gives that

$$\begin{aligned}
\sum \mathrm{E}[X_{n,N,\varepsilon}(t_1,t_2)X_{n,N,\varepsilon}(t_3,t_4)] \to&\, \tau_2(C_\alpha(t_1,t_3)C_\alpha(t_2,t_4) + C_\alpha(t_1,t_4)C_\alpha(t_2,t_3)) \\
&+ \tau_1(C_\alpha(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\alpha(t_1,t_4)C_\varepsilon(t_2,t_3)) \\
&+ (C_\varepsilon(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\varepsilon(t_1,t_4)C_\varepsilon(t_2,t_3)) \\
&:= \Gamma(t_1,t_2,t_3,t_4),
\end{aligned}$$

and

$$\begin{aligned}
\sum \mathrm{E}[X_{n,N,\alpha}(t_1,t_2)X_{n,N,\varepsilon}(t_3,t_4)] \to&\, \zeta_2(C_\alpha(t_1,t_3)C_\alpha(t_2,t_4) + C_\alpha(t_1,t_4)C_\alpha(t_2,t_3)) \\
&+ \zeta_1(C_\alpha(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\alpha(t_1,t_4)C_\varepsilon(t_2,t_3)) \\
&+ (C_\varepsilon(t_1,t_3)C_\varepsilon(t_2,t_4) + C_\varepsilon(t_1,t_4)C_\varepsilon(t_2,t_3)) \\
&:= \Gamma(t_1,t_2,t_3,t_4).
\end{aligned}$$

Thus the first requirement for Lemma 1 holds. The second requirement holds by the arguments of part (b). To show the third requirement holds, consider

$$\mathrm{E}\,\|X_{n,N}\|^4 = \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 \mathrm{E}\,\|Y_n \otimes Y_n - d_nC_\alpha - C_\varepsilon\|^4.$$

Applying Lemma 2 we have (taking $\delta = 2$)

$$\begin{aligned}
\mathrm{E}\,\|X_{n,N}\|^4 &\le \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 \mathrm{E}\,\|Y_n\|^8 \\
&\le \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 (\mathrm{E}\,\|Y_n\|^2)^4 2^4 \frac{\Gamma(\frac{9}{2})}{\sqrt{\pi}} \\
&\le \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 (\mathrm{E}\,\|Y_n\|^2)^4 105.
\end{aligned}$$

Summing over n we have

$$\begin{aligned}
&\sum \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 (\mathrm{E}\,\|Y_n\|^2)^4 \\
=&\sum \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 (d_n \,\mathrm{E}\,\|\alpha\|^2 + \mathrm{E}\,\|\varepsilon\|^2)^4 \\
\le& \max\{\mathrm{E}\,\|\alpha\|^2, \mathrm{E}\,\|\varepsilon\|^2\}^4 \sum \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 (d_n + 1)^4.
\end{aligned}$$

Examining the summand we have

$$\begin{aligned}
&\sum \left(\frac{w_n^2}{(\sum w_n^2)} + \frac{u_n^2}{(\sum u_n^2)}\right)^2 (d_n + 1)^4 \\
=&\sum \frac{w_n^4 (d_n+1)^4}{(\sum w_n^2)^2} + \sum \frac{u_n^4 (d_n+1)^4}{(\sum u_n^2)^2} + \sum \frac{w_n^2 u_n^2 (d_n+1)^4}{(\sum w_n^2)(\sum u_n^2)}.
\end{aligned}$$

The first two summands tend to zero by assumption while the third summand, by the Cauchy–Schwarz inequality, is bounded by

$$\sum \frac{w_n^2 u_n^2 (d_n+1)^4}{(\sum w_n^2)(\sum u_n^2)} \le \sqrt{\left(\sum \frac{w_n^4 (d_n+1)^4}{(\sum w_n^2)^2}\right)\left(\sum \frac{u_n^4 (d_n+1)^4}{(\sum u_n^2)^2}\right)} \to 0,$$

by assumption. Therefore $\sum \mathrm{E}\,\|X_{n,N}\|^4 \to 0$, the third condition for Lemma 1 is satisfied, and the claim holds.

d) Examining the condition $a_1^2 - a_0 a_2 \neq 0$, by the Cauchy–Schwarz inequality we have

$$a_1^2 \le a_0 a_2,$$

with equality if and only if the $d_n^2 = a + b(d_n + 1)^2$ for some fixed a and b and for all n. However, such a relationship can hold if and only if the d_n are constant. Thus the assumption is really a statement about the equality of the d_n. Clearly one cannot estimate C_α and C_ε in such a situation.

Finally, we show that in the special case $C_\alpha = C_\varepsilon$, the defined estimate is the best

quadratic unbiased estimate. We have already shown that

$$\mathrm{E}\,\|\hat{C}_\alpha - C_\alpha\|^2 = c_0 \sum w_n^2 + 2c_1 \sum w_n^2 d_n + c_2 \sum w_n^2 d_n^2.$$

Where

$$\begin{aligned} c_0 &= \mathrm{E}\,\|\alpha_1\|^4 - \|C_\alpha\|^2 \\ c_1 &= \mathrm{E}\,\|\alpha_1\|^2\,\mathrm{E}\,\|\varepsilon_1\|^2 - \langle C_\alpha, C_\varepsilon\rangle \\ c_2 &= \mathrm{E}\,\|\varepsilon_1\|^4 - \|C_\varepsilon\|^2. \end{aligned}$$

Thus, unfortunately, the optimal weights will depend on C_α and C_ε. However, if we consider the case where $C_\alpha = C_\varepsilon$ and the functions are Gaussian, then an application of Isserlis's theorem [13] shows the weights satisfy

$$c_0 = c_2 = c_1,$$

and the optimal weights are the ones that minimize

$$\sum w_n^2 + 2\sum w_n^2 d_n + \sum w_n^2 d_n^2.$$

Using the method of Lagrange multipliers we wish to minimize the function

$$\sum w_n^2 + 2\sum w_n^2 d_n + \sum w_n^2 d_n^2 + \gamma_1\left(\sum w_n\right) + \gamma_2\left(\sum w_n d_n - 1\right).$$

The derivative of the target function, with respect to w_n, is

$$2w_n d_n^2 + 4w_n d_n + 2w_n + \gamma_1 + d_n\gamma_2.$$

So we get

$$w_n = \frac{-\gamma_1 - d_n\gamma_2}{2d_n^2 + 4d_n + 2} = \frac{-\gamma_1 - d_n\gamma_2}{2(d_n+1)^2}.$$

From the constraints this gives

$$0 = -\gamma_1 \sum \frac{1}{2(d_n+1)^2} - \gamma_2 \sum \frac{d_n}{2(d_n+1)^2} = \frac{-\gamma_1 a_0 - \gamma_2 a_1}{2}$$

and

$$1 = -\gamma_1 \sum \frac{d_n}{2(d_n+1)^2} - \gamma_2 \sum \frac{d_n^2}{2(d_n+1)^2} = \frac{-\gamma_1 a_1 - \gamma_2 a_2}{2}.$$

The first equation yields

$$\gamma_1 = -\gamma_2 \frac{a_1}{a_0}.$$

Plugging into the second gives

$$2 = \gamma_2 \frac{a_1^2}{a_0} - \gamma_2 a_2,$$

which implies.

$$\gamma_2 = \frac{2}{a_1^2/a_0 - a_2} = \frac{2a_0}{a_1^2 - a_0 a_2} \qquad \text{and} \qquad \gamma_1 = \frac{-2a_1}{a_1^2 - a_0 a_2}.$$

So the optimal weights are given by

$$w_n = \frac{a_1 - d_n a_0}{(a_1^2 - a_0 a_2)(d_n + 1)^2},$$

which completes the proof.