Statistical inference for functions of the covariance matrix in stationary Gaussian vector time series

Ian L. Dryden[*], Alfred Kume[+], Huiling Le[*] and Andrew T.A. Wood[*]
(∗) University of Nottingham
(+) University of Kent

Abstract

We consider inference for functions of the marginal covariance matrix under a general class of stationary multivariate temporal Gaussian models. The main application which motivated this work involves the estimation of configurational entropy from molecular dynamics simulations in computational chemistry, where current methods of entropy estimation involve calculations based on the sample covariance matrix. The class of Gaussian models we consider, referred to as Gaussian Independent Principal Components models, is characterised as follows: the temporal sequence corresponding to each principal component (PC) is permitted to have general (temporal) dependence structure, but sequences corresponding to distinct PCs are assumed independent. In many contexts, this model class has the potential to achieve a good balance between flexibility and tractability: distinct PCs are permitted to have different, and quite general, dependence structures, but, as we shall see, estimation and large-sample inference are quite feasible, even in high-dimensional settings. We derive the limiting large-sample Gaussian distribution for the sample covariance matrix, and also results for functions of the sample covariance matrix, which provide a basis for approximate inference procedures, including confidence calculations for scalar quantities of interest. The results are applied to the molecular dynamics application, and the asymptotic properties of a configurational entropy estimator are given. Rotation and translation are removed by initial Procrustes registration, so that entropy is calculated from the size-and-shape of the configuration. An improved estimator based on maximum likelihood is suggested, and some further applications are also discussed.

Keywords: Autoregressive, Central Limit Theorem, Configurational Entropy, Gaussian, Moments, Principal components, Procrustes, Sample covariance, Shape, Size, Temporal.

1 Introduction

The sample covariance matrix is frequently used for statistical inference even when temporally correlated multivariate observations are available. The main application which motivated this work involves the estimation of configurational entropy from molecular dynamics simulations

in computational chemistry, where current methods of entropy estimation involve calculations based on the sample covariance matrix; see e.g. Schlitter (1993) and Harris et al. (2001). For example, entropy calculations were used by Harris et al. (2001) to explain why a particular DNA molecule binds with two ligands, rather than a single ligand. Other applications include the study of sample principal components analysis of multivariate temporal data (Section 3.4), and size-and-shape analysis of temporally correlated planar data (Section 5.1).

In this paper we develop inference procedures for functions of the covariance matrix under a general class of stationary temporally correlated Gaussian models. Models exhibiting long-range dependence are included in the class, as well as more standard short-range dependence models. These models, which are appropriate for temporally correlated vector observations, are referred to as Gaussian Independent Principal Components models and are characterised as follows: the temporal sequence corresponding to each principal component (PC) is permitted to have general (temporal) dependence structure, if desired different from that of the other PCs, but sequences corresponding to distinct PCs are assumed independent. In many contexts, this model class has the potential to achieve a good balance between flexibility and tractability: distinct PCs are permitted to have different, and quite general, dependence structures, but, as we shall see, estimation and large-sample inference are quite feasible, even in high-dimensional settings.

The plan of the paper is as follows. In Section 2 we define the class of stationary Gaussian Independent Principal Component models. In Section 3.1 we present a central limit theorem for a general function of the sample covariance matrix. This provides a basis for constructing approximate confidence regions for functions of the population covariance matrix. In Section 3.2 we determine the leading bias term which allows us to derive approximate bias-corrected confidence intervals, and in Section 3.3 we briefly consider long-range dependence. In Section 3.4 principal component analysis is discussed when temporal correlations are present. In Section 4 we describe the molecular dynamics application that motivated this work, and we investigate Schlitter's (1993) absolute configurational entropy estimator. We also show how long-range dependence leads to a simple asymptotic power law for the expectation of the entropy estimator. Rotation and translation are removed by initial Procrustes registration, so that entropy is calculated from the size-and-shape of the configuration. We suggest an improved estimator based on maximum likelihood, and compare the estimators in a numerical example. In Section 5 we briefly discuss another application, in planar size-and-shape analysis, and we conclude with a discussion. All proofs are given in the Appendix.

2 The stationary Gaussian IPC model

2.1 Preliminaries

We shall consider the situation where a sequence of p-vectors $X_1, \ldots, X_n$ is available at n time points. For example, the vectors could contain observations at p sites in space or p co-ordinates of

a geometrical object. We write $X_i = (X_i(1), \ldots, X_i(p))^T$, $i = 1, \ldots, n$. It is assumed throughout the paper that the X_i sequence is jointly Gaussian. We also assume stationarity: for any integers $k > 0$, $1 \leq i_1 < \ldots < i_k$ and h,

$$\{X_{i_1}, \ldots, X_{i_k}\} \quad \text{has the same distribution as} \quad \{X_{i_1+h}, \ldots, X_{i_k+h}\}.$$

Consider the marginal mean vector and marginal covariance matrix,

$$E[X_i] = \mu, \quad \text{var}(X_i) = \Sigma_S, \ i = 1, \ldots, n,$$

respectively. We call μ the *spatial mean*, and Σ_S the *spatial covariance matrix*, since in many of our applications the vector measurements are collected in space or on geometrical objects. From the spectral decomposition we have $\Sigma_S = Q\Lambda Q^T$, where the columns of Q are eigenvectors of Σ_S and $\text{diag}(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix containing the corresponding eigenvalues.

2.2 The IPC model

We now specify the Independent Principal Components (IPC) model. The temporal covariance structure between the vectors is specified using the transformed vectors of population PC scores

$$Z_i = (Z_i(1), \ldots, Z_i(p))^T = Q^T(X_i - \mu) \sim N_p(0, \Lambda), \ i = 1, \ldots, n, \tag{1}$$

$$\text{cov}(Z_i(r), Z_j(s)) = \begin{cases} \rho_r(i-j)\lambda_r & r = s \\ 0 & r \neq s \end{cases} \tag{2}$$

where $\rho_r(0) = 1, \ r = 1, \ldots, p$. We write Ξ_r for the $n \times n$ *temporal correlation matrix* of population PC score r, which has (i, j)th entry $\rho_r(i - j)$. Hence under this model the population PC scores are mutually independent but there are possibly different temporal correlation structures for each PC score. In terms of the original measurements we have

$$X = (X_1(1), \ldots, X_n(1), \ldots, X_1(p), \ldots, X_n(p))^T \sim N_{np}(\mu \otimes 1_n, \Omega), \tag{3}$$

where $\Omega = (Q\Lambda^{1/2} \otimes I_n)\text{diag}(\Xi_1, \ldots, \Xi_p)(Q\Lambda^{1/2} \otimes I_n)^T$, 1_n is a n-vector of ones, I_n is the $n \times n$ identity matrix and $\otimes$ denotes the Kronecker product.

Remark 2.1. Our main reason for assuming stationarity in the IPC model is to simplify the exposition. However, the stationarity assumption is not essential for the developments in this paper and can be weakened. The Gaussian assumption is more difficult to relax. In particular, the limit theory and the resulting approximate inference procedures described in Section 3 become much more complex. See, for example, Arcones (1994) for relevant limit theory under one type of departure from the Gaussian assumption.

The sample covariance matrix of the original vectors is

$$\hat{\Sigma}_S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T - \bar{X}\bar{X}^T, \tag{4}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. One of the principal goals of this paper is to establish in the asymptotic properties of the estimator (4) of Σ_S under model (1)–(3).

We shall sometimes find it more convenient to work with the population PC scores $Z_i, i = 1, \ldots, n$, and transforming from Z_i to $X_i = QZ_i + \mu$ (and vice versa) is straightforward.

Under the stationary Gaussian IPC model (1)-(3), the population PC scores have joint distribution

$$Z = (Z_1(1), \ldots, Z_n(1), \ldots, Z_1(p), \ldots, Z_n(p))^T \sim N_{np}(0, \mathrm{diag}(\lambda_1\Xi_1, \ldots, \lambda_p\Xi_p)). \quad (5)$$

and the distribution of the sample mean of the PC scores is

$$\overline{Z}(r) = \frac{1}{n}\sum_{i=1}^{n} Z_i(r) \sim N(0, \gamma_{n,r}) \ , \quad r = 1, \ldots, p, \quad (6)$$

independently, where

$$\gamma_{n,r} = \frac{1}{n^2}\lambda_r \sum_{j=1}^{n}\sum_{l=1}^{n} \rho_r(j-l). \quad (7)$$

The sample covariance matrix of the Z sequence is given by

$$\hat{\Lambda} = Q^T\hat{\Sigma}_S Q = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})(Z_i - \bar{Z})^T = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T - \bar{Z}\bar{Z}^T. \quad (8)$$

An important special case is the separable model where $\rho_r(i-j) = \rho(i-j)$ does not depend on r. We write $\Sigma_T = \Xi_1 = ... = \Xi_p$ for the common temporal correlation matrix in the separable case and so

$$X \sim N_{np}(\mu \otimes 1_n, \Sigma_S \otimes \Sigma_T). \quad (9)$$

2.3 Parameter estimation for the IPC model

Suppose that, for $r = 1, \ldots, p$, the correlation sequence $\rho_r(.)$ depends on a parameter vector θ_r, assumed to be $t_r \times 1$. Write $\theta = (\theta_1^T, \ldots, \theta_p^T)^T$. Then the IPC model has the following parameters: μ, Q, Λ and θ. These have, respectively, p, $p(p-1)/2$, p and $t_0 = \sum_{r=1}^{p} t_r$ component parameters. Thus the model has $\frac{1}{2}p(p+3) + t_0$ parameters, provided there are no no functional dependencies, as we shall assume for the remainder of this section.

It turns out that, under the Gaussian IPC model, the parameters split into $p+2$ blocks which are mutually orthogonal with respect to expected Fisher information: $\{\mu\}$, $\{Q\}$ and $\{\lambda_r, \theta_r\}$, $r = 1, \ldots, p$. This fact greatly simplifies the asymptotic covariance structure of the maximum likelihood estimators of the model parameters. Note that, for given "current" estimates of μ and Q, updating the maximum likelihood estimates of the λ_r and θ_r reduces to p independent optimization procedures, each involving a scalar time series. Moreover, under the stationary

Gaussian assumption, it is reasonable to estimate μ by the sample mean, which is asymptotically efficient. Then an alternating procedure may be used in which we update the estimates of the λ_r and θ_r for fixed Q, and update the estimate of Q for fixed λ_r and θ_r. The more difficult part of this procedure is the updating of Q; an algorithm for doing this is proposed in Section 4.2. A simpler alternative is to estimate Q using the matrix whose columns are eigenvectors of $\hat{\Sigma}_S$ but, although this estimator of Q is consistent, it is not fully efficient. In Section 4.2 we discuss parameter estimation when λ_r and θ_r are the parameters of an AR(2) model for the rth PC, $r = 1, \ldots, p$.

3 Asymptotic results and inference for the sample covariance matrix

We begin by introducing some notation. Let $A = (a_{ij})_{i,j=1}^{p}$ denote a $p \times p$ matrix and write $P = p(p+1)/2$. We denote by $\mathrm{vech}(A)$ the $P \times 1$ vector $(a_{11}, a_{12}, a_{22}, a_{13}, ..., a_{p-1,p}, a_{pp})^T$ consisting of the elements in the upper triangle of A. Note that it is not essential to use this ordering of the elements; any ordering could be used, provided it is used consistently. For a vector or matrix A, we define the Euclidean norm, $||A|| = \{\mathrm{tr}(AA^T)\}^{1/2}$, where $\mathrm{tr}(.)$ denotes the trace of a square matrix. For any random vectors U and V, we define $\mathrm{cov}(U, V) = E(UV^T) - E(U)E(V^T)$, and we use $\mathrm{cov}(U)$ as an abbreviation for $\mathrm{cov}(U, U)$.

3.1 Central Limit Theorem

Our first result is a central limit theorem for the sample covariance matrix under the summability condition (10) on the correlation sequences.

Theorem 3.1. *Suppose that*

$$\sum_{h=-\infty}^{\infty} |\rho_r(h)|^2 < \infty, \quad r = 1, ...p. \tag{10}$$

Then, under the stationary Gaussian IPC model (1)–(3),

$$n^{1/2}\mathrm{vech}(\hat{\Sigma}_S - \Sigma_S) \xrightarrow{D} N_P(0, \mathcal{C}\mathcal{V}\mathcal{C}^T) \tag{11}$$

where $\mathcal{V}$ $(P \times P)$ is a diagonal matrix with diagonal elements

$$\mathcal{V}\{(r,s);(r,s)\} = \begin{cases} 2\sum_{h=-\infty}^{\infty} \lambda_r^2 \rho_r(h)^2 & \text{if } r = s \\ \sum_{h=-\infty}^{\infty} \lambda_r \lambda_s \rho_r(h)\rho_s(h) & \text{if } r < s \end{cases}$$

and $\mathcal{C}$ $(P \times P)$ is a matrix with elements

$$\mathcal{C}\{(r,s);(i,j)\} = \begin{cases} q_{ri}q_{si} & \text{if } 1 \le i = j \le p \\ q_{ri}q_{sj} + q_{rj}q_{si} & \text{if } 1 \le i < j \le p \end{cases}$$

where the q_{ij} are the elements of Q.

Remark 3.1. It is interesting to note that Theorem 3.1 holds even if the $\rho_r(h)$ sequences exhibit long-range dependence, i.e. if some or all of the sums $\sum_{h=-\infty}^{\infty} |\rho_r(h)|$ are infinite. This is because $\hat{\Sigma}_S$ is a quadratic function of the data and therefore has Hermite rank 2. See Arcones (1994) and the references therein for further details of Hermite rank.

By applying the delta method (e.g. Mardia et al., 1979, p. 51) we have the following result for a multivariate function of the sample covariance matrix.

Corollary 3.2 *Let $g = (g_1, \ldots, g_t)^T$ be a multivariate t-dimensional function defined on $\mathbb{R}^P$ which is continuously differentiable at* $\mathrm{vech}(\Sigma_S)$*. Under the conditions of Theorem 3.1,*

$$n^{1/2}\left[g\{\mathrm{vech}(\hat{\Sigma}_S)\} - g\{\mathrm{vech}(\Sigma_S)\}\right] \xrightarrow{D} N_q(0, D\mathcal{V}_\Sigma D^T).$$

where $\mathcal{V}_\Sigma = \mathcal{C}\mathcal{V}\mathcal{C}^T$ and $(D)_{ij} = \partial g_i/\partial(\mathrm{vech}(\Sigma_S))_j$, $i = 1, \ldots, t, j = 1, \ldots, P$.

Remark 3.2. In the context of Corollary 3.2, suppose that $t = 1$, so that g is a real-valued function. Let $\theta = g\{\mathrm{vech}(\Sigma_S)\}$ and $\hat{\theta} = g\{\mathrm{vech}(\hat{\Sigma}_S)\}$, and write $\hat{D}$ for the gradient of g evaluated at $\mathrm{vech}(\hat{\Sigma}_S)$; both of these estimators are consistent under the assumptions of Proposition 3.5. Write $\hat{\mathcal{V}}$ for a consistent estimator of $\mathcal{V}$ based on consistent estimators $\hat{\lambda}_r$ of λ_r and $\hat{\rho}_r(.)$ of $\rho_r(.)$; and, finally, let $\hat{\mathcal{C}}$ denote a consistent estimator of $\mathcal{C}$ based on a consistent estimator $\hat{Q}$ of Q. Then an approximate 95% confidence interval for θ is given by

$$(\hat{\theta} - 1.96n^{-1/2}\hat{\sigma}_{\hat{\theta}}, \hat{\theta} + 1.96n^{-1/2}\hat{\sigma}_{\hat{\theta}}), \quad \text{where} \quad \hat{\sigma}^2_{\hat{\theta}} = \hat{D}\hat{\mathcal{C}}\hat{\mathcal{V}}\hat{\mathcal{C}}^T\hat{D}^T \tag{12}$$

is a consistent estimator of $\sigma^2_\theta = D\mathcal{C}\mathcal{V}\mathcal{C}^T D^T$. More will be said on the calculation of $\hat{\sigma}^2_{\hat{\theta}}$ in Section 4.

3.2 Asymptotic Bias

We now consider the leading term in the asymptotic bias of both $\hat{\Sigma}_S$ and a general smooth function $g\{\mathrm{vech}(\hat{\Sigma}_S)\}$ of $\hat{\Sigma}_S$. A stronger summability condition is needed in this case if the bias is to be of order $O(n^{-1})$.

Proposition 3.3 *Suppose that*

$$\sum_{h=-\infty}^{\infty} |\rho_r(h)| < \infty \tag{13}$$

Then under the stationary Gaussian IPC model (1)–(3),

$$E\{\mathrm{vech}(\hat{\Sigma}_S - \Sigma_S)\} = -n^{-1}\mathcal{C}\mathrm{vech}(\Upsilon) + R_n, \tag{14}$$

where $\mathcal{C}$ is the $P \times P$ matrix defined in Theorem 3.1, R_n is a vector remainder term which satisfies $||R_n|| = o(n^{-1})$, and $\Upsilon = \mathrm{diag}(\upsilon_1, \ldots, \upsilon_p)$ has diagonal elements

$$\upsilon_r = \lambda_r \sum_{h=-\infty}^{\infty} \rho_r(h). \tag{15}$$

For sufficiently smooth functions of $\hat{\Sigma}_S$, we have the following result.

Corollary 3.4 *Suppose that $g : \mathbb{R}^P \to \mathbb{R}$ is a function whose second partial derivatives at $\mathrm{vech}(\Sigma_S)$ are all continuous. Let $(a_n)_{n\geq 1}$ be any sequence of positive numbers converging to zero in such a way that*

$$P\left[||\hat{\Sigma}_S - \Sigma_S|| > a_n\right] \to 0 \quad \textit{as} \quad n \to \infty. \tag{16}$$

Then, under the assumptions of Proposition 3.3,

$$E\left[\left\{g\{\mathrm{vech}(\hat{\Sigma}_S)\} - g\{\mathrm{vech}(\Sigma_S)\}\right\} I\left(||\hat{\Sigma}_S - \Sigma_S|| \leq a_n\right)\right] = n^{-1}b_0 + o(n^{-1}) \tag{17}$$

where $I(.)$ denotes the indicator function,

$$b_0 = -\left[\bigtriangledown^T g\{\mathrm{vech}(\Sigma_S)\}\right] \mathcal{C}\,\mathrm{vech}(\Upsilon) + \frac{1}{2}\mathrm{tr}[\mathcal{V}_\Sigma \bigtriangledown \bigtriangledown^T g\{\mathrm{vech}(\Sigma_S)\}], \tag{18}$$

$\mathcal{V}_\Sigma$ is as defined in Corollary 3.2, Υ is the diagonal matrix defined in Proposition 3.3 with elements given by (15), and $\bigtriangledown g\{\mathrm{vech}(\Sigma_S)\}$ and $\bigtriangledown \bigtriangledown^T g\{\mathrm{vech}(\Sigma_S)\}$ are, respectively, the gradient and Hessian of g evaluated at $\mathrm{vech}(\Sigma_S)$.

Remark 3.3. An analogous result holds for multivariate functions $g : \mathbb{R}^P \to \mathbb{R}^t$, $t > 1$.

Remark 3.4. Assumption (16) ensures that $\hat{\Sigma}_S$ lies in a sequence of shrinking neighbourhoods of Σ_S with probability is converging to 1. It is used so that we can avoid having to make moment assumptions about derivatives of g. However, (16) can be avoided if we impose suitable moment assumptions on these derivatives.

A bias-corrected version of the approximate 95% confidence interval (12) is given by

$$(\hat{\theta} - n^{-1}\hat{b}_0 - 1.96n^{-1/2}\hat{\sigma}_{\hat{\theta}}, \hat{\theta} - n^{-1}\hat{b}_0 + 1.96n^{-1/2}\hat{\sigma}_{\hat{\theta}}) \tag{19}$$

where $\hat{b}_0$ is a consistent estimator of b_0 defined in (18) and $\hat{\sigma}^2_{\hat{\theta}}$ is defined in (12).

3.3 Long-range dependence

We now consider long-range dependence. We focus on correlation functions which asymptotically follow a power law. Specifically, we assume in this subsection that

$$\rho_r(h) \sim \beta_r h^{-\alpha} \quad \text{as} \quad h \to \infty, \tag{20}$$

where $\alpha > 0$ does not depend on r. Note that when $\alpha > 1/2$ and $\alpha > 1$, the conditions for Theorem 3.1 and Proposition 3.3, respectively, are satisfied. In this subsection we focus on values of α which give longe-range dependence, i.e. $0 < \alpha \leq 1$. Write $B = \mathrm{diag}(\beta_1, \ldots, \beta_p)$.

Proposition 3.5 *For a covariance function which satisfies (20),*

$$E\left\{\mathrm{vech}(\hat{\Sigma}_S - \Sigma_S)\right\} = \mathcal{C}\,\mathrm{vech}(B\Lambda) \begin{cases} n^{-1}\log n + o(n^{-1}\log n) & \text{if } \alpha = 1 \\ \{(1-\alpha)(2-\alpha)\}^{-1} n^{-\alpha} + o(n^{-\alpha}) & \text{if } \alpha \in (0,1) \end{cases}$$

and

$$\left|\left|\mathrm{cov}\left\{\mathrm{vech}(\hat{\Sigma}_S - \Sigma_S)\right\}\right|\right| = \begin{cases} O(n^{-1}) & \text{if } \alpha > 1/2 \\ O(n^{-1}\log n) & \text{if } \alpha = 1/2 \\ O(n^{-2\alpha}) & \text{if } \alpha \in (0, 1/2). \end{cases}$$

Using a Taylor expansion again, we obtain a similar result for smooth functions of $\hat{\Sigma}_S$.

Corollary 3.6 *Suppose $g : \mathbb{R}^P \to \mathbb{R}$ is a function whose second partial derivatives are all continuous at $\mathrm{vech}(\Sigma)$, and assume that $(a_n)_{n\geq 1}$ is any sequence which converges to zero and satisfies (16). Then, under the conditions of Proposition 3.5,*

$$E\left[\left\{g\{\mathrm{vech}(\hat{\Sigma}_S)\} - g\{\mathrm{vech}(\Sigma_S)\}\right\} I\left(||\hat{\Sigma}_S - \Sigma_S|| \leq a_n\right)\right]$$
$$= 2b_1 \begin{cases} n^{-1}\log n + o(n^{-1}\log n) & \text{if } \alpha = 1 \\ \{(1-\alpha)(2-\alpha)\}^{-1} n^{-\alpha} + o(n^{-\alpha}) & \text{if } 0 < \alpha < 1. \end{cases}$$

where

$$b_1 = -\left[\bigtriangledown^T g\{\mathrm{vech}(\Sigma_S)\right]\, \mathcal{C}\,\mathrm{vech}(B\Lambda). \tag{21}$$

Remark 3.5. We briefly indicate without proof what happens to the limit theory for $\hat{\Sigma}_S - \Sigma_S$ when $0 < \alpha \leq 1/2$. When $\alpha = 1/2$, a central limit theorem holds, with norming factor $(n/\log n)^{1/2}$ rather than $n^{1/2}$. When $0 < \alpha < 1/2$, a so-called non-central-limit theorem holds; see Arcones (1994, Theorem 6) and the references therein for results concerning the limiting distribution theory which arises. Convergence rates under (20) may be determined from Proposition 3.5. An additional complication is that $\bar{Z}\bar{Z}^T$ is no longer negligible when $0 < \alpha < 1/2$.

3.4 Example: Principal components analysis of temporal data

Principal components analysis is often carried out in applications where the observation vectors are temporally correlated. We now discuss relevant asymptotic results under the stationary Gaussian IPC model.

Let $\hat{q}_j$, $j = 1, \ldots, p$, denote the eigenvectors of $\hat{\Sigma}_S$ and write

$$G_n = n^{1/2}(\hat{\Sigma}_S - \Sigma_S).$$

Assume that the eigenvectors of Σ_S are $q_1, \ldots, q_p$ corresponding to eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_p \geq 0$, i.e. the population eigenvalues are assumed distinct. Then, under the assumptions of Theorem 3.1, $\mathrm{vech}(G_n) \xrightarrow{D} \mathrm{vech}(G)$ as $n \to \infty$, where $\mathrm{vech}(G)$ has the Gaussian distribution given by the right-hand side of (11). Moreover, it follows from Corollary 3.2 that, as $n \to \infty$,

$$n^{1/2}(\hat{q}_j - q_j) \xrightarrow{D} \sum_{k \neq l} \frac{q_k^T G q_j}{\lambda_j - \lambda_k} q_k \quad \text{and} \quad n^{1/2}(\hat{\lambda}_j - \lambda_j) \xrightarrow{D} \mathrm{tr}(G q_j q_j^T).$$

The above expressions may be obtained by standard perturbation arguments; see, for example, Watson (1983, Appendix B).

It follows from Corollary 3.4 and Corollary 3.6 that, if the correlations satisfy (20), then

$$E[\hat{\lambda}_j - \lambda_j] = \begin{cases} O(n^{-1}) & \alpha > 1 \\ O(n^{-1}\log n) & \alpha = 1 \\ O(n^{-\alpha}) & 0 < \alpha < 1, \end{cases}$$

and the same order statements hold for $||E(\hat{q}_j - q_j)||$. Corresponding results can be obtained when there are repeated (population) eigenvalues.

We shall see later that the PCs based on maximum likelihood estimation of Σ_S, after identifying the form of the temporal correlation structure, provide an alternative (and improved) method of principal components analysis when the temporal model is correctly specified.

4 Entropy and molecular dynamics simulations

4.1 Asymptotic properties of Schlitter's configurational entropy estimator

Molecular dynamics simulations are a widely-used and powerful method of gaining an understanding of the properties of molecules, particularly biological molecules such as DNA. The simulations are undertaken with a computer package (e.g. AMBER) and involve a deterministic model being specified for the molecule. The model consists of point masses (atoms) connected

by springs (bonds) moving in an environment of water molecules, also treated as point masses and springs. At each time step the equations of motion are solved to provide the next position of the configuration in space. The simulations are very time-consuming to run - for example several weeks of computer time may be needed to generate a few nanoseconds of data.

A major objective of the simulation is the estimation of the configurational entropy of the molecule. The configurational entropy is invariant under location and rotation of the molecule, and the remaining geometrical information is called the 'size-and-shape' of the molecule. Schlitter's (1993) definition of the absolute configurational entropy, based on the covariance matrix of the Cartesian coordinates of atoms calculated by molecular dynamics simulations, is given by

$$S_\infty = \frac{k^*}{2} \log \mid I + c^* M \Sigma_S \mid$$

where M is a diagonal matrix mass, k^* Boltzmann constant, $c^* = k^* T e^2/(2h\pi)^2$, T is the temperature in Kelvin and h is Planck's constant. This formula was derived as an approximation to the configurational entropy where each atom follows a one dimensional quantum-mechanical harmonic oscillator.

Suppose all the atoms have the same atomic mass m and define $c = mk^* T e^2/(2h\pi)^2$. In this case $S_\infty = \frac{k^*}{2} \log \mid I + c\Sigma_S \mid$. An estimate of the entropy is (c.f. Schlitter, 1993)

$$S_n = \frac{k^*}{2} \log \mid I + c\hat{\Sigma}_S \mid,$$

where $\hat{\Sigma}_S$ is the sample covariance matrix given by (4). We have three aims in this subsection: to study the asymptotic behaviour of $E(S_n)$ under the Gaussian model (1)-(3) with correlation function (20), to provide a confidence interval for S_∞, and to suggest a better method of entropy estimation under this model based on maximum likelihood.

In Harris et al. (2001) an empirical observation of the rate at which S_n converges to S_∞ as n tends to infinity is given by the approximation:

$$S_\infty \approx S_n - \frac{a}{n^\alpha}. \tag{22}$$

with $a > 0$. For many simulations of proteins and nucleic acids a value of α between about 0.6 and 0.7 seems reasonable. The following result provides a theoretical basis for (22) under the assumption that long-range dependence is present when $0 < \alpha < 1$.

Theorem 4.1 *For the stationary Gaussian IPC model (1)–(3) with correlation function (20),*

$$E[S_n] = S_\infty - \begin{cases} c_1 n^{-1} + o(n^{-1}) & \alpha > 1 \\ c_2 n^{-1} \log_e n + o(n^{-1} \log_e n) & \alpha = 1 \\ c_3 n^{-\alpha} + o(n^{-\alpha}) & 0 < \alpha < 1. \end{cases},$$

where

$$c_1 = \frac{ck^*}{2} \left\{ \left(\sum_{r=1}^{p} \frac{\upsilon_r}{1 + c\lambda_r} + \right) + \sum_{r,s=1}^{p} \frac{c\mathcal{V}\{(r,s);(r,s)\}}{2(1 + c\lambda_r)(1 + c\lambda_s)} \right\} \tag{23}$$

$$c_2 = ck^* \sum_{r=1}^{p} \frac{\lambda_r \beta_r}{1 + c\lambda_r} \tag{24}$$

$$c_3 = \frac{ck^*}{(1-\alpha)(2-\alpha)} \sum_{r=1}^{p} \frac{\lambda_r \beta_r}{1 + c\lambda_r}, \tag{25}$$

$\mathcal{V}$ is the diagonal matrix defined in the statement of Theorem 3.1, the λ_r are the eigenvalues of Σ_S, and the υ_r are defined in (15).

An approximate confidence interval for S_∞ can be obtained using (12) or its bias-corrected version (19). The relevant partial derivatives are given by $D = \partial S_\infty(\Sigma_S)/\partial \text{vech}(\Sigma_S)$ where

$$\frac{\partial S_\infty(\Sigma_S)}{\partial (\Sigma_S)_{ii}} = \frac{ck^*}{2}(I_p + c\Sigma_S)^{ii} \quad \text{and} \quad \frac{\partial S_\infty(\Sigma_S)}{\partial (\Sigma_S)_{ij}} = ck^*(I_p + c\Sigma_S)^{ij} \quad (i < j).$$

In the above, we have used A^{ij} to denote the elements of A^{-1}.

4.2 Maximum likelihood estimation of entropy

We consider the Gaussian model (1)-(3) and use maximum likelihood to estimate the parameters. The maximum likelihood estimator (m.l.e.) of Σ_S is denoted by $\check{\Sigma}_S$ and the m.l.e. of entropy is

$$\check{S}_\infty = \frac{k^*}{2} \log |I + c\check{\Sigma}_S|.$$

In particular, if $\theta = (S_\infty, \phi^T)^T$, where ϕ is a vector of nuisance parameters, is used to denote the parameters of the distribution and the likelihood function is written as $L(S_\infty, \phi)$, then, as $n \to \infty$,

$$-2 \log \left\{ \sup_{\phi} L(S_\infty, \phi) / \sup_{S_\infty, \phi} L(S_\infty, \phi) \right\} \xrightarrow{D} \chi^2_1,$$

using Wilks' Theorem. The result can be used to obtain confidence intervals based on profile likelihood. However, in practice the constrained maximization over ϕ with S_∞ fixed can be very time-consuming for high-dimensional problems, and we therefore consider an alternative approach.

AR(2) maximum likelihood estimation - separable case

We first consider maximum likelihood estimation in the separable model (9) where the temporal covariance structure is given by a second-order autoregressive [AR(2)] model. For the separable

AR(2) model, the inverse of the temporal correlation matrix is

$$\Sigma_T^{-1} = \sigma_a^{-2} \begin{pmatrix} 1 & -\psi_1 & -\psi_2 & 0 & \dots & 0 & 0 & 0 \\ -\psi_1 & 1+\psi_1^2 & -\psi_1(1-\psi_2) & -\psi_2 & \dots & 0 & 0 & 0 \\ -\psi_2 & -\psi_1(1-\psi_2) & 1+\psi_1^2+\psi_2^2 & -\psi_1(1-\psi_2) & \dots & 0 & 0 & 0 \\ 0 & -\psi_1 & -\psi_1(1-\psi_2) & 1+\psi_1^2+\psi_2^2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -\psi_1(1-\psi_2) & 1+\psi_1^2 & -\psi_1 \\ 0 & 0 & 0 & 0 & \dots & -\psi_2 & -\psi_1 & 1 \end{pmatrix},$$

where $\sigma_a^2 = (1+\psi_2)((1-\psi_2)^2 - \psi_1^2)/(1-\psi_2)$. This matrix is persymmetric (i.e. symmetric about both diagonals) and the determinant is (cf. Siddiqui, 1958) is given by

$$|\Sigma_T^{-1}(\psi_1, \psi_2)| = \sigma_a^{-2n}\{(1-\psi_2^2)^2 - (1+\psi_2)^2\psi_1^2\}.$$

If $\psi_2 = 0$ then this reduces to the AR(1) case. In general, the stationarity conditions for the AR(2) model are

$$\begin{cases} \psi_1 + \psi_2 < 1 \\ \psi_2 - \psi_1 < 1 \\ |\psi_2| < 1 \end{cases}.$$

Let us write Y for the $p \times n$ matrix with the ith column of Y given by X_i, $i = 1, \dots, n$. For given μ, ψ_1 and ψ_2, the density function of Y is

$$f(Y) = \frac{1}{(2\pi)^{pn/2}(\lambda_1\lambda_2 \dots \lambda_p)^{n/2}|\Sigma_T|^{p/2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma_S^{-1} - (Y-\mu 1_n^T)\Sigma_T^{-1}(Y-\mu 1_n^T)^T)\},$$

where $\lambda_1, \lambda_2 \dots \lambda_p$ are the eigenvalues of Σ_S. If μ, ψ_1, ψ_2 are known, then the maximum likelihood estimator (m.l.e.) of Σ_S is

$$\hat{\Sigma}_S(\mu, \psi_1, \psi_2) = \frac{1}{n}(Y-\mu 1_n^T)\Sigma_T^{-1}(Y-\mu 1_n^T)^T, \tag{26}$$

and if the eigenvectors q_r are also known then the m.l.e. of λ_r is

$$\hat{\lambda}_r = \frac{1}{n} q_r^T Y \Sigma_T^{-1} Y^T q_r \quad , \quad r = 1, \dots, p. \tag{27}$$

From (26) we see that $\mathrm{tr}(\Sigma_S^{-1}(Y-\mu 1_n^T)\Sigma_T^{-1}(Y-\mu 1_n^T)^T) = np$. So, modulo a constant,

$$\begin{aligned} l &= -\frac{n}{2}\log\left[\det\{n^{-1}(Y-\mu 1_n^T)\Sigma_T^{-1}(Y-\mu 1_n^T)^T\}\right] + \frac{p}{2}\log\left[\sigma_a^{-2n}((1-\psi_2^2)^2 - (1+\psi_2)^2\psi_1^2)\right], \\ &= -\frac{n}{2}\sum_{r=1}^{p}\log\hat{\lambda}_r + \frac{p}{2}\log\left[\sigma_a^{-2n}((1-\psi_2^2)^2 - (1+\psi_2)^2\psi_1^2)\right], \end{aligned}$$

where $\det$ denotes determinant.

The m.l.e. of μ would be equal to $\bar{X}$ if all rows of Σ_T^{-1} had the same sum. Since n is large and all but four of the row sums of Σ_T^{-1} are equal, the sample mean will be a very good approximation to $\hat{\mu}$. Hence we take $\hat{\mu} \approx \bar{X}$.

AR(2) maximum likelihood estimation - nonseparable case

Let us write Σ_{Tr} for the temporal correlation matrix for the rth score based on an AR(2) model with parameters ψ_{1r}, ψ_{2r}, $r = 1, \ldots, p$.

Given a random sample X_i, $i = 1, \ldots, n$, the joint density function of $Y = [X_1, \ldots, X_n]$ is

$$f(Y) = \frac{1}{(2\pi)^{pn/2}(\lambda_1\lambda_2 \ldots \lambda_p)^{n/2}(\prod_{r=1}^{p} |\Sigma_{Tr}|^{1/2})} \exp\{-\frac{1}{2}\sum_{r=1}^{p} \lambda_r^{-1} q_r^T (Y - \mu 1_n^T)\Sigma_{Tr}^{-1}(Y - \mu 1_n^T)^T q_r\},$$

where $\lambda_1, \ldots \lambda_p$ are marginal variances of the PCs defined by eigenvectors $q_1, \ldots, q_p$. Again we take $\hat{\mu} \approx \bar{X}$.

In order to carry out approximate maximization of the likelihood we consider the following algorithm. Note that the algorithm may not work well in all situations, but it does work well for our situation where there is a strong decay in the eigenvalues.

Approximate MLE computation algorithm

1. Obtain initial estimates of the PC eigenvectors (q_r) from the sample covariance matrix $\hat{\Sigma}_S$ and calculate the PC score vectors $Z_1, \ldots, Z_n$.
2. Estimate $\psi_{1r}, \psi_{2r}, \lambda_r$ based on the AR(2) model for the PC score r, assuming the eigenvectors q_r are fixed.
3. Evaluate $\hat{q}_1$ as the eigenvector of

   $$(Y - \bar{X})\hat{\Sigma}_{T1}^{-1}(Y - \bar{X})^T$$

   corresponding to the smallest positive eigenvalue, where $\hat{\Sigma}_{T1}$ is based on ψ_{11}, ψ_{21}.
4. For $r = 2, 3, \ldots, p$ take $\hat{q}_r$ to be the eigenvector of

   $$P_r(Y - \bar{X})\hat{\Sigma}_{Tr}^{-1}(Y - \bar{X})^T P_r^T$$

   with smallest positive eigenvalue, where $\hat{\Sigma}_{Tr}$ is based on ψ_{1r}, ψ_{2r}, and $P_r = (I_p - \sum_{l=1}^{r-1} \hat{q}_l \hat{q}_l^T)$ is a projection matrix.
5. Estimate $\psi_{1r}, \psi_{2r}, \lambda_r$ based on the AR(2) model for the PC scores, assuming $\hat{Q} = [\hat{q}_1, \ldots, \hat{q}_p]$ is fixed. Note that we do not order the eigenvalues here.
6. Repeat steps 3-5 until convergence or a fixed number of iterations.
7. An approximation for the m.l.e. is the value of the parameters at the highest value of the log-likelihood observed.

The log-likelihood does not necessarily increase at each iteration, but in practice there is usually an increase from the initial starting values in the first few iterations. The algorithm alternates between a) estimating q_r's given the other parameters, and b) estimating the other parameters given the q_r's. The above algorithm effectively explores part of the parameter space near to the sample covariance eigenvectors, which are consistent estimates of the q_r's.

An alternative algorithm that we have experimented with is a Markov chain Monte Carlo algorithm for simulating from a Bayesian model with vague priors, and with simulated annealing. The above approximate MLE algorithm provides better point estimates of the entropy (with higher likelihood) in our implementation.

Finally, we have also explored an algorithm which follows the steepest change in the space of orthogonal matrices $Q = [q_1, \ldots, q_p]$ given the other parameters. One dimensional maximisations are carried out at each iteration. However, in high-dimensional settings, the likelihood increases extremely slowly with this algorithm, and so for practical purposes we consider the approximate MLE algorithm.

4.3 Example: Synthetic dodecamer duplex DNA

We consider the statistical modelling of a specific DNA molecule configuration in water. In particular, we concentrate on the simple case of 22 phosphorus atoms in the synthetic dodecamer duplex DNA which has sequence

1st strand: CTTTTGCAAAAG
2nd strand: GAAAACGTTTTC

The $k = 22$ phosphorous atom locations are recorded in Angstroms (in three dimensions) and are observed over 4 nanoseconds (4×10^{-9}s) with $n = 4000$ observations. Our data are (multivariate) time series in the size-and-shape space $S\Sigma_3^k$ where $k = 22$. For discussion of size-and-shape space, see Dryden and Mardia, 1998, Chapter 8). The observations $X_1, \ldots, X_n$ have been Procrustes rotated to remove rotation and translation (cf. Dryden and Mardia, 1998, Section 5.4.1), and are considered as vectors in $\mathbb{R}^{66}$. Due to the Procrustes registration there are 6 constraints on the data (3 translation and 3 rotation), and so there are $p = 3k - 6$ non-zero eigenvalues of both $\hat{\Sigma}_S$ and Σ_S. Note that the methodology described in subsection 4.2 may be applied directly, even though there are linear constraints in the data: in effect, we simply project the data onto the subspace of dimension p generated by the eigenvectors corresponding to positive eigenvalues.

First of all we calculate the principal components of shape. The PC scores 1-4 are displayed in Figure 1. Note that from Section 3.4 the bias in the eigenvectors is of order $O(n^{-1})$ for $\alpha > 1$, under assumption (20).

Our aim is to estimate the configurational entropy for the DNA using a suitable temporal covariance structure. From the ACF/PACF plots of the PC scores in Figure 2 there are clearly strong correlations present. Note that the autocorrelation structure is somewhat different in each plot.

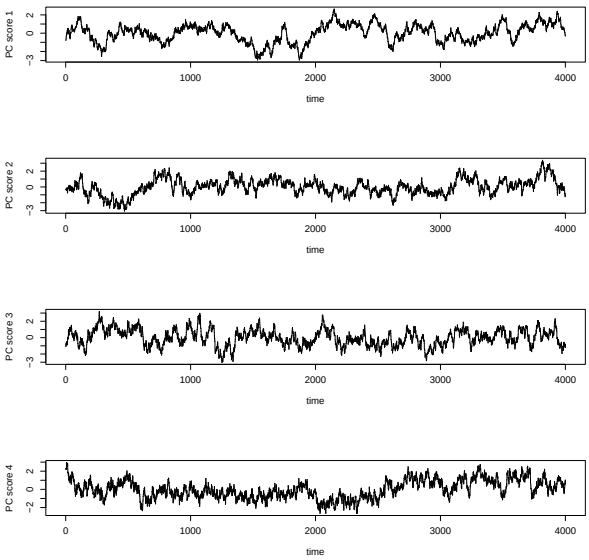

Figure 1: PCs 1-4 for the DNA data.

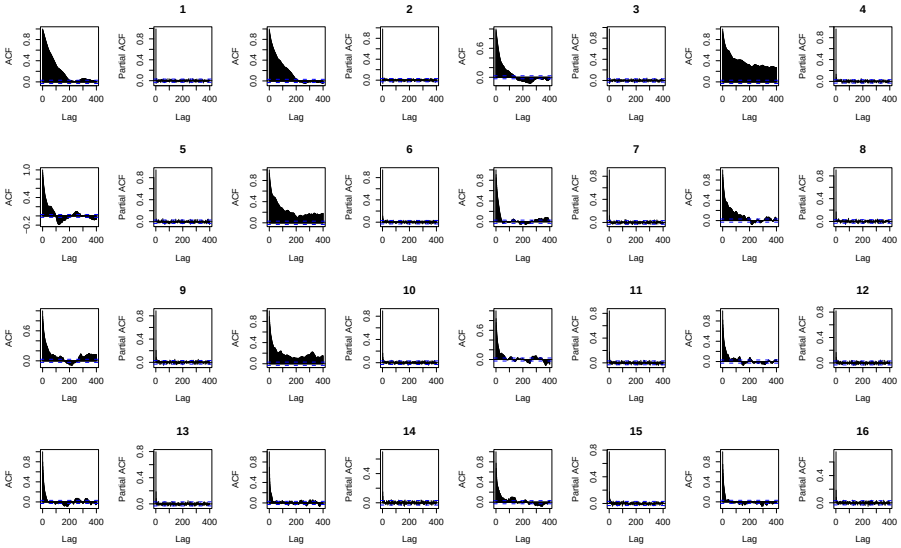

Figure 2: The autocorrelation and partial autocorrelation functions for the first 16 PC scores for the With Water data.

The first few PCs show stronger autocorrelation, but in general an exponential correlation seems reasonable. The partial autocorrelation structure has just a few lags present, perhaps indicating a low order autoregressive model, such as AR(2), might be suitable. We shall consider three models:

I Non-separable stationary Gaussian IPC model with each population PC score following an AR(2) model with parameters $\psi_{1r}, \psi_{2r}, \lambda_r$, $r = 1, \ldots, p$.

II Separable Gaussian model with all components having a common AR(2) model with parameters $\psi_1, \psi_2, \lambda_r$, $r = 1, \ldots, p$.

III Temporal indepedence Gaussian model $\psi_1 = 0, \psi_2 = 0, \lambda_r$, $r = 1, \ldots, p$.

Each model has the additional parameters μ and the p eigenvectors of Σ_S, namely $q_1, \ldots, q_p$, that are used to calculate the PC scores.

We fit the models by maximum likelihood. For models I and II the m.l.e.s are written as $\hat{\mu}$ and spatial covariance m.l.e. is denoted by $\hat{\Sigma}_S^{(AR)}$. The estimated entropy is then given by $S_{AR} = \frac{k^*}{2} \log_e |I_p + c\hat{\Sigma}_S^{(AR)}|$. We write $S_{AR,I}, S_{AR,II}$ for the estimators under models I and II respectively. For model I we approximate the m.l.e.s of the eigenvectors of $\Sigma_S^{(AR)}$ using the algorithm stated at the end of the previous section. For model III the m.l.e. of Σ_S is the sample covariance matrix $\hat{\Sigma}_S$.

In Figure 3 we see the $S_{AR,I}$, $S_{AR,II}$ and S_n estimators obtained from the series of length n (starting from the end of the series) for the datasets. We see that $S_{AR,I}$, $S_{AR,II}$ and S_n increase with n over this time scale. It should be expected that the S_{AR} estimators are larger than S_n, especially for smaller n, as there are strong positive autocorrelations present. A bias corrected estimator S_B (Harris et al.,2001) is also obtained, where the bias is estimated using least squares fits through plots of S_n versus n by fitting the equation (22). The bias corrected estimator is a little larger than $S_{AR,I}$ here.

Note that if the eigenvectors are not estimated by maximum likelihood but rather they are fixed at the sample covariance eigenvectors then the estimation over the remaining parameters leads to estimates of entropy under model I almost identical to S_n (also displayed in Figure 3), and the estimate under model II is almost identical to $S_{AR,II}$.

Under all three models, the m.l.e. of entropy has the same asymptotic properties as the $\alpha > 1$ case. For $n = 4000$ the approximate standard error for $(2/k^*)S_{AR,I}$ obtained under model I is $\hat{\sigma}_\theta = 10.765$ using (12) and the calculations in the Appendix; see (42).

Various standard procedures were used to test for the presence of long-range dependence in the DNA data. Although the findings were not conclusive, it did appear that long-range dependence is not present in these data.

Of course the question remains as to which estimator is to be preferred. Given a long enough simulation the m.l.e. under the correct model and the Schlitter estimator should be approximately

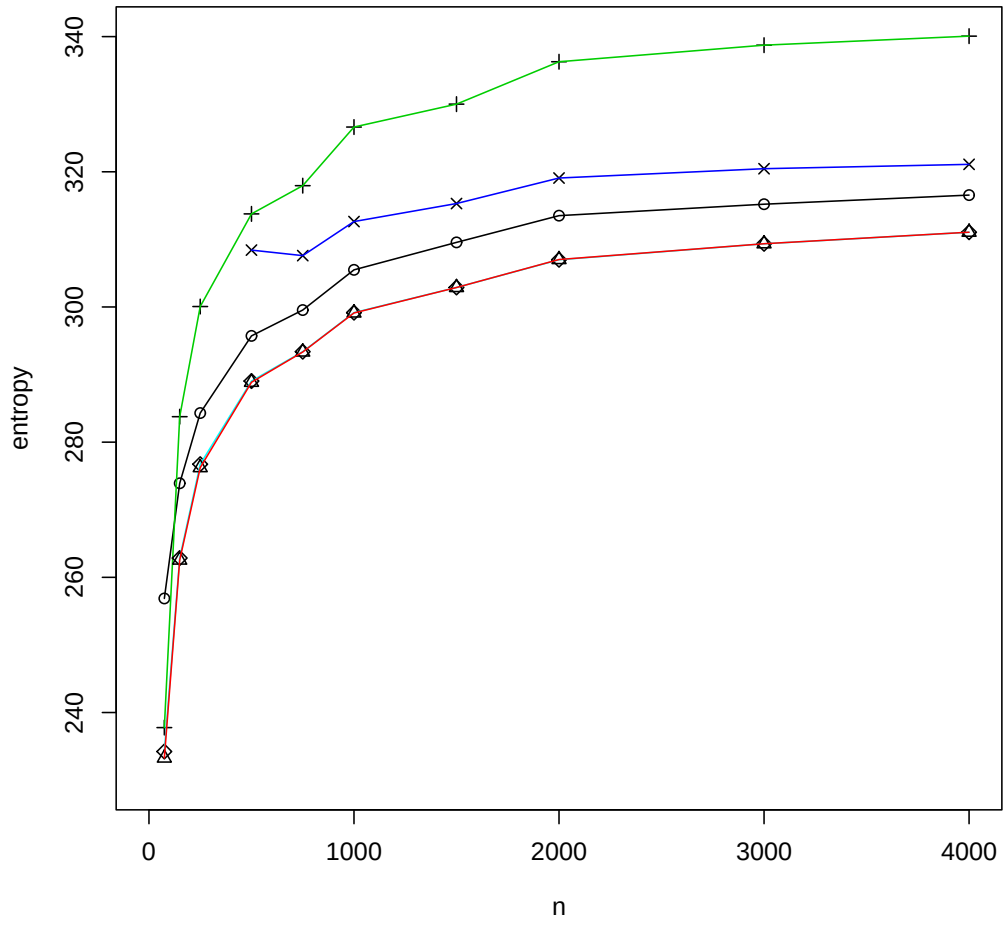

Figure 3: The estimators of entropy versus n. The plots show $(2/k^*)S_{AR,I}$ (circles; black lines), $(2/k^*)S_{AR,II}$ (+; green lines), $(2/k^*)S_n$ (triangles; red lines) and $(2/k^*)S_B$ (x; blue lines). Also, the maximum likelihood estimator under Model I using sample covariance eigenvectors is marked with diamonds, and is very similar to $(2/k^*)S_n$.

unbiased. We carried out a simulation study where data are simulated from a non-separable AR(2) model. The true AR parameters are taken to be the same as those fitted to the scores when using the sample eigenvectors from the DNA dataset. In particular $(2/k^*)S_\infty = 311.06$. In this simulation study Procrustes registration was not carried out. The estimators for samples of size n are given in Figure 4.

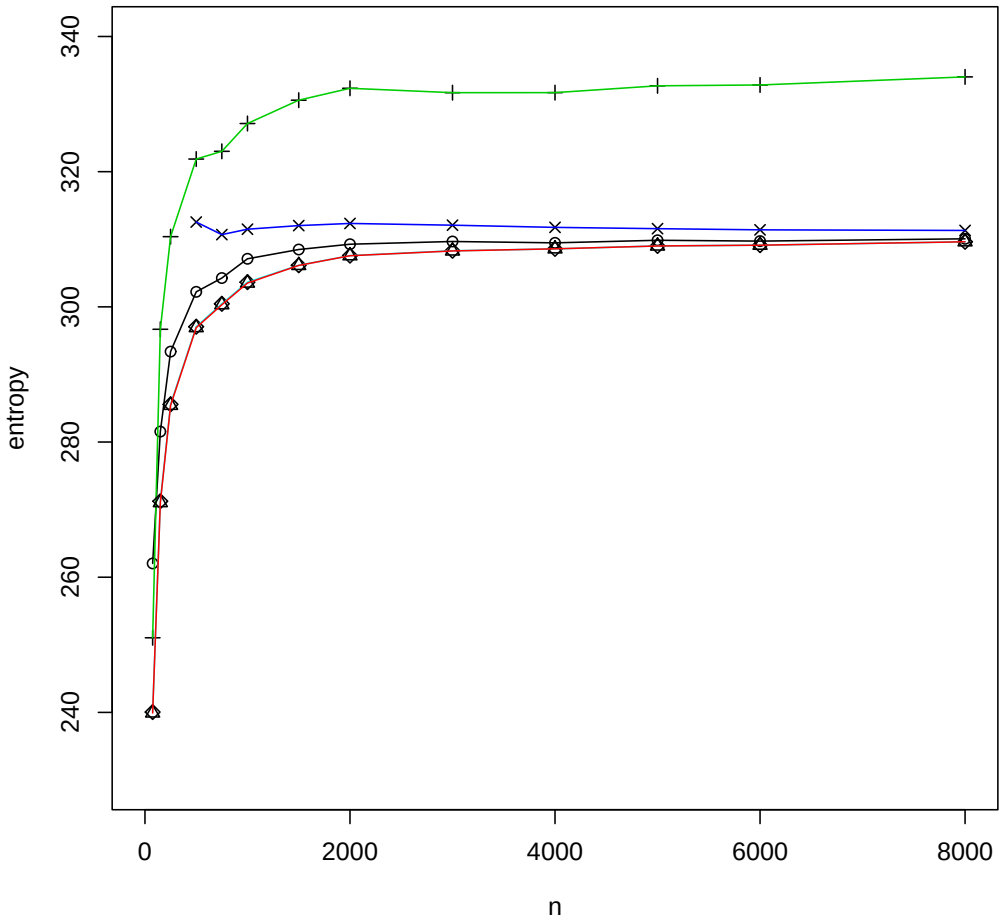

Figure 4: The estimators of entropy versus n for simulated data. The plots show $(2/k^*)S_{AR,I}$ (circles; black lines), $(2/k^*)S_{AR,II}$ (+; green lines), $(2/k^*)S_n$ (triangles; red lines) and $(2/k^*)S_B$ (x; blue lines). Also, the maximum likelihood estimator under Model I using sample covariance eigenvectors is marked with diamonds, and is very similar to $(2/k^*)S_n$.

It is clear that $S_{AR,II}$ is biased but the other three estimators are reasonable for large n. The estimator $S_{AR,I}$ is less biased than $S_{AR,II}$ particularly for smaller n. The bias corrected estimator S_B also performs well.

5 Discussion

5.1 Planar size and shape analysis

There are other application areas for which the developments in this paper are relevant. A possible candidate model for planar size-and-shape analysis is the zero mean complex Gaussian distribution (Dryden and Mardia, 1998, p. 189). Let z^o denote a k-vector of complex co-ordinates and let $Y = HY^o$ be the Helmertized version, where H is the Helmert sub-matrix (Dryden and Mardia, p. 34). The zero mean complex Gaussian model is

$$Y \sim \mathbb{C}N_{k-1}(0, \Sigma_C),$$

where Σ_C is complex Hermitian. Since this distribution is invariant under rotations of z, it is suitable as a size-and-shape distribution. Let us assume Σ_C has complex eigenvectors $q_1, \ldots, q_{k-1}$ corresponding to real eigenvalues $\lambda_1 > \ldots > \lambda_{k-1} > 0$. The modal shape is q_1 and the modal size is λ_1. Consider observations $Y_1, \ldots, Y_n$ available from the general Gaussian model with the restriction that it has complex Gaussian structure (cf. Goodman, 1963). We construct the sample complex Hermitian covariance matrix

$$\hat{\Sigma}_C = \frac{1}{n} \sum_{i=1}^{n} Y_i Y_i^*,$$

where Y^* denotes the transpose of the complex conjugate of Y. Let $\hat{q}_j$, $j = 1, \ldots, p$, denote the sample eigenvectors of $\hat{\Sigma}_C$ and let

$$F_n = n^{1/2}(\hat{\Sigma}_C - \Sigma_C).$$

By analogy with Theorem 3.1 and Corollary 3.2 in the real case, under the complex Gaussian model with correlations which satisfy a condition similar to (10), we obtain the following, as $n \to \infty$: $F_n \xrightarrow{D} F$, where $\mathrm{vech}(F)$ has a complex multivariate Gaussian distribution analogous to (11), but with Hermitian covariance matrix,

$$n^{1/2}(\hat{q}_j - q_j) \xrightarrow{D} \sum_{k \neq l} \frac{q_k^* F q_j}{\lambda_j - \lambda_k} q_k \quad \text{and} \quad n^{1/2}(\hat{\lambda}_j - \lambda_j) \xrightarrow{D} \mathrm{tr}(F q_j q_j^*).$$

Under these assumptions, we can construct confidence intervals for shape q_1 and size λ_1 under the temporally correlated model. Also note that the estimators of shape and size based on the m.l.e. of Σ_S under a particular family of temporal correlation models will give a more efficient estimator under that family when the model is correct than if no temporal correlation is assumed.

5.2 Further points

Inspection of the periodograms for the PC scores in the DNA example indicates that some form of periodicity may be present. Periodic type behaviour can be exhibited by AR(2) models, lending

some extra weight to our choice of temporal model. Alternatively we could work with explicit periodic models. However, given that the periods themselves appear random we believe that the AR(2) model will provide reasonable estimators for entropy, which is the main aim.

The DNA strand in our example is symmetric in labelling - strand 1 and strand 2 could be interchanged and nucleotides letters (A,C,G,T) labelled in reverse order. So, we can consider symmetric PCA, where the dataset size is doubled by including both the original strand labelling and the data with the alternative strand labelling. When examining the effect of the PCs it is clear that there is little difference between the symmetric and standard PCA.

All our modelling has assumed Gaussian data. It would be good to develop the work for non-Gaussian models, even though in our applications there is no reason to doubt the Gaussian assumption. However, inference is likely to be rather more difficult in the non-Gaussian case.

An alternative method to using molecular dynamic simulations is to use Markov chain Monte Carlo methods for simulating from the Gibbs distributions at the molecular level. Entropy can then be directly calculated from such models. This approach is very complicated although it would be of interest to link such an analysis with the molecular dynamics simulations.

References

Arcones, M. A. (1994). Limit theorems for nonlinear functions of a stationary Gaussian field of vectors. *Ann. Probab.*, 22: 2242–2274.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.

Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis*. Wiley, Chichester.

Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *Annals of Mathematical Statistics*, 34:152–177.

Harris, S. A., Gavathiotis, E., Searle, M. S., Orozco, M., and Laughton, C. A. (2001). Cooperativity in drug-DNA recognition: a molecular dynamics study. *Journal of the American Chemical Society*, 123:12658–12663.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Schlitter, J. (1993). Estimation of absolute entropies of macromolecules using the covariance matrix. *Chemical Physics Letters*, 215:617–621.

Siddiqui, M. M. (1958). On the inversion of the sample covariance matrix in a stationary autoregressive process. *Ann. Math. Statist.*, 29:585–588.

Watson, G.S. (1983). *Statistics on Spheres*. University of Arkansas Lecture Notes in the Mathematical Sciences, Vol. 6. John Wiley, New York.

Appendix: Proofs of the results

The following standard result is used repeatedly: if X, Y, Z, W are zero-mean jointly Gaussian random variables then

$$\mathrm{cov}(XY, ZW) = \sigma_{XZ}\sigma_{YW} + \sigma_{XW}\sigma_{YZ}, \tag{28}$$

where $\sigma_{AB} = \mathrm{cov}(A, B)$. The following elementary lemma, whose proof is omitted, is used in the proof of Theorem 3.1.

Lemma A1

Suppose that the sequence a_h, $-\infty < h < \infty$, satisfies $\sum_{h=-\infty}^{\infty} |a_h| < \infty$. Then there exists a sequence $b_h \geq 1$, $-\infty < h < \infty$, such that $b_h \to \infty$ as $|h| \to \infty$ and $\sum_{h=-\infty}^{\infty} |a_h| b_h < \infty$.

Proof of Theorem 3.1

Using (6),

$$E(\bar{Z}\bar{Z}^T) = \mathrm{diag}(\gamma_{n,r} : \ r = 1, \ldots, p) \tag{29}$$

where (7) is equal to

$$\gamma_{n,r} = n^{-1}\lambda_r \sum_{h=-n+1}^{n-1} \left(1 - \frac{|h|}{n}\right) \rho_r(h) \tag{30}$$

Using the Cauchy-Schwarz inequality we have, for any sequence of positive real numbers $(b_h)_{h=-\infty}^{\infty}$,

$$\begin{aligned} |\gamma_{n,r}| &= \left| n^{-1}\lambda_r \sum_{h=-n+1}^{n-1} \left(1 - \frac{|h|}{n}\right) \rho_r(h) \right| \\ &\leq n^{-1}\lambda_r \sum_{h=-n+1}^{n-1} (b_h^{1/2}|\rho_r(h)|)(b_h)^{-1/2} \\ &\leq n^{-1}\lambda_r \left(\sum_{h=-n+1}^{n-1} b_h \rho_r(h)^2 \right)^{1/2} \left(\sum_{h=-n+1}^{n-1} b_n^{-1} \right)^{1/2}. \end{aligned}$$

Using Lemma A1, we may choose the sequence b_h such that $1 \leq b_h \to \infty$ as $|h| \to \infty$, and $\sum_{h=-\infty}^{\infty} b_h \rho_r(h)^2 < \infty$. Moreover, since $b_h^{-1} \to 0$ as $|h| \to \infty$, it follows that, for such a choice of the b-sequence,

$$\sum_{h=-n+1}^{n-1} b_h^{-1} = o(n).$$

Therefore $|\gamma_{n,r}| = o(n^{-1/2})$ under condition (10). It follows that, under (10),

$$||E(\bar{Z}\bar{Z}^T)|| = o(n^{-1/2}).$$

Also, under model (1)-(3),

$$\mathrm{cov}\{\mathrm{vech}(\bar{Z}\bar{Z}^T)\} = \mathrm{diag}[\mathrm{var}\{\bar{Z}(r)\bar{Z}(s)\} : \ 1 \leq r \leq s \leq p],$$

where, using (6) and (28),

$$\mathrm{var}\{\bar{Z}(r)\bar{Z}(s)\} \quad = \quad \begin{cases} 2\gamma_{n,r}^2 & \text{if } \ r = s \\ \\ \gamma_{n,r}\gamma_{n,s} & \text{if } \ r < s, \end{cases}$$

from which it follows that

$$||\mathrm{cov}\{\mathrm{vech}(\bar{Z}\bar{Z}^T)\}|| = o(n^{-1}).$$

Consequently, $||n^{1/2}\bar{Z}\bar{Z}^T|| = o_p(1)$, and therefore

$$n^{1/2}(\hat{\Lambda} - \Lambda) = n^{1/2}(U_n - \Lambda) + R_n \tag{31}$$

where $U_n = n^{-1}\sum_{i=1}^n Z_i Z_i^T$ and $||R_n|| = o_p(1)$. Using (28) again,

$$\mathrm{cov}\{n^{1/2}\mathrm{vech}(U_n - \Lambda)\} \to \mathcal{V} \quad \text{as} \quad n \to \infty. \tag{32}$$

Moreover, using (32), the Cramér-Wold device and Theorem 4 of Arcones (1994), in which the relevant Hermite rank is 2, we obtain

$$n^{1/2}\{\mathrm{vech}(U_n - \Lambda)\} \xrightarrow{D} N_P(0, \mathcal{V}). \tag{33}$$

Finally, (11) follows from (31), (33) and the fact that

$$\mathrm{vech}(\hat{\Sigma}_S - \Sigma_S) = \mathcal{C}\ \mathrm{vech}(\hat{\Lambda} - \Lambda). \tag{34}$$

Proof of Corollary 3.2

This is a standard application of the delta method.

Proof of Proposition 3.3

This is a consequence of the following:

$$E(\hat{\Lambda} - \Lambda) = -\mathrm{diag}(\gamma_{n,r} : \ r = 1, \ldots, p), \tag{35}$$

the fact that, under assumption (13),

$$\gamma_{n,r} \sim n^{-1}\sum_{h=-\infty}^{\infty} \rho_r(h) = n^{-1}\upsilon_r \quad \text{as} \quad n \to \infty,$$

and the identity (34).

Proof of Corollary 3.4

This result follows directly from a second-order Taylor expansion and Proposition 3.3.

Proof of Proposition 3.5

Consider (30) and (35). By assumption, $\rho_r(h) \sim \beta_r |h|^{-\alpha}$ as $|h| \to \infty$, where $0 < \alpha \leq 1$. Therefore, when $0 < \alpha < 1$,

$$\begin{aligned}
\gamma_{n,r} &= n^{-\alpha}\lambda_r n^{-1} \sum_{h=-n+1}^{n-1} \left(1 - \frac{|h|}{n}\right) n^{\alpha}\rho_r(h) \\
&\sim n^{-\alpha}\lambda_r\beta_r n^{-1} \sum_{h=-n+1}^{n-1} \left(1 - \frac{|h|}{n}\right) \left(\frac{|h|}{n}\right)^{-\alpha} \\
&\sim n^{-\alpha}\lambda_r\beta_r \int_{-1}^{1} (1 - |x|)|x|^{-\alpha} dx \\
&= 2n^{-\alpha}\lambda_r\beta_r \frac{1}{(1-\alpha)(2-\alpha)},
\end{aligned}$$

as $n \to \infty$. In the case $\alpha = 1$, a slightly more delicate approximation argument gives

$$\gamma_{n,r} \sim 2\lambda_r\beta_r n^{-1} \log n \quad \text{as} \quad n \to \infty.$$

The first part of Proposition 3.5 now follows directly from the identity (34).

To establish the second part of Proposition 3.5, consider the identity

$$\hat{\Lambda} - \Lambda = U_n - \Lambda - \bar{Z}\bar{Z}^T,$$

where $U_n = n^{-1}\sum_{i=1}^{n} Z_i Z_i^T$ as before. It follows that

$$\operatorname{cov}\{\operatorname{vech}(\hat{\Lambda})\} = \operatorname{cov}\{\operatorname{vech}(U_n)\} + \operatorname{cov}\{\operatorname{vech}(\bar{Z}\bar{Z}^T)\} - A - A^T, \tag{36}$$

where $A = \operatorname{cov}\{\operatorname{vech}(U_n), \operatorname{vech}(\bar{Z}\bar{Z}^T)\}$. From the IPC assumption,

$$\operatorname{cov}\{\operatorname{vech}(U_n)\} = \operatorname{diag}[\operatorname{var}\{U_n(r,s)\} : \ 1 \leq r \leq s \leq p]$$

and

$$\operatorname{cov}\{\operatorname{vech}(\bar{Z}\bar{Z}^T)\} = \operatorname{diag}[\operatorname{var}\{\bar{Z}(r)\bar{Z}(s)\} : \ 1 \leq r \leq s \leq p].$$

Using (28), we obtain

$$\operatorname{cov}\{U_n(r,r)\} = 2n^{-1} \sum_{h=-n+1}^{n-1} \left(1 - \frac{|h|}{n}\right) \rho_r(h)^2; \tag{37}$$

when $r \neq s$,

$$\mathrm{cov}\{U_n(r,s)\} = n^{-1} \sum_{h=-n+1}^{n-1} \left(1 - \frac{|h|}{n}\right) \rho_r(h)\rho_s(h); \tag{38}$$

using (6) and (28), we obtain

$$\mathrm{var}\{\bar{Z}(r)\bar{Z}(r)\} = 2\gamma_{n,r}^2; \tag{39}$$

and, when $r \neq s$,

$$\mathrm{var}\{\bar{Z}(r)\bar{Z}(s)\} = \gamma_{n,r}\gamma_{n,s}. \tag{40}$$

Moreover, A is also a diagonal matrix with elements $\mathrm{cov}\{U_n(r,s), \bar{Z}(r)\bar{Z}(s)\}$ which, by the Cauchy-Schwartz inequality, satisfy

$$|\mathrm{cov}\{U_n(r,s), \bar{Z}(r)\bar{Z}(s)\}| \leq \left[\mathrm{var}\{U_n(r,s)\}\,\mathrm{var}\{\bar{Z}(r)\bar{Z}(s)\}\right]^{1/2}. \tag{41}$$

Finally, using the results obtained for $\gamma_{n,r}$ in the first part of the proof of Proposition 3.5, and obtaining similar asymptotic expressions for (37) and (38) by approximating the sums by integrals, we find that (37)-(41) are all of the appropriate order, so by (36) the proof of the second part of Proposition 3.5 is now complete.

Proof of Corollary 3.6

This result follows from a first-order Taylor expansion, which is all that is needed, because the second-order term is of strictly smaller order than the leading term, by Proposition 3.5.

Proof of Theorem 4.1

For a symmetric matrix M,

$$\log \mid I_p + M \mid = \mathrm{tr}(M) - \frac{1}{2}\mathrm{tr}(M^2) + O(\|M\|^3),$$

and therefore

$$\begin{aligned} \log \mid I_p + c\hat{\Sigma}_S \mid &= \log \mid I_p + c\Sigma_S + c(\hat{\Sigma}_S - \Sigma_S), \mid \\ &= \log \mid I_p + c\Sigma_S \mid + \log_e \mid I_p + A \mid \end{aligned}$$

where

$$A = cF(\hat{\Sigma}_S - \Sigma_S)F$$

and $F = (I_p + c\Sigma_S)^{-1/2}$ are both symmetric. We have

$$\begin{aligned} S_n &= S_\infty + \frac{k^*}{2}\mathrm{tr}(A) - \frac{k^*}{4}\mathrm{tr}(A^2) + O_p(||A||^3) \\ &= S_\infty + \frac{k^*}{2}\mathrm{tr}(A) - \frac{k^*}{4}\mathrm{tr}[\{A - E(A)\}^2 + AE(A) + E(A)A - \{E(A)\}^2] + O_p(||A||^3). \end{aligned}$$

Therefore

$$E(S_n) = S_\infty + \frac{k^*}{2} E\{\mathrm{tr}(A)\} + \frac{k^*}{4} E\left(\mathrm{tr}[\{A - E(A)\}^2]\right) + o\left[E\{\mathrm{tr}(A)\}\right].$$

Now

$$\begin{aligned} E\{\mathrm{tr}(A)\} &= \mathrm{tr}\{E(A)\} \\ &= -\mathrm{tr}\{cF\,\mathrm{diag}(\gamma_{n,r} : \ r = 1, \ldots, p)F\} \\ &= -c \sum_{r=1}^{p} \frac{\gamma_{n,r}}{1 + c\lambda_r}, \end{aligned}$$

which gives the first term in (23), corresponding to $\alpha > 1$, and also gives (24) and (25) when $\alpha = 1$ and $\alpha \in (0, 1)$, respectively, since

$$\gamma_{n,r} \sim \begin{cases} \upsilon_r n^{-1} & \alpha > 1 \\ 2\lambda_r \beta_r n^{-1} \log n & \alpha = 1 \\ \frac{2\lambda_r \beta_r}{(1-\alpha)(2-\alpha)} n^{-\alpha} & \alpha \in (0, 1). \end{cases}$$

Moreover, when $\alpha > 1$,

$$\begin{aligned} E\left(\mathrm{tr}[\{A - E(A)\}^2]\right) &= \mathrm{tr}\left(E[\{A - E(A)\}^2]\right) \\ &= c^2 \sum_{r,s=1}^{p} \frac{\mathrm{var}\{\hat{\Lambda}(r,s)\}}{(1 + c\lambda_r)(1 + c\lambda_s)} \\ &\sim n^{-1} c^2 \sum_{r,s=1}^{p} \frac{\mathcal{V}\{(r,s);(r,s)\}}{(1 + c\lambda_r)(1 + c\lambda_s)}. \end{aligned}$$

This gives the second term in (23), corresponding to $\alpha > 1$. Note that, by the second part of Proposition 3.5, this second-order term is asymptotically negligible when $0 < \alpha \leq 1$. The proof is now complete.

Covariance sums for AR(2) models.

Simple expressions for the diagonal elements of $\mathcal{V}$ in Theorem 3.1 may be derived in the AR(2) case. If ψ_{1r} and ψ_{2r} are the parameters of an AR(2) process for the rth PC score then the auto-correlation function has the form

$$\rho_r(0) = 1$$
$$\rho_r(h) = a_{1r} \xi_{1r}^{|h|} + a_{2r} \xi_{2r}^{|h|}$$

where

$$a_{1r} = \frac{\psi_{1r}/(1 - \psi_{2r}) - \xi_{2r}}{\xi_{1r} - \xi_{2r}} \quad \text{and} \quad a_{2r} = \frac{\xi_{1r} - \psi_{1r}/(1 - \psi_{2r})}{\xi_{1r} - \xi_{2r}},$$

and ξ_{1r}, ξ_{2r} are the solutions of

$$\xi^2 - \psi_{1r}\xi - \psi_{2r} = 0,$$

i.e.

$$\xi_{1r}, \xi_{2r} = \frac{1}{2}\left(\psi_{1r} \pm \sqrt{\psi_{1r}^2 + 4\psi_{2r}}\right).$$

Then for $1 \leq r \leq s \leq p$,

$$\sum_{h=-\infty}^{\infty} \rho_r(h)\rho_s(h) = a_{1r}a_{1s}\frac{1+\xi_{1r}\xi_{1s}}{1-\xi_{1r}\xi_{1s}} + a_{1r}a_{2s}\frac{1+\xi_{1r}\xi_{2s}}{1-\xi_{1r}\xi_{2s}} + a_{2r}a_{1s}\frac{1+\xi_{2r}\xi_{1s}}{1-\xi_{2r}\xi_{1s}} + a_{2r}a_{2s}\frac{1+\xi_{2r}\xi_{2s}}{1-\xi_{2r}\xi_{2s}} \tag{42}$$

See Cox and Miller (1965, Chapter 7.2) for similar calculations. Hence, the estimate $\hat{\mathcal{V}}$ required for the confidence interval calculation in equation (12) can be computed easily in this particular case.