Inferring the eigenvalues of covariance matrices from limited, noisy data

Richard Everson, Stephen Roberts

Abstract

The eigenvalue spectrum of covariance matrices is of central importance to a number of data analysis techniques. Usually the sample covariance matrix is constructed from a limited number of noisy samples. We describe a method of inferring the true eigenvalue spectrum from the sample spectrum. Results of Silverstein, which characterise the eigenvalue spectrum of the noise covariance matrix, and inequalities between the eigenvalues of Hermitian matrices are used to infer probability densities for the eigenvalues of the noise-free covariance matrix, using Bayesian inference. Posterior densities for each eigenvalue are obtained, which yield error estimates. The evidence framework gives estimates of the noise variance and permits model order selection by estimating the rank of the covariance matrix. The method is illustrated with numerical examples.

Keywords

sample covariance, eigenvalue spectrum, Bayesian evidence, model order selection.

I. Introduction

The covariance matrix and its spectrum of eigenvalues are of great interest in the analysis and modelling of experimental data. Principal Components Analysis [1], [2], the Karhunen-Loève decomposition [3], [4] and related techniques such as Independent Components Analysis (ICA) [5] model N-dimensional data vectors $\mathbf{x}(t)$ as an admixture of M $(M \leq N)$ decorrelated (or in the case of ICA, statistically independent) sources, $\mathbf{s}(\mathbf{t})$, which are linearly mixed by $A \in \mathbb{R}^{\mathbb{N}\times\mathbb{M}}$; thus

$$\mathbf{x}(t) = A\mathbf{s}(\mathbf{t}). \tag{1}$$

Without loss of generality we may assume that the sources each have mean zero and unit variance, so that

$$\langle \mathbf{s}(\mathbf{t})\mathbf{s}(\mathbf{t}) \rangle = I_M. \tag{2}$$

Richard Everson is with Department of Computer Science, Exeter University. Exeter. UK. E-mail: R.M.Everson@exeter.ac.uk
Stephen Roberts is with Department of Engineering Science, University of Oxford. Oxford. UK. E-mail: sjrob@robots.ox.ac.uk

Before the data are examined the number of sources, M, is usually unknown and determination of M is a model order selection problem [6], [7]. The number may, in principle at least, be deduced from the rank of the covariance matrix

$$C_{\mathbf{x}} = \left\langle \mathbf{x}(t)\mathbf{x}^{\mathsf{T}}(t) \right\rangle \tag{3}$$

where $\langle \cdot \rangle$ denotes the expectation. In the absence of noise

$$C_{\mathbf{x}} = A \left\langle \mathbf{s}(\mathbf{t})\mathbf{s}(\mathbf{t})^{\mathsf{T}} \right\rangle A^{\mathsf{T}} = AA^{\mathsf{T}}, \tag{4}$$

which clearly has rank equal to the number of sources. In this case the eigenvalue spectrum of $C_{\mathbf{x}}$ is comprised of M positive eigenvalues and $N - M$ zeroes, thus $\omega_1 \geq \omega_2 \geq \ldots \omega_M > \omega_{M+1} = \cdots = \omega_N = 0$.[1] Methods such as ICA, which assume that the sources are statistically independent, use higher order (i.e., greater than second order) statistics to estimate A^{-1} [5], [8], [9]. Note however that the *number* of sources is still determined by the rank of AA^{T} because statistical independence implies linear decorrelation (provided that the sources each have mean zero).

Inevitably the data are contaminated by noise and equation (1) might be replaced by

$$\mathbf{x}(t) = A\mathbf{s}(\mathbf{t}) + \sigma \mathbf{v}(\mathbf{t}) \tag{5}$$

where $\mathbf{v}(\mathbf{t})$ denotes an N-dimensional random noise vector whose elements are independently and identically distributed (i.i.d.) noise, with mean zero and unit variance. Consequently

$$C_{\mathbf{x}} = AA^{\mathsf{T}} + 2\sigma A \left\langle \mathbf{s}(\mathbf{t})\mathbf{v}(\mathbf{t})^{\mathsf{T}} \right\rangle + \sigma^2 I_N \tag{6}$$

Since the noise and the signal are assumed to be uncorrelated the data covariance is the sum of AA^{T} and the noise covariance:

$$C_{\mathbf{x}} = AA^{\mathsf{T}} + \sigma^2 I_N \tag{7}$$

The effect of the noise, therefore, is merely to raise all the eigenvalues of $C_{\mathbf{x}}$ by σ^2, so that $C_{\mathbf{x}}$ now has full rank and $N - M$ eigenvalues equal to σ^2. The noise variance is readily found from the smallest ω_n and so the rank of AA^{T} is

[1] Throughout this article we adopt the convention of listing eigenvalues in order of *decreasing* magnitude.

easily determined.

The subject of this paper is the determination of the eigenvalue spectrum $\{\lambda_n\}$ of AA^{T} when the number of observations T is limited. In this case the noise covariance matrix is not diagonal, and the sample covariance matrix is

$$\hat{C}_{\mathbf{x}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}(t)\mathbf{x}^{\mathsf{T}}(t), \tag{8}$$

which converges to the true covariance matrix, $C_{\mathbf{x}}$, in the limit of infinite observations [4]. With limited data equation 7 is replaced by

$$\hat{C}_{\mathbf{x}} = AA^{\mathsf{T}} + \sigma^2\hat{C}_{\mathbf{v}} \tag{9}$$

where $\hat{C}_{\mathbf{v}}$ is the sample noise covariance matrix.

Results of Silverstein [10] characterise the eigenvalue spectrum of the noise covariance matrix and inequalities between the eigenvalues of Hermitian matrices are used to infer probability densities for the eigenvalues of AA^{T}, using Bayesian inference. Section II summarises Silverstein's work on the eigenvalues of sample covariance matrices; these are incorporated in a Bayesian model in section III. The use of the evidence to infer the number of non-zero eigenvalues and the noise variance is discussed in section IV.

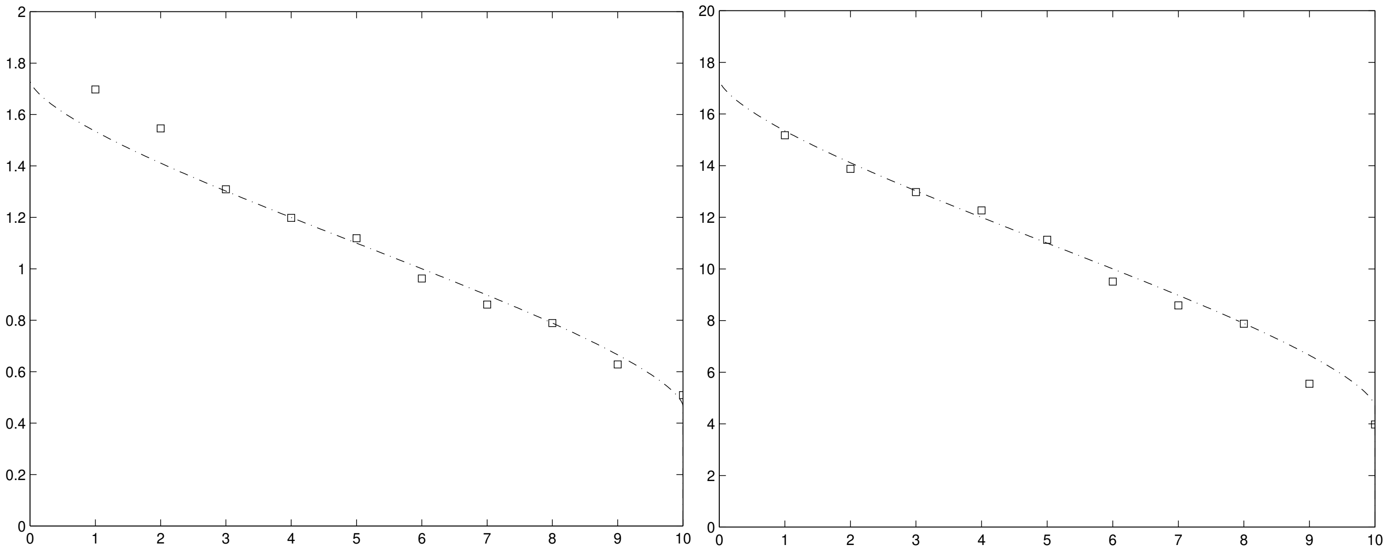
II. Eigenvalues of sample covariance matrices

Silverstein [10] has proved a remarkable result characterising the eigenvalue spectrum of a sample covariance matrices. For each N let V_N be an $N \times T$ matrix consisting of i.i.d., mean 0, variance 1 random variables v_{ij} with distribution common for all N. Let $T/N \to y > 0$ as $N \to \infty$ and denote the sample covariance matrix by $\hat{C}_N = (1/T)V_N V_N^{\mathsf{T}}$. The result is stated in terms of the empirical distribution function, F_N of the eigenvalues of $\hat{C}_N$; that is, for every η, $F_N(\eta) = \frac{1}{N}\times$ (number of eigenvalues of $\hat{C}_{\mathbf{v}} \leq \eta$).

With slight changes in notation (and correction of a typo) we quote his result:

.0.a Theorem: (Silverstein). If there exists a $\delta > 0$ such that $\left\langle |v_{11}|^{2+\delta} \right\rangle < \infty$, then for every $\eta \in \mathbb{R}$, $F_N(\eta) \xrightarrow{as} F_y(\eta)$ as $N \to \infty$, where for $0 < y \leq 1$

$$F_y'(\eta) = f_y(\eta) = \begin{cases} \frac{1}{2\pi y\eta}\sqrt{(\eta - b_-)(b_+ - \eta)} & \text{if } b_- < \eta < b_+ \\ 0 & \text{otherwise} \end{cases} \tag{10}$$



a b

Fig. 1. Eigenvalues of sample covariance matrices. **a:** Eigenvalues, η_n, of a sample covariance matrix constructed from $T = 100$ random vectors of dimension $N = 10$. The dashed line is η plotted versus $n = N(1 - F(\eta))$, which is the cumulative probability that there are n eigenvalues greater than η. **b:** The 10 non-zero eigenvalues of a sample covariance matrix constructed from $T = 10$ Gaussian-distributed random vectors each of dimension $N = 100$. Here the dashed line is η versus $n = T(1 - F(\eta))$.

where $b_\pm = (1 \pm \sqrt{y})^2$, and for $1 < y < \infty$,

$$F_y(\eta) = (1 - \frac{1}{y}) I_{[0,\infty)}(\eta) + \frac{1}{y} \int_{b_-}^{\eta} f_y(t) dt, \tag{11}$$

where $I_S(\eta) = 1$ if $\eta \in S$ and zero otherwise.

The first term on the right hand side of (11) represents the $N - T$ zero eigenvalues which must occur when there are fewer samples than the dimension of the sample vectors. This is commonly the case in the analysis of ensembles of images, each of which has a great many pixels [11].

When the number of samples is very large so that $y \ll 1$, equation (10) reproduces the usual approximation that the eigenvalues of $\hat{C}_{\mathbf{v}}$ are all unity. However, as y approaches 1 the smallest eigenvalue decreases towards zero (being equal to b_-) and the largest eigenvalue (equal to b_+) increases. It is worthy of note that in this regime both the mean and the mode of f are greater than 1, so that in addition to spreading the range of the eigenvalues, limited sampling inflates the effect of noise. For $y = 1$ there is a single zero eigenvalue (because the v_{ij} have zero mean), and as y becomes large all of the T non-zero eigenvalues approach y. Here again, a limited number of samples magnifies the apparent noise.

Although Silverstein's theorem is true in the limit $N \to \infty$, numerical experiments show that it is a good approximation even for N as small as 10. As an illustration, we display in figure 1a the eigenvalues of a single sample covariance matrix constructed from $T = 100$ random vectors of dimension $N = 10$, so that $y = 0.1$. The elements of the random vectors are

Gaussian distributed with zero mean and unit variance. Also shown is the line η versus $n = N(1 - F(\eta))$, which is the cumulative probability that there are n eigenvalues greater than η. As the graph shows there is fairly good agreement between Silverstein's asymptotic result and the eigenvalues of this single, low-dimensional realisation. The figure also illustrates that the largest eigenvalue is substantially larger than the noise variance ($\sigma^2 = 1$). Figure 1b shows the 10 non-zero eigenvalues from a covariance matrix with $y = 10$, constructed from $T = 10$ random vectors ($\sigma^2 = 1$) each of dimension 100. Again there is reasonable agreement with the asymptotic result, and the eigenvalues are located around y, in this case 10 times the noise variance.

III. Eigenvalues of AA^T

Although the covariance matrices add,

$$\hat{C}_{\mathbf{x}} = AA^\mathsf{T} + \sigma^2 \hat{C}_{\mathbf{v}}, \tag{12}$$

the eigenvalues, ω_n, of $\hat{C}_{\mathbf{x}}$ are nonlinear functions of λ_n and η_n, the eigenvalues of AA^T and $\hat{C}_{\mathbf{v}}$ respectively. Since all the matrices involved are symmetric, bounds on the ω_n are given by inequalities attributed to Weyl, which are quoted in the Appendix (see also [12], [13]). Listing the eigenvalues in decreasing order of magnitude, Weyl's inequalities imply that

$$\lambda_{n+N-j} \ + \ \sigma^2\eta_j \ \ \leq \ \ \omega_n \ \ \leq \ \ \lambda_{n-k+1} \ + \ \sigma^2\eta_k \tag{13}$$

where $1 \leq k \leq n \leq j \leq N$.

When $N > T$, so that $\hat{C}_{\mathbf{v}}$ is of full rank, Silverstein's result shows that the η_i are bounded by

$$\sigma^2 b_- \ \leq \ \eta_i \ \leq \ \sigma^2 b_+. \tag{14}$$

Combining (13) and (14) gives bounds on ω_n:

$$\lambda_n + \sigma^2 b_- \ \leq \ \omega_n \ \leq \ \lambda_n + \sigma^2 b_+. \tag{15}$$

If there are fewer than N samples $\hat{C}_{\mathbf{v}}$ has $N - T$ zero eigenvalues, and the upper and lower bounds in (15) each become a pair of inequalities, either of which may provide the tightest bound.

Although equation 15 provides rigorous bounds on ω_n, (and therefore on λ_n given ω_n) better, probabilistic, estimates may be obtained by considering the probability densities of the η_i.

A. Bayesian Inference

In order to estimate probability density functions for the λ_n we shall adopt a Bayesian point of view, using Bayes' rule in the form

$$p(\boldsymbol{\lambda}|\,\boldsymbol{\omega},\boldsymbol{\theta}) = \frac{p(\boldsymbol{\omega}|\,\boldsymbol{\lambda},\boldsymbol{\theta})\pi(\boldsymbol{\lambda}|\,\boldsymbol{\theta})}{p(\boldsymbol{\omega}|\,\boldsymbol{\theta})} \tag{16}$$

where $\boldsymbol{\lambda},\boldsymbol{\omega}$ denote N-dimensional random vectors formed from the eigenvalues λ_n and ω_n. Prior belief about the density of eigenvalues of AA^{T}, given a vector of parameters $\boldsymbol{\theta}$, is embodied in the *prior* density, $\pi(\boldsymbol{\lambda}|\,\boldsymbol{\theta})$. The parameters in this problem are $\boldsymbol{\theta} = (\sigma^2, M)^{\mathsf{T}}$, the (usually unknown) noise variance and rank of AA^{T}. Having observed the data, namely the eigenvalues $\boldsymbol{\omega}$ of $\hat{C}_{\mathbf{x}}$, the *posterior* density $p(\boldsymbol{\lambda}|\,\boldsymbol{\omega},\boldsymbol{\theta})$ for $\boldsymbol{\lambda}$ may be calculated using the *likelihood*, $p(\boldsymbol{\omega}|\,\boldsymbol{\lambda},\boldsymbol{\theta})$. The form of the likelihood is determined by the model, equation (5), and is calculated below. The denominator $p(\boldsymbol{\omega}|\,\boldsymbol{\theta})$, which may be determined by the requirement that the posterior density integrates to 1, is known as the *evidence* and is useful in determining the noise variance and the rank of AA^{T}; this is the subject of section IV and for now we omit the dependence of these densities on $\boldsymbol{\theta}$, for notational simplicity.

A.1 Likelihood

We model the ω_n as being conditionally independent given $\boldsymbol{\lambda}$, so that the (pseudo) likelihood is expressed as the product:

$$p(\boldsymbol{\omega}|\,\boldsymbol{\lambda}) = \prod_{n=1}^{N} p(\omega_n|\,\boldsymbol{\lambda}). \tag{17}$$

The likelihood $p(\omega_n|\,\boldsymbol{\lambda})$ is determined by model equation (5), and may be estimated from (13) as

$$\begin{aligned} p(\omega_n|\,\boldsymbol{\lambda}) &= P(\omega_n \geq \lambda_{n+N-j} + \sigma^2\eta_j \text{ and } \omega_n \leq \lambda_{n-k+1} + \sigma^2\eta_k) \qquad 1 \leq k \leq n \leq j \leq N && (18) \\ &= \prod_{j=n}^{N} P(\frac{\omega_n - \lambda_{n+N-j}}{\sigma^2} \geq \eta_j) \prod_{k=1}^{n} P(\frac{\omega_n - \lambda_{n-k+1}}{\sigma^2} \leq \eta_k). && (19) \end{aligned}$$

The probabilities appearing in (19) are no more than the cumulative densities for the eigenvalues η_n, which may be calculated, using elementary methods from order statistics, in the following way. The non-zero eigenvalues of $\hat{C}_{\mathbf{v}}$ may be regarded as $R = \min(T, N)$ realisations of the random variable η, whose density function is $f_y(\eta)$, given by equations (10) and (11). The cumulative density function, $F_n(\eta)$, of the nth *largest* eigenvalue η_n is the probability that at least $R - n$ of the eigenvalues are less than or equal to η. Thus

$$F_n(\eta) = P(\eta_n \leq \eta) = \sum_{i=R-n+1}^{R} \binom{R}{i} \left[F_y(\eta)\right]^i \left[1 - F_y(\eta)\right]^{R-i}, \tag{20}$$

which is readily calculated from $f_y(\eta)$.

Combining (19) and (20) gives

$$p(\omega_n|\,\boldsymbol{\lambda}) = \prod_{j=n}^{N} F_j\left(\frac{\omega_n - \lambda_{n+N-j}}{\sigma^2}\right) \prod_{k=1}^{n} \left\{1 - F_k\left(\frac{\omega_n - \lambda_{n-k+1}}{\sigma^2}\right)\right\} \tag{21}$$

$$= \prod_{j=n}^{N} F_j(\tilde{\omega}_n - \tilde{\lambda}_{n+N-j}) \prod_{k=1}^{n} \bar{F}_k(\tilde{\omega}_n - \tilde{\lambda}_{n-k+1}), \tag{22}$$

where for notational brevity we have written $\tilde{\omega}_n = \omega_n/\sigma^2$, $\tilde{\lambda}_n = \lambda_n/\sigma^2$ and $\bar{F}(\eta) = 1 - F(\eta)$. The likelihood is thus seen to be the product of $N - n + 1$ factors estimating lower bounds on ω_n and n factors estimating upper bounds. The lower bound factors, F_j, are cumulative density functions and so increase monotonically from zero at small (possibly negative) ω_n to unity when ω_n is sufficiently large. Conversely the upper bound factors, $\bar{F}_k$, decrease from unity to zero. Since there is always at least one F_j and one $\bar{F}_k$, the likelihood $p(\omega_n|\,\boldsymbol{\lambda})$ is zero for sufficiently small ω_n, rises monotonically with increasing ω_n to a maximum and then decreases monotonically to zero at large ω_n.

The full likelihood, equation (17), is therefore

$$p(\boldsymbol{\omega}|\,\boldsymbol{\lambda}) = \prod_{n=1}^{N} \prod_{j=n}^{N} F_{n+N-j}(\tilde{\omega}_n - \tilde{\lambda}_j) \prod_{k=1}^{n} \bar{F}_{n-k+1}(\tilde{\omega}_n - \tilde{\lambda}_k) \tag{23}$$

and identifying terms that depend upon λ_n gives the likelihood of $\boldsymbol{\omega}$ conditioned on λ_n:

$$p(\boldsymbol{\omega}|\,\lambda_n) = \prod_{j=1}^{n} F_{N-n+j}(\tilde{\omega}_j - \tilde{\lambda}_n) \prod_{k=n}^{N} \bar{F}_{k-n+1}(\tilde{\omega}_k - \tilde{\lambda}_n). \tag{24}$$

Provided that the prior for $\boldsymbol{\lambda}$ factorises as $p(\boldsymbol{\lambda}) = \prod_{n=1}^{N} p(\lambda_n)$, the posterior densities for each λ_n may be calculated separately using equation (24):

$$p(\lambda_n | \boldsymbol{\omega}) = \frac{p(\boldsymbol{\omega} | \lambda_n) p(\lambda_n)}{p(\boldsymbol{\omega})} \tag{25}$$

$$= \prod_{j=1}^{n} F_{N-n+j}(\tilde{\omega}_j - \tilde{\lambda}_n) \prod_{k=n}^{N} \bar{F}_{k-n+1}(\tilde{\omega}_k - \tilde{\lambda}_n) \, \frac{p(\lambda_n)}{p(\boldsymbol{\omega})}. \tag{26}$$

This factorisation of the posterior density $p(\boldsymbol{\lambda} | \boldsymbol{\omega}) = \prod_{n=1}^{N} p(\lambda_n | \boldsymbol{\omega})$ is a consequence of the form of the estimates for $p(\omega_n | \boldsymbol{\lambda})$ and the factorisation of the $\boldsymbol{\lambda}$ prior. Note that, unlike the likelihood (22), it is the factors $\bar{F}_{n-k+1}$ that shape the lower bounds of the posterior density, while the cumulative densities F_{N-n+j} determine the posterior probabilities for large λ_n.

Although there are $N+1$ of these factors, in practice only a few of them are relevant, because many of them are rendered impotent by the fact that another factor is zero everywhere that they are non-zero. Utilisation of this fact greatly speeds up computation of the likelihood. Since $b_- \leq \eta_n \leq b_+$ for all n, it is clear that $F_n(\eta) = 0$ for $\eta < b_-$ and $F_n(\eta) = 1$ for $\eta > b_+$. Consider the factors associated with the lower bounds (i.e., the $\bar{F}_{k-n+1}$ in (26)) and in particular the factor $\bar{F}_1(\tilde{\omega}_n - \tilde{\lambda}_n)$. This term is zero for $\tilde{\lambda}_n < \lambda^* = \tilde{\omega}_n - b_+$. Any other term which is 1 for $\tilde{\lambda}_n = \lambda^*$ (and therefore for any $\tilde{\lambda}_n > \lambda^*$) cannot play a role in shaping the likelihood, because $\bar{F}_1(\tilde{\omega}_n - \tilde{\lambda}_n)$ and therefore $p(\boldsymbol{\omega} | \lambda_n)$ is zero for $\tilde{\lambda}_n < \lambda^*$, and when $\tilde{\lambda}_n > \lambda^*$ the contribution of unity to the product (26) is irrelevant. The only potent contributions are therefore those F_{k-n+1} for which $\tilde{\omega}_k - b_- \geq \lambda^*$. Similar considerations show that the only potent upper bound factors are those for which $\tilde{\omega}_j - b_+ \leq \tilde{\omega}_n - b_-$. If λ_n is separated from its neighbours by at least $(b_+ - b_-)\sigma^2 = 4\sqrt{N/T}\sigma^2$ then only λ_n plays a role in determining ω_n.

The most time-consuming part of the likelihood calculation is the numerical integration of $f_y(\eta)$ to obtain $F_y(\eta)$. Each likelihood estimate requires $F(\eta)$ for different values of η, but great economies may be made by tabulating $F_y(\eta)$ (once) on a relatively fine mesh and then interpolating to the required η.[2]

A.2 Prior

In order to complete the Bayesian scheme a prior density, $\pi(\boldsymbol{\lambda} | \boldsymbol{\theta})$, must be chosen to reflect belief and prior knowledge about the eigenvalues of AA^{T}. Since AA^{T} is positive semi-definite, $\lambda_n \geq 0$. Non-negativity of the eigenvalues is not enforced by the inequalities (13), which are applicable to the wider class of Hermitian matrices. The prior should

[2]Matlab scripts implementing these calculations are available from `http://www.dcs.ex.ac.uk/academics/reverson`

therefore encode this knowledge about λ_n, thus

$$\pi(\lambda_n | \boldsymbol{\theta}) = 0 \quad \forall \lambda_n < 0. \tag{27}$$

In some instances, when N is large, λ_n may be regarded as scale variables, and may be expected to decay like $\lambda_n = a^{-kn}$ for some constants k and a. In this case a gamma distribution centred around a^{-kn} is a reasonable model for $\pi(\lambda_n | \boldsymbol{\theta})$. Note however that it may be important to allow for the possibility of zero eigenvalues by adopting a prior which is non-zero for $\lambda_n = 0$.

When N is relatively small and the data are thought to have been generated by a small number of roughly equal-powered sources the λ_n are all expected to be about the same size. In the absence of further information a uniform prior between 0 and some outer scale is most uninformative. Note that we have found a Jeffrey's prior (which gives equal weight to λ_n on a logarithmic scale; see e.g., [14]) to be unsuitable for this situation because: (1) it places too much weight at small scales; but (2) does not permit the possibility of an exactly zero eigenvalue. Since $\omega_n > \lambda_n$ a suitable outer scale for λ_n is ω_n, and we therefore choose

$$\pi(\lambda_n | \boldsymbol{\theta}) = \begin{cases} \frac{1}{\omega_n} & \text{if } 0 \leq \lambda_n < \omega_n \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

In this context the ω_n are hyper-parameters and formally their values may be found through a hierarchical Bayesian methodology.

In any case, since the likelihood is compact ($p(\omega_n | \boldsymbol{\lambda}, \boldsymbol{\theta}) = 0$ for ω_n outside the interval $[\lambda_n + \sigma^2 b_-, \lambda_n + \sigma^2 b_+]$) the precise form of the prior is not crucial provided that it is sufficiently broad that it does not unwarrentedly prejudice the posterior.

Finally, we shall wish to examine the hypothesis that the number of sources, M, is less than N. To do this we choose $\pi(\lambda_n | \sigma, M) = \delta(\lambda_n)$ for $n > M$. In summary we have

$$\pi(\boldsymbol{\lambda} | \sigma, M) = \prod_{n=1}^{M} \frac{1}{\omega_n} I_{[0,\omega_n]}(\lambda_n) \prod_{n=M+1}^{N} \delta(\lambda_n). \tag{29}$$

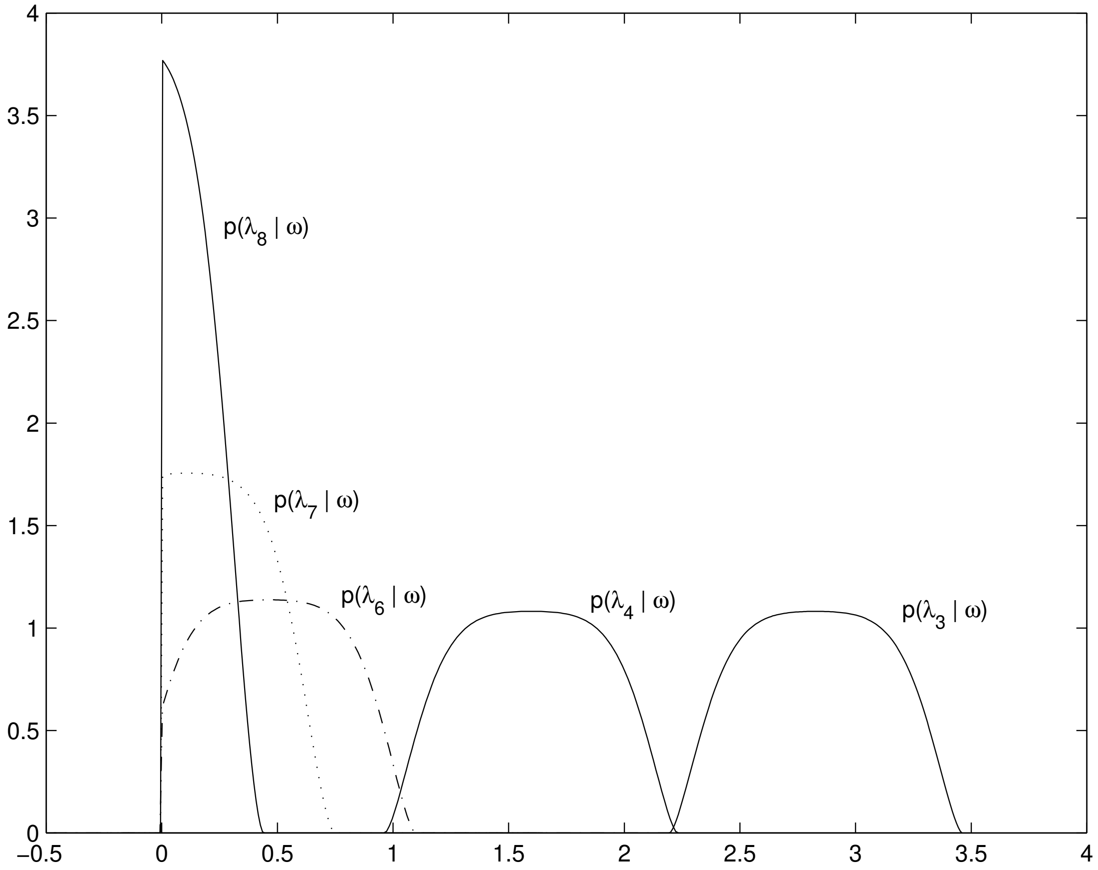


Fig. 2. Posterior densities for eigenvalues 3, 4, 6, 7 and 8.

B. Example

As an illustration we apply the method to covariance matrices corresponding to $M = 6$ sources and $T = 100$ observations of an $N = 10$-dimensional vector. The noise covariance matrix was constructed from unit variance Gaussian noise vectors. The noise power is given by $Tr\hat{C}_{\mathbf{v}} = 10.5$, and the signal power by $TrAA^{\mathsf{T}} = 19.0$. The eigenvalues of $\hat{C}_{\mathbf{v}}$ (which would usually be unknown) are those shown in figure 1b.

Figure 2 shows the posterior densities for some of the eigenvalues. This calculation used the known noise variance and the flat prior (28). The modes of the posteriors for the non-zero λ_n are close to the real values, while the posterior densities for λ_7 and λ_8 correctly indicate the eigenvalues to be zero or very close to zero.

Figure 3 shows the modes and standard deviations of posteriors for all the eigenvalues. In all cases the mode of the posterior is close to the true eigenvalue and an advantage of the Bayesian method is that error estimates are placed on the eigenvalues.

Also shown in figure 3 are the bounds $\omega_n - \sigma^2 b_+ \leq \lambda_n \leq \omega_n - \sigma^2 b_-$, which arise directly from the Weyl inequalities (13), Silverstein's result (10) and the fact that eigenvalues of covariance matrices are non-negative. Though these bounds

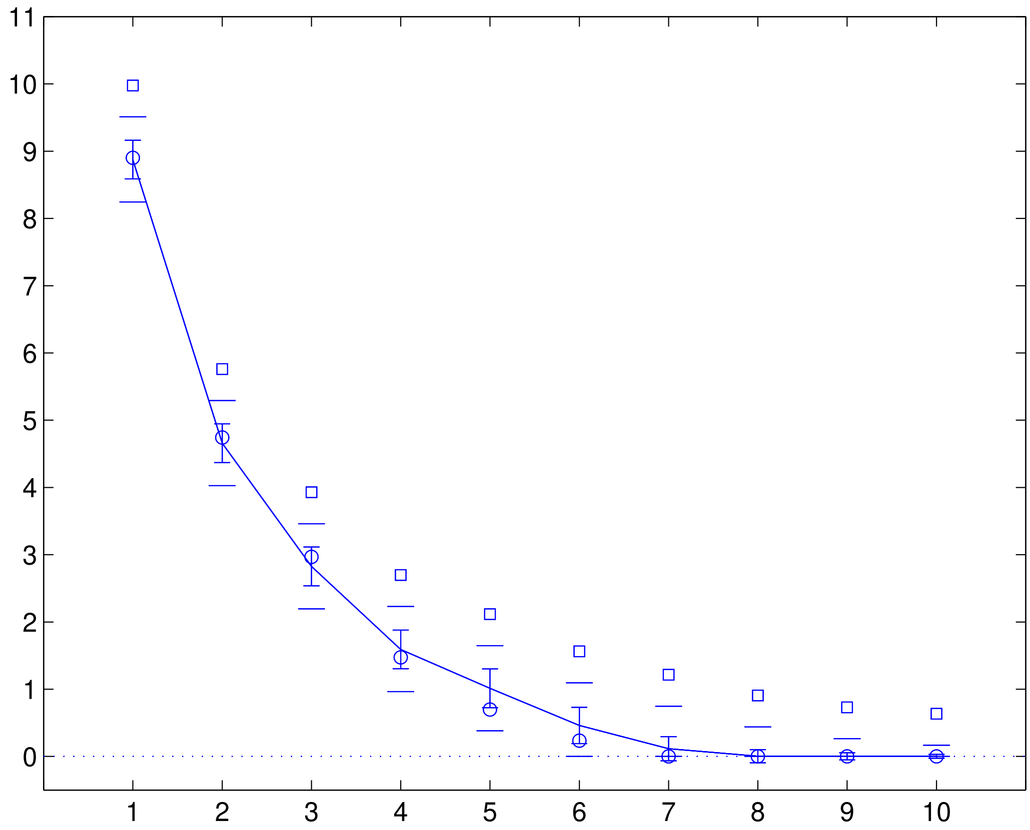


Fig. 3. Modes of the posterior densities. The solid line joins the modes of $p(\lambda_n | \omega_n)$ and error bars indicate the standard deviation of $p(\lambda_n | \boldsymbol{\omega})$. Circles show the true eigenvalues, λ_n, and squares mark the eigenvalues, ω_n, of the noise-corrupted covariance matrix. The longer horizontal lines indicate the non-probabilistic bounds arising from the Weyl inequalities and Silverstein's result.

are much looser their computation is extremely simple. Note in particular that the bounds placed on λ_7, the first zero eigenvalue, are rather loose, whereas the probabilistic estimates indicate that it is very close to zero.

IV. Evidence and model order selection

The noise variance, σ^2, and the rank M of AA^T, which are usually unknown parameters may be estimated from the *evidence*. This follows from the fact that posterior probability of $\boldsymbol{\theta} = (\sigma, M)^\mathsf{T}$ given $\boldsymbol{\omega}$ may be expressed as

$$p(\sigma, M | \boldsymbol{\omega}) = \frac{p(\boldsymbol{\omega} | \sigma, M)\pi(\sigma, M)}{p(\boldsymbol{\omega})}. \tag{30}$$

Since $p(\boldsymbol{\omega})$ is constant, the most probable σ and M are those which maximise the numerator of (30). If the prior $\pi(\sigma, M)$ is uninformative this is equivalent to maximising $p(\boldsymbol{\omega} | \sigma, M)$, which is therefore known as the evidence.

While σ is a continuous variable, M may only assume integer values and choosing the M for which there is most evidence thus constitutes selecting a model order, i.e., the rank of the matrix A. For predictive purposes the Bayesian approach is to integrate (marginalise) over all σ and one might also integrate over the model order M. In many cases

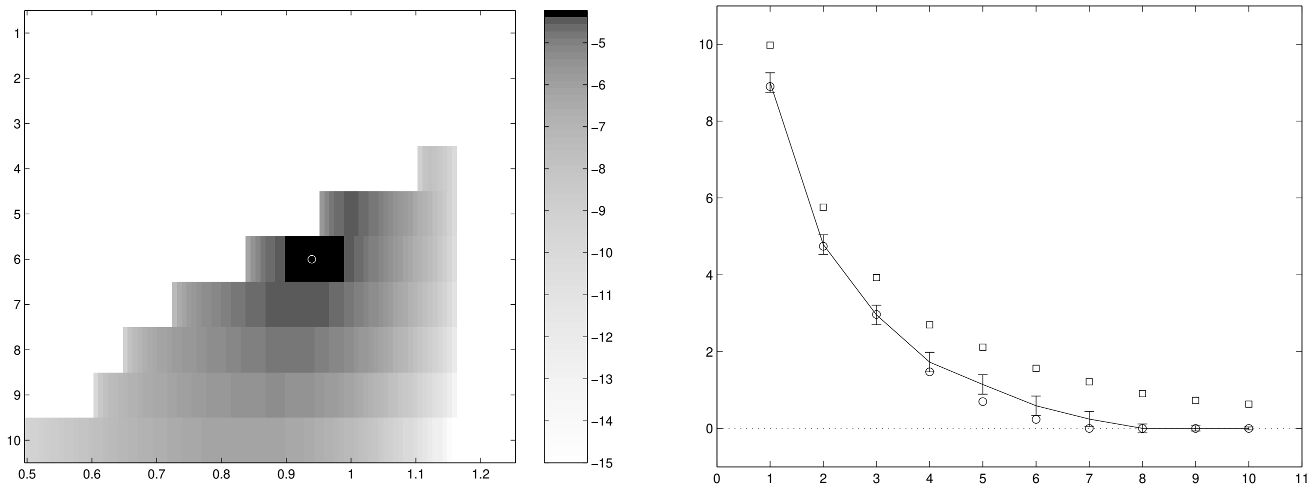


Fig. 4. Evidence for noise variance and rank. **a:** Gray scale shows $log_{10}p(\sigma, M|\,\boldsymbol{\omega})$; white indicates that $p(\sigma, M|\,\boldsymbol{\omega}) = 0$. The maximum at $\sigma = 0.94$, $M = 6$ is indicated by the white circle. **b:** Modes of posterior densities for λ_n calculated with the σ for which there is maximum evidence. Circles show the true eigenvalues, λ_n, and squares mark the eigenvalues ω_n of the noise-corrupted covariance matrix.

$p(\sigma, M|\,\boldsymbol{\omega})$ is sharply peaked in both σ and M so that choosing the modal values forms a good approximation to the full integration.

The minimum message length (MML) criterion [15], and the minimum description length (MDL) criterion [6] each seek to select the model order by determining the shortest string that describes the data in terms of the model and the data given the model. This balances model complexity (measured as the length, $-\log \pi(M)$, of a string describing the model), with the length, $-\log p(\boldsymbol{\omega}|\,M)$, of an additional string required to describe the data once the model is known. Since the length of a message describing the model is proportional to the model order, the MML criterion may be viewed as maximising $p(M|\,\boldsymbol{\omega})$ with the prior $\pi(M) = e^{-M}$.

Since σ may be regarded as a scale variable a Jeffrey's prior for $\pi(\sigma, M)$ may be appropriate in some circumstances. For now we choose the MML prior $\pi(\sigma, M) = e^{-M}$. Using (29), we have

$$p(\sigma, M|\,\boldsymbol{\omega}) = \prod_{n=1}^{M} \frac{1}{\omega_n} \int_0^{\omega_n} p(\boldsymbol{\omega}|\,\lambda_n, \sigma) d\lambda_n \prod_{n=M+1}^{N} p(\boldsymbol{\omega}|\,\lambda_n, \sigma)|_{\lambda_n=0}. \tag{31}$$

As figure 4 shows for the example discussed in §III-B, there is most evidence for $M = 6$ and $\sigma = 0.94$. The rank of AA^{T} has been correctly identified and the noise variance is close to the true value of unity. Figure 4 shows the eigenvalue spectrum at $\sigma = 0.94$.

Particularly when the rank of A is small there is, however, a tendency for the evidence calculation to underestimate σ and overestimate the rank M. The reason for this can be seen by examining equation (31). The terms in the second product are the likelihoods evaluated at $\lambda_n = 0$, each of which may be interpreted as the evidence that λ_n is zero. As argued in section III-A.1 the support for $p(\boldsymbol{\omega}|\,\lambda_n)$ is no larger than $[\omega_n - \sigma^2 b_+, \omega_n - \sigma^2 b_-]$ and the likelihood attains a maximum somewhere in this interval. When $\sigma^2 < \omega_n/b_+$ there is zero evidence for λ_n being zero because $p(\boldsymbol{\omega}|\,\lambda_n,\sigma)|_{\lambda_n=0} = 0$. As σ^2 becomes larger than ω_n/b_+ the likelihood (and therefore the evidence for $\lambda_n = 0$) increases, achieves a maximum for $\sigma \approx \sqrt{\omega_n}$, and then decreases. When $\sigma^2 > \omega_n/b_-$ the support for $p(\boldsymbol{\omega}|\,\lambda_n)$ lies entirely in the negative half axis and there is zero evidence for $\lambda_n = 0$.

Note that this gives an immediate estimate of the maximum noise variance, namely $\sigma^2 < \omega_N/b_-$; if the noise were any larger ω_N would be negative which contradicts the positive definiteness of $\hat{C}_{\mathbf{x}}$. This is, however, an overestimate of σ because the non-zero eigenvalues of AA^{T} increase ω_N away from η_N.

The use of the Hermitian properties of the covariance matrices (without exploiting their positive definiteness) leads to the prediction of negative λ_n, which was eliminated above by a prior which truncated the likelihood at zero. Here it leads to decreasing evidence for $\lambda_n = 0$ as σ^2 exceeds $\approx \omega_n$. The rate of this decrease may be too rapid because the model fails to account for the increasing probability that λ_n is exactly zero, but treats $\lambda_n = 0$ in exactly the same way as, for example, $\lambda_n = 0.1$. It would be possible to choose priors $\pi(\lambda_n)$ that assign a proportion of the probability mass to $\lambda_n = 0$; however, in the absence of *a priori* information it is difficult to choose the proportion to be assigned to the spike at zero and we refrain introducing additional hyper-parameters.

This point is illustrated in a second example, which was chosen to be difficult to the point of pathological. Two fragments ($T = 200$ samples, less than 1/50th second) of music were linearly mixed and Gaussian observational noise ($\sigma = 1$) added to form an observation sequence $\mathbf{x}(t)$ in $N = 20$ dimensions. The noise power, $Tr\hat{C}_{\mathbf{v}} = 19.75$ and the power of the noiseless signal was $Tr(T^{-1}\sum_t^T A\mathbf{s}(\mathbf{t})\mathbf{s}(\mathbf{t})^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}) = \mathbf{7.76}$. Also the vast majority of the signal power is represented by $\lambda_1 = 7.22$, while the other non-zero eigenvalue is well below the noise power, $\lambda_2 = 0.54$.

The eigenvalues λ_n and ω_n are shown in Figure 5a. Apart from the first eigenvalue, the spectrum is dominated by the noise. The abrupt drop in the true eigenvalues between λ_2 and $\lambda_3 = 0$ is obscured in the ω_n and it is difficult to tell by eye that the underlying rank is 2.

Näive application of the MDL criterion, based on a linear model with Gaussian distributed errors, suggests that the model order is 1.

Rajan and Rayner's scheme [7] suggests M = 13. The evidence (Figure 5b) is maximum when $\sigma = 0.87$ and $M = 7$. The reason for this overestimate of M and concomitant underestimate of σ is apparent from figures 5c and 5d, which show $\omega_n^{-1}\int_0^{\omega_n} p(\boldsymbol{\omega}|\,\lambda_n,\sigma)d\lambda_n$ and $p(\boldsymbol{\omega}|\,\lambda_n,\sigma)|_{\lambda_n=0}$ as functions of σ and n. The evidence for any σ and M is obtained (30) by multiplying together the values down the column corresponding to σ in figure 5c as far as M and then continuing

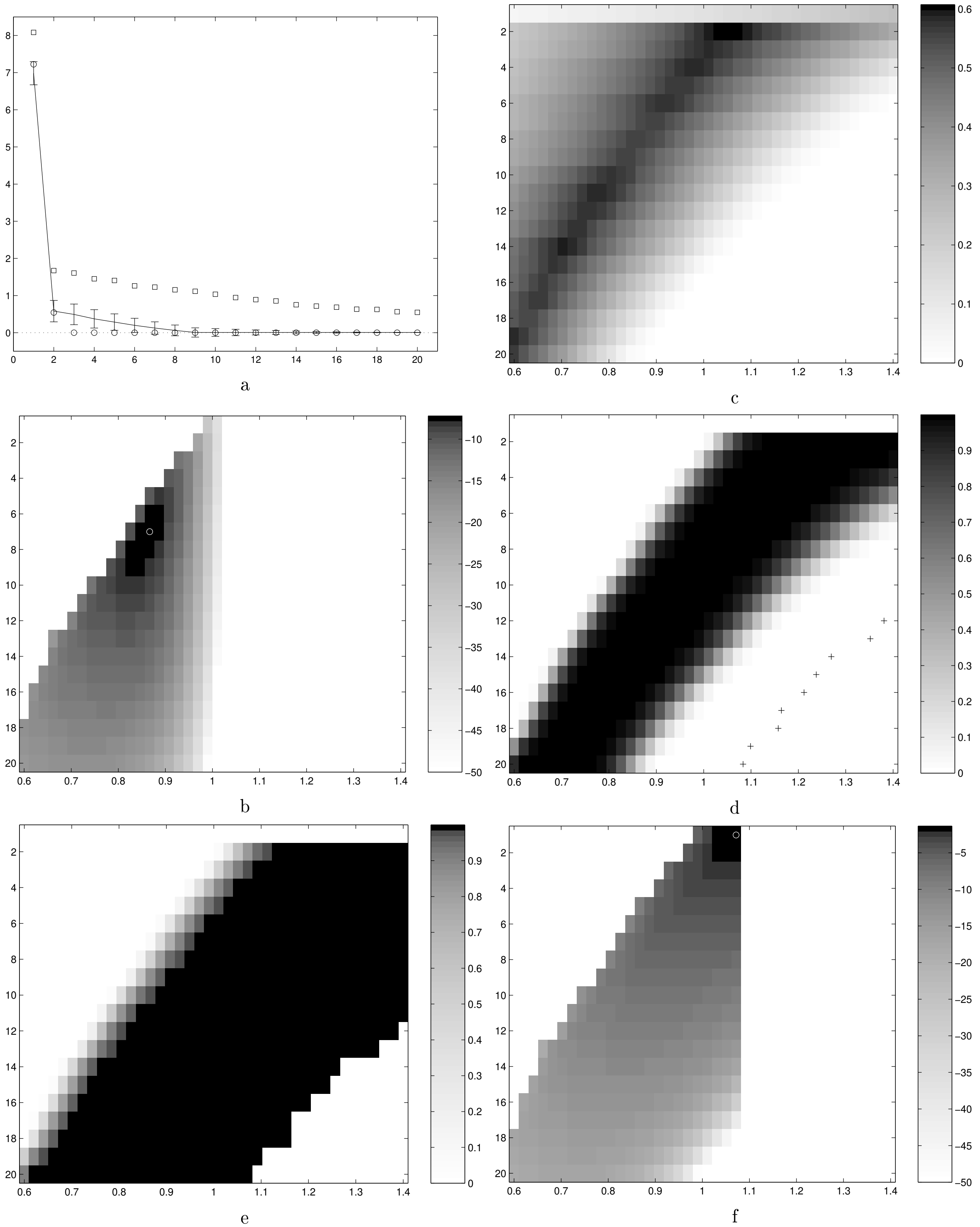


Fig. 5. Evidence for noise variance and model order for a mixture of two fragments of music mixed into 20 dimensions. **a:** True eigenvalues (λ_n, circles), noise-corrupted eigenvalues (ω_n, squares) and modes of posterior densities for $\sigma = 1$. The first eigenvalues ($\lambda_1 = 13.73$, $\omega_1 = 14.67$) are not plotted. **b:** Evidence $log_{10}p(\sigma, M|\,\boldsymbol{\omega})$. The maximum at $\sigma = 0.87$, $M = 8$ is indicated by a white circle. **c:** Evidence conditioned on λ_n. **d:** Evidence, $p(\boldsymbol{\omega}|\,\lambda_n, \sigma)|_{\lambda_n=0}$, that $\lambda_n = 0$. Crosses mark the locus $\omega_n - \sigma^2/b_- = 0$. **e:** Modified evidence, $\rho_n(\sigma)$, that $\lambda_n = 0$. **f:** Overall evidence, $log_{10}p(\sigma, M|\,\boldsymbol{\omega})$. The white circle indicates the maximum at $\sigma = 1.07, M = 1$.

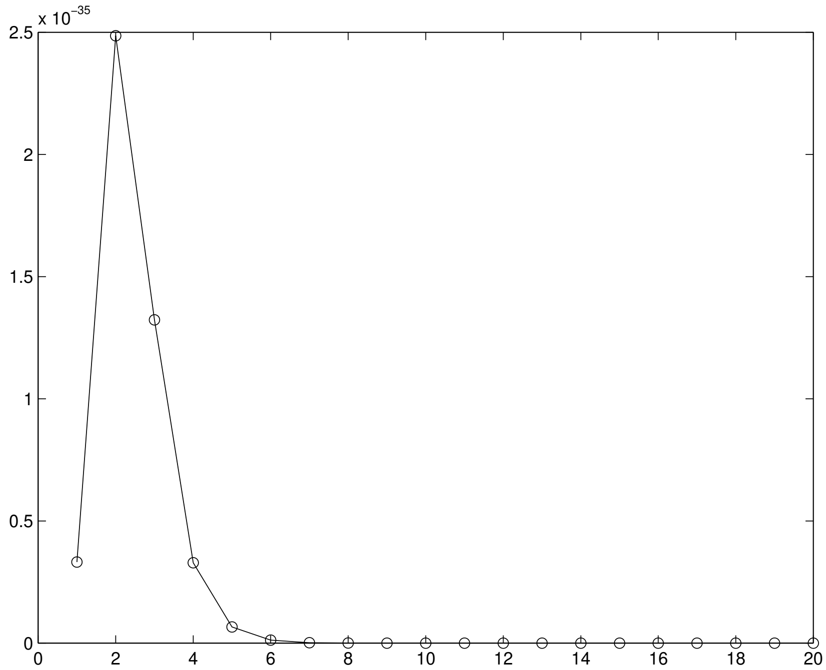


Fig. 6. Evidence for model order $p(M, \sigma | \boldsymbol{\omega})$ evaluated at $\sigma = 1.0061$.

down the corresponding column in figure 5d. As figure 5d illustrates, the evidence for $\lambda_{20} = 0$ is large when $\sigma \approx 0.7$, but is very small ($\approx 10^{-15}$) when $\sigma = 1$. In fact when $\sigma = 0.9$, for example, the evidence that $\lambda_{20} = 0$ is less than the evidence that $\lambda_{19} = 0$, even though we know that $\lambda_{19} \geq \lambda_{20}$.

Denoting the evidence that $\lambda_n = 0$ by $\rho_n(\sigma)$ $(n > M)$, we might expect that $\rho_n(\sigma) \geq \rho_{n-1}(\sigma)$ when $\sigma \leq \omega_n/b_-$ and $\rho_n(\sigma) = 0$ if $\sigma > \omega_n/b_-$. That is, at any noise variance we expect there to be at least as much evidence that $\lambda_n = 0$ as evidence that $\lambda_{n-1} = 0$, unless $\sigma^2 > \omega_n/b_-$ which is an infeasible value for σ. We might therefore model the evidence that $\lambda_n = 0$ as

$$\rho_n(\sigma) = \begin{cases} \max_{n' \leq n} p(\lambda_{n'} | \boldsymbol{\omega}, \sigma)|_{\lambda_{n'}=0} & \text{if } \sigma \leq \sqrt{\omega_n/b_-} \\ 0 & \text{if } \sigma > \sqrt{\omega_n/b_-}. \end{cases} \tag{32}$$

Figure 5e shows $\rho_n(\sigma)$, and figure 5f shows the overall evidence, which is maximum at $\sigma = 1.07$ and $M = 1$ – closer to the correct model order and variance. Here the fact that $\rho_n(\sigma)$ does not decrease with increasing σ, until $\sigma = \sqrt{\omega_n/b_-}$, when ρ_n drops abruptly to zero, means that M is selected at the largest feasible value of σ for ω_N: namely $\sigma = \sqrt{\omega_N/b_-}$. Since this is an upper bound for σ, it is generally an overestimate of the actual variance and consequently often leads to an underestimate of the rank M.

A better estimate of σ is obtained from a least squares fit of the tail of the observed spectrum to the Silverstein noise spectrum, η_n, assuming that the effect of λ_n on the tail of the spectrum is negligible. In this example a least-squares fit yields $\sigma = 1.0061$. The evidence for different model orders at this σ is plotted in figure 6 and correctly suggests that the rank is 2.

V. Discussion

We have presented a method for estimating the eigenvalues and hence the rank of a covariance matrix when the observed covariance matrix is heavily contaminated with noise and the number of data samples is limited. The Bayesian approach yields error bounds for the eigenvalues and the model order and noise variance may be estimated using the evidence.

Silverstein's expression for the eigenvalue density of sample covariance matrices is valid for all i.i.d. noise. Consequently this method is applicable to all sorts of zero mean noise, not just Gaussian noise.

We have concentrated on the regime $y = N/T < 1$. Apart from the $N - T$ zero eigenvalues in the covariance matrix the theory is unchanged when $N > T$, and we point out that the ω_n may be efficiently calculated from the $T \times T$ matrix $K_{pq} = \frac{1}{T}\mathbf{x}^{\mathsf{T}}(t_p)\mathbf{x}(t_q)$, which has the same eigenvalues as $\hat{C}_{\mathbf{x}}$ [16]. This method is also applicable to singular value decomposition (SVD) because the singular values of the data matrix are just the square roots of the ω_n.

Many data analysis methods (PCA, SVD, ICA) do not explicitly model the noise as in (5) but implicitly use the noiseless model (1). Exciting exceptions are Tipping and Bishop's "probabilistic PCA" [17] and EM formulations of ICA [18], [19]

Model order estimation and the blind estimation of noise variance are notoriously difficult problems and many methods have been developed to attack them. A novel feature of our approach is the explicit incorporation in the model of the number of data samples and the expected statistics of the noise. Methods which fail to model the number of data samples (such as näive MDL based on a linear generative model with Gaussian latent variables) perform poorly because the eigenvalues of the sample covariance matrix are not merely raised by σ^2 except in the limit of infinite data. Rajan and Rayner [7] also give a Bayesian scheme for SVD model order determination. They do not explicitly model the noise but assume that the projections onto the singular vectors with small singular values are dominated by noise. Zarowski's approach [20] is similar to ours in that he has modelled the singular values of noisy data by assuming that the singular values of the noise free data are perturbed noise drawn from *ad hoc* distributions. He then uses the MDL criterion to decide the rank of the noiseless data.

We have discussed a particular example in which this scheme has difficulty in correctly assessing the model order and noise variance. In this context model order estimation is equivalent to determining the number of eigenvalues that are exactly zero, and small changes in the estimated variance can lead to large changes in the model order. The principal obstacle here is the inadequacy of modelling the likelihood of a zero eigenvalue. Since the covariance matrices are positive semi-definite zero is a distinguished value, but it is not treated specially by the inequality relations between eigenvalues of general Hermitian matrices. Note that brute-force sampling approaches are computationally completely infeasible even for small problems. We have presented two methods (a modification of the evidence for $\lambda_n = 0$ and estimation of σ by least squares fitting) that improve the estimates. More robust results are obtained when the signal to noise ratio is larger or prior information exists about the variance or eigenvalue spectrum.

Finally, it is important to recognise that we have assumed that the noise and signal are uncorrelated. This assumption was made so that the cross term $2\sigma A \left\langle \mathbf{s(t)v(t)}^\mathsf{T} \right\rangle$ could be discarded (cf equation 6). While this is certainly true with many data samples, spurious correlations with few samples may affect ω_n. Indeed, since $2\sigma A \left\langle \mathbf{s(t)v(t)}^\mathsf{T} \right\rangle$ is indefinite the ω_n may be decreased. Also if the sources are not perfectly decorrelated $\left\langle A\mathbf{s(t)s(t)}^\mathsf{T}\mathbf{A}^\mathsf{T} \right\rangle$ may have larger rank than the actual number of sources.

Acknowledgements

We are grateful for discussions with Dave Denison, Dirk Husmeier, Will Penny, Iead Rezek and Amos Storkey. The comments of two anonymous referees are appreciated. Part of this work was supported by funding from British Aerospace plc, to whom the authors are most grateful.

Appendix

I. Appendix: Weyl inequalities

Here we quote Weyl's theorem relating the eigenvalues of two Hermitian matrices to the eigenvalues of the sum of the matrices. Proofs are given in [12], [13] for example.

Theorem (Weyl)

Let A, B be N by N Hermitian matrices and let the eigenvalues $\lambda_i(A)$, $\lambda_i(B)$ and $\lambda_i(A+B)$ be arranged in *decreasing* order ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$). Then for each j, k, n ($1 \leq k \leq n \leq j \leq N$) we have

$$\lambda_{N+n-j}(A) + \lambda_j(B) \;\leq\; \lambda_n(A+B) \;\leq\; \lambda_{n-k+1}(A) + \lambda_k(B) \tag{33}$$

Note that the statement of the theorem is apparently different from the usual statement because the eigenvalues are listed in decreasing order.

References

[1] H. Hotelling, "Analysis of complex statistical variables in principal components," *J. Educ Psy.*, vol. 24, pp. 417–441, 498–520, 1933.

[2] I.T. Joliffe, *Principal Component Analysis*, Springer-Verlag, 1986.

[3] K. Karhunen, "Zur Spektraltheorie Stochasticher," *Prozesse Ann. Acad. Sci. Fennicae*, vol. 37, 1946.

[4] M.M. Loève, *Probability Theory*, Van Nostrand, Princeton, N.J., 1955.

[5] A.J. Bell and T.J. Sejnowski, "An information maximization Approach to Blind Separation and Blind Deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[6] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[7] J.J. Rajan and P.J.W. Rayner, "Model order selection for the singular value decomposition and the discrete Karhunen-Loeve transform using a Bayesian approach," *IEE Proc. Vis. Image Signal Process.*, vol. 144, no. 2, pp. 116–123, 1997.

[8] T-W. Lee, M. Girolami, A.J. Bell, and T.J. Sejnowski, "A Unifying Information-theoretic Framework for Independent Component Analysis," *International Journal on Mathematical and Computer Modeling*, 1998, (In press). Available from `http://www.cnl.salk.edu/~tewon/Public/mcm.ps.gz`.

[9] R.M. Everson and S.J. Roberts, "ICA: A flexible non-linearity and decorrelating manifold approach," *Neural Computation*, vol. 11, no. 8, 1999, Available from `http://www.dcs.ex.ac.uk/academics/reverson`.

[10] J.W. Silverstein, "Eigenvalues and eigenvectors of large dimensional sample covariance matrices," *Contemporary Mathematics*, vol. 50, pp. 153–159, 1986.

[11] L. Sirovich and R.M. Everson, "Analysis and management of large scientific databases," *International Journal of Supercomputing Applications*, vol. 6, no. 1, pp. 50–68, 1992.

[12] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[13] B.N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[14] P.M. Lee, *Bayesian Statistics: An introduction*, Oxford University Press, 1989.

[15] C.S. Wallace and D.M. Boulton, "An information measure for classification," *Computer Journal*, vol. 11, no. 2, pp. 195–209, 1968.

[16] L. Sirovich, "Turbulence and the dynamics of coherent structures, pt i: Coherent structures, p t ii: Symmetries and transformations, pt. iii: Dynamics and scaling," *Quarterly of Applied Mathematics*, vol. XLV, no. 3, pp. 561–590, 1987.

[17] M.E. Tipping and C.M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[18] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Nerual Computation*, vol. 11, no. 2, pp. 305–346, 1999.

[19] H. Attias, "EM algorithms for independent components analysis," in *Neural Networks for Signal Processing VIII*, A. Costantinides, S.-Y. Kung, M. Niranjan, and E. Wilson, Eds. 1998, pp. 132–141, IEEE.

[20] C.J. Zarowski, "The MDL criterion for rank determination via effective singular values," *IEEE Trans. Sig. Proc.*, vol. 46, no. 6, pp. 1741–1744, 1998.