Journal of Machine Learning Research 9 (2008) 2733-2759 Submitted 10/07; Revised 7/08; Published 12/08

Automatic PCA Dimension Selection for High Dimensional Data and Small Sample Sizes

David C. Hoyle DAVID.HOYLE@MANCHESTER.AC.UK
North West Institute for BioHealth Informatics,
University of Manchester, Faculty of Medical and Human Sciences,
University Place (East), Oxford Rd., Manchester, M13 9PL, UK.

Editor: Chris Williams

Abstract

Bayesian inference from high-dimensional data involves the integration over a large number of model parameters. Accurate evaluation of such high-dimensional integrals raises a unique set of issues. These issues are illustrated using the exemplar of model selection for principal component analysis (PCA). A Bayesian model selection criterion, based on a Laplace approximation to the model evidence for determining the number of signal principal components present in a data set, has previously been show to perform well on various test data sets. Using simulated data we show that for d-dimensional data and small sample sizes, N, the accuracy of this model selection method is strongly affected by increasing values of d. By taking proper account of the contribution to the evidence from the large number of model parameters we show that model selection accuracy is substantially improved. The accuracy of the improved model evidence is studied in the asymptotic limit $d \to \infty$ at fixed ratio $\alpha = N/d$, with $\alpha < 1$. In this limit, model selection based upon the improved model evidence agrees with a frequentist hypothesis testing approach.

Keywords: PCA, Bayesian model selection, random matrix theory, high dimensional inference

1. Introduction

The generation of high dimensional data is fast becoming a common place occurrence. Examples range from genomics and molecular biology, for example high-throughput single nucleotide polymorphism (SNP) genotyping scans (Price et al., 2006) and microarray gene expression studies (Golub et al., 1999), to geophysical imaging, for example hyperspectral image data (Landgrebe, 2002). Intuitive visualization of the data and construction of novel features from the data are key tasks in processing such high-dimensional data. This often involves dimensionality reduction, for which a number of algorithms exist. Principal component analysis (PCA) is a ubiquitous method of data analysis and dimensionality reduction (Joliffe, 1986). Its utility and success stems from the simplicity of the method - one simply calculates the eigenvectors and eigenvalues of the sample covariance matrix $\hat{C}$ of the data set. A subset of the eigenvectors of $\hat{C}$, the principal components, are then selected to represent the data. A 'kernelized' version has been formulated - kernel PCA (Scholköpf, Smola, and Müller, 1998), and building on probabilistic formulations (Roweis, 1998; Tipping and Bishop, 1999a) it has also been extended to a mixture of principal component analysers (Tipping and Bishop, 1999b). In the latter case a number of local linear models are embedded in the high dimensional data space, with the properties of each local model being determined from the local responsibility-weighted covariance matrix.

Clearly selection of the correct number of principal components is crucial to the success of PCA in representing a data set. Identification of the appropriate signal dimensionality is just a model selection process to which the techniques of Bayesian model selection can be applied via a suitable approximation of the Bayesian evidence (MacKay, 1992). What is the most suitable method of approximating the evidence for high-dimensional data and what are the inherent problems? These are the research questions we address and a roadmap for the paper is given below,

• In Section 2 we motivate why high-dimensional small sample size data sets present a challenge for Bayesian model selection.

• In Section 3 we summarize the behaviour of the eigenvectors and eigenvalues of sample covariance matrices formed from high-dimensional small sample size data sets.

• In Section 4.1 we review the formalism of Bayesian model selection for PCA, and evaluate through simulation the model selection accuracy of an existing approximation to the Bayesian evidence.

• In Section 4.2 we develop an improved approximation to the Bayesian evidence specifically for high dimensional data.

• In Section 5 we evaluate the asymptotic properties of the improved approximation to the model evidence.

• In Section 6 the model selection performance of the improved approximation to the model evidence is compared with a frequentist hypothesis testing approach to model selection.

2. The Challenge of High-Dimensional Data for Bayesian Model Selection

A number of Bayesian formulations of PCA have followed from the probabilistic formulation of Tipping and Bishop (1999a), with the necessary marginalization being approximated through both Laplace approximations (Bishop, 1999a; Minka, 2000, 2001a) and variational bounds (Bishop, 1999b). More recently, work within the statistics research community has used a Bayesian variational approach to derive an explicit conditional probability distribution for the signal dimension given the data (Šmídl and Quinn, 2007). However, these results have only been tested on low dimensional data with relatively large sample sizes. A somewhat more tractable expression for the signal dimension posterior was also obtained by Minka (2000, 2001a) and it is that Bayesian formulation of PCA that we draw upon. By performing a Laplace approximation (Wong, 1989), that is, expanding about the maximum posterior solution, Minka derived an elegant approximation to the probability, the model evidence $p(D|k)$, of observing a data set D given the number of principal components k (Minka, 2000, 2001a). The signal dimensionality of the given data set is then estimated by the value of k that maximizes $p(D|k)$. As with any Bayesian model selection procedure, if the data has truly been generated by a model of the form proposed, then one is guaranteed to select the correct model dimensionality as the sample increases to an infinite size. Minka's dimensionality selection method performs well when tested on data sets of moderate size and dimensionality. Indeed, the Laplace approximation incorporates the leading order term in an asymptotic expansion of the Bayesian evidence, with the sample size N playing the role of the 'large' parameter, and so we would expect the Laplace approximation to be increasingly accurate as $N \to \infty$. In real-world data sets, such as those emanating from molecular biology experiments, the number of variables d is often very much greater than the sample size N, with $d \sim 10^4$ yet $N \sim 10$ or $N \sim 10^2$ not uncommon (Hoyle and Rattray, 2003). Typically, data sets with a sample size of $N = 100$ might be considered as large enough to be well approximated by the asymptotic limit $N \to \infty$, and therefore the Laplace

approximation to be appropriate. However, though retaining only a small number of terms from the asymptotic expansion of the evidence would be increasingly accurate as $N \to \infty$, individual expansion coefficients may be significant due to the large data dimensionality d. This suggests that for real finite sample size data sets, higher order terms in the asymptotic expansion not encapsulated within the Laplace approximation will make significant contributions to the evidence, and model selection based upon a simplistic application of the Laplace approximation will perform poorly. What then defines a 'large' sample size N is clearly dependent on the data dimensionality d. We would expect the conjectures about the previously derived Laplace approximation to the evidence to be increasingly true when the data dimensionality is very much larger than the sample size, that is, $N \ll d$, the situation encountered for many modern data sets. For high dimensional data, rather than considering the evidence to be close to its value obtained in the asymptotic limit $N \to \infty$ at fixed d, it may be more appropriate to consider the evidence as being close to its value in the distinguished limit $d, N \to \infty$ at fixed $\alpha = N/d$. Within this paradigm, developing a suitable Gaussian approximation requires us to identify all contributions to the evidence that would scale extensively, that is increase linearly with N, as $N, d \to \infty$ at fixed α. This would be increasingly important for $\alpha < 1$, where the contribution to the evidence resulting from many features can be significant. Ideally we should re-formulate the evidence as an integration over a set of variables which remains finite in number in the distinguished limit.

To be more explicit, consider that the Bayesian approach to model selection in PCA starts from the probability $p(D|k,\boldsymbol{\theta})p(\boldsymbol{\theta}|k)$ and integrates over the model parameters $\boldsymbol{\theta}$ to obtain the evidence $p(D|k)$. This integration is often evaluated by the aforementioned Laplace approximation - expansion about the maximum of $p(D|k,\boldsymbol{\theta})p(\boldsymbol{\theta}|k)$ and evaluation of the consequent tractable Gaussian integrals. For high-dimensional data the model parameters may consist of a small set of parameters, $\boldsymbol{\theta}_k$, of order of the signal dimensionality k, and a much larger set of parameters, $\boldsymbol{\theta}_d$, of order of the data dimensionality. For example, the latter may be the principal vectors, in the d-dimensional space, that form part of the model. Overall we can write $\boldsymbol{\theta} = (\boldsymbol{\theta}_d, \boldsymbol{\theta}_k)$. Integration over $\boldsymbol{\theta}_d$ provides a significant contribution to $p(D|k)$ due simply to the large number of individual model parameters that we are integrating over. In this scenario, the values of $\boldsymbol{\theta}_k$ obtained from maximizing $\int p(D|k,\boldsymbol{\theta}_d,\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_d,\boldsymbol{\theta}_k|k)d\boldsymbol{\theta}_d$ and $p(D|k,\boldsymbol{\theta}_d,\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_d,\boldsymbol{\theta}_k|k)$ do not coincide. In fact for large values of d they may be significantly different. The more accurate estimates of $\boldsymbol{\theta}_k$ are naturally obtained from the maximum of $\int p(D|k,\boldsymbol{\theta}_d,\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_d,\boldsymbol{\theta}_k|k)d\boldsymbol{\theta}_d$, and consequently the more accurate estimates of the evidence $p(D|k)$ are obtained by expanding about this maximum.

The distorting effects of high dimensionality upon covariance matrix eigenvalue spectra and eigenvectors are well known from random matrix theory (RMT) studies (Johnstone, 2006). The RMT studies inform us about the expected sample covariance eigenvalue spectrum in the limit $d \to \infty$ (at fixed α), and consequently the limits of any model selection procedure based upon the observed eigenvalue spectra. As PCA is based upon the eigenvalues and eigenvectors of $\hat{\boldsymbol{C}}$, understanding their behaviour for small sample sizes and high data dimensions is key to understanding the behaviour of the existing model selection criterion, including the Bayesian model selection approach of Minka. Results from RMT studies are summarized in Section 3.

3. High-Dimensional Sample Covariance Matrices

We envisage a scenario where one has N, d-dimensional data vectors $\boldsymbol{\xi}_\mu, \mu = 1, \ldots, N$, with sample mean $\bar{\boldsymbol{\xi}}$, which are drawn from a multi-variate Gaussian distribution with covariance $\boldsymbol{C}$. The eigen-

values of $\boldsymbol{C}$ we denote by Λ_i, $i = 1,\ldots,d$. The sample data vectors $\boldsymbol{\xi}_\mu$ contain both signal and noise components so we represent,

$$\boldsymbol{C} = \sigma^2 \boldsymbol{I} + \sum_{m=1}^{S} \sigma^2 A_m \boldsymbol{B}_m \boldsymbol{B}_m^T \quad , \quad \boldsymbol{B}_m^T \boldsymbol{B}_{m'} = \delta_{mm'} \quad , \quad A_m \geq 0 \; \forall m \; , \tag{1}$$

corresponding to a population covariance $\boldsymbol{C}$ that contains a small number, S, of orthogonal signal components, $\{\boldsymbol{B}_m\}_{m=1}^S$, but that is otherwise isotropic. Here, σ^2 represents the variance of the additive noise component of the sample data vectors. Such models have been termed "spiked" covariance models within the statistics research literature (Johnstone, 2001), due to the small number of δ-function spikes in the population covariance eigenspectrum. In this case the population eigenvalues are $\Lambda_i = \sigma^2(1+A_i)$, $i \leq S$ and $\Lambda_i = \sigma^2$, $i > S$. The signal strengths $\sigma^2 A_m$ merely determine the population covariance eigenvalues corresponding to signal directions, and so the number of signal components S is commonly estimated by some process of inspection of the ordered eigenvalues $\lambda_i, i = 1,\ldots,d$, of the sample covariance matrix $\hat{\boldsymbol{C}} = N^{-1}\sum_\mu (\boldsymbol{\xi}_\mu - \bar{\boldsymbol{\xi}})(\boldsymbol{\xi}_\mu - \bar{\boldsymbol{\xi}})^T$.

When the sample size is greater than the dimensionality, that is, $N > d$, the sample covariance eigenvalues λ_i may be reasonable estimators of the population covariance eigenvalues Λ_i, and indeed are asymptotically unbiased estimators, that is, $\lambda_i \to \Lambda_i$ as $N \to \infty$ for fixed dimensionality d (Anderson, 1963). However, for small sample sizes $N \leq d$ the sample covariance $\hat{\boldsymbol{C}}$ is singular with a $d - N + 1$ degenerate zero eigenvalue. Similarly, the non-zero sample covariance eigenvalues, λ_i, $i = 1,\ldots,N-1$, can display considerable bias. This is reflected in the expected eigenspectrum, $\rho(\lambda)$, which is simply defined as the expectation over data sets of the empirical eigenvalue density,

$$\rho(\lambda) \; = \; \mathrm{E}_{\boldsymbol{\xi}} \left(\frac{1}{d} \sum_{i=1}^{d} \delta(\lambda - \lambda_i) \right) \; .$$

Here $\delta(x)$ is the Dirac δ-function, and we have used $\mathrm{E}_{\boldsymbol{\xi}}(\cdot)$ to denote expectation over the ensemble of sample data sets. The empirical eigenvalue density is considered to be a self-averaging quantity, such that as $N \to \infty$ the eigenvalue density from any individual sample covariance matrix is well represented by the ensemble average. Therefore, for large sample covariance matrices studying the behaviour of the expected sample covariance eigenvalue distribution provides us with insight into the behaviour of individual sample covariance matrices and consequently the behaviour of any model selection algorithms based upon the sample covariance eigenvalues.

When no signal components are present, that is, $\boldsymbol{C} = \sigma^2 \boldsymbol{I}$, and in the limit $d \to \infty$ with $\alpha = N/d$ fixed, the expected distribution of sample eigenvalues tends to the Marčenko-Pastur distribution (Marčenko and Pastur, 1967),

$$\begin{aligned} \rho(\lambda) \; = \; \rho_{bulk}(\lambda) \quad &= \quad (1-\alpha)\Theta(1-\alpha)\delta(\lambda) \\ &+ \quad \frac{\alpha}{2\pi\lambda\sigma^2} \sqrt{\max[0, (\lambda - \lambda_{min})(\lambda_{max} - \lambda)]} \; , \end{aligned} \tag{2}$$

where $\lambda_{max} = \sigma^2(1+\alpha^{-\frac{1}{2}})^2$, $\lambda_{min} = \sigma^2(1-\alpha^{-\frac{1}{2}})^2$, and $\Theta(x)$ is the Heaviside step function. Figure 1 shows examples of the Marčenko-Pastur distribution for different values of α. It should be noted that although the mean sample eigenvalue is an unbiased estimator of σ^2, that is, $\int_{0^-}^{\infty} d\lambda \, \lambda \rho_{bulk}(\lambda) = \sigma^2$, the individual non-zero sample covariance eigenvalues lie in the interval $[\lambda_{min}, \lambda_{max}]$ and so for $\alpha < 1$ are highly biased estimators of the corresponding population eigenvalues.

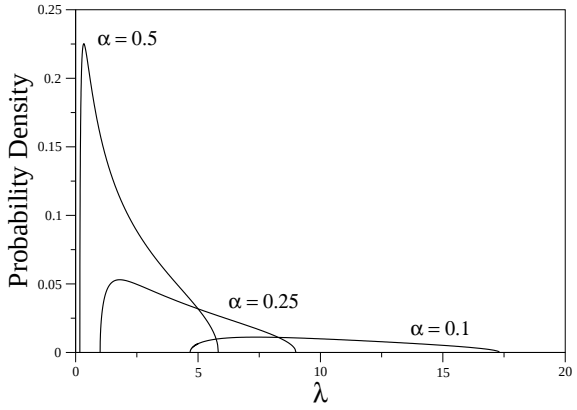

Figure 1: The Marčenko-Pastur limiting distribution for sample covariance eigenvalues, at $\alpha = 0.1, 0.25, 0.5$. In all cases $\sigma^2 = 1$. We have shown only the part of the distribution pertaining to non-zero eigenvalues. For $\alpha < 1$ there is also a δ-function peak at $\lambda = 0$ due to the singular nature of the sample covariance matrix - see main text.

Hoyle and Rattray (2004a) studied the expected behaviour of the sample covariance eigenvalue spectrum for "spiked" covariance models in the asymptotic limit $d \to \infty$ at fixed α, by using techniques from statistical physics. Similar results have been obtained within the statistics research community (Baik and Silverstein, 2006). As the addition of a small number, S, of signal directions provides a relatively small perturbation to an isotropic population covariance, the majority, or bulk of eigenvalues are still distributed according to the Marčenko-Pastur law. For this reason we have used $\rho_{bulk}(\lambda)$ to denote the Marčenko-Pastur distribution. For the "spiked" covariance models of Equation (1) the expected eigenvalue distribution $\rho(\lambda)$ is modified from $\rho_{bulk}(\lambda)$. At finite but large values of d and N the expected sample covariance eigenvalue density can be approximated by,

$$\begin{aligned}\rho(\lambda) &= (1-\alpha)\Theta(1-\alpha)\delta(\lambda) + \frac{1}{d}\sum_{m=1}^{S}\delta(\lambda-\lambda_u(A_m))\Theta(\alpha-A_m^{-2}) \\ &+ \left(1-d^{-1}\sum_{m=1}^{S}\Theta(\alpha-A_m^{-2})\right)\frac{\alpha}{2\pi\lambda\sigma^2}\sqrt{\max[0,(\lambda-\lambda_{min})(\lambda_{max}-\lambda)]}\ , \qquad (3)\end{aligned}$$

where $\lambda_u(A) = \sigma^2(1+A)(1+(\alpha A)^{-1})$. A number of interesting features are present in this spectrum. A transition occurs at $\alpha = A_m^{-2}$, such that for $\alpha > A_m^{-2}$ a sample eigenvalue located at $\lambda = \lambda_u(A_m)$ can be resolved separately from the remaining Marčenko-Pastur bulk of eigenvalues. Thus for S signal components within the "spiked" covariance model we can observe up to S transitions in the sample covariance eigenspectrum, on increasing α. The first transition point $\alpha = A_1^{-2}$ corresponds to the transition point in learning the leading signal direction $\boldsymbol{B}_1$. The scenario of learning a single signal component $\boldsymbol{B}_1$ of strength A_1 has been studied by Reimann et al. (1996), who

considered the behaviour (as $d \to \infty$ at fixed α) of the expectation value of R_1^2, where $R_1 = \boldsymbol{B}_1 \cdot \boldsymbol{J}_1$ is the overlap between the first principal component $\boldsymbol{J}_1$ of the sample covariance and $\boldsymbol{B}_1$. One observes the phenomenon of retarded learning whereby $R_1^2 = 0$ for $\alpha < A_1^{-2}$ and $R_1^2 > 0$ for $\alpha > A_1^{-2}$. This has been generalized to learning multiple orthogonal signals and one observes a separate retarded learning transition at $\alpha = A_m^{-2}$ for each of the overlaps $R_m^2 = (\boldsymbol{B}_m \cdot \boldsymbol{J}_m)^2$, where $\boldsymbol{J}_m$ is the m^{th} principal component (Hoyle and Rattray, 2007). That the ability to detect the signal components is reflected in the sample covariance eigenvalue structure (with retarded learning transitions coinciding with transitions in the eigenspectrum) demonstrates the utility of the sample covariance eigenspectrum for model selection. It also highlights that if the true signal dimensionality is S then asymptotically we have at most only S sample covariance eigenvalues separated from the Marčenko-Pastur bulk distribution, dependent on the value of α. If, for the given value of α, we have $\hat{S}$ eigenvalues separated from the Marčenko-Pastur bulk distribution, then the asymptotic equivalence of the observed sample covariance eigenspectra when $\boldsymbol{C}$ contains S signals or $\hat{S} \leq S$ signals means that no correct Bayesian model selection procedure can, asymptotically, select greater than $\hat{S}$ principal components (applying an Occam's Razor like argument), since both models are equally capable of explaining the observed eigenspectra. Equally, for sufficiently small α it is impossible, asymptotically, to distinguish the sample spectrum from one which has been generated from a model containing no signal structure, that is, from a population covariance $\boldsymbol{C} = \sigma^2 \boldsymbol{I}$. Within these constraints placed by the expected behaviour of the observed eigenspectra we now attempt to derive a suitable Bayesian model selection procedure that performs well in the distinguished asymptotic limit $N, d \to \infty$ at fixed α.

4. Bayesian Model Selection

In this section we summarize the Bayesian model selection procedure for PCA. We start in Section 4.1 by reproducing the formulation of the Bayesian model evidence as outlined by Minka (2000, 2001a) and the subsequent Laplace approximation. In Section 4.2 we re-express the evidence in a form that is more suitable for application of a Gaussian approximation when $d, N \to \infty$ at fixed $\alpha < 1$.

4.1 Laplace Approximation of Minka

The data vectors $\boldsymbol{\xi}_\mu$ are modelled as being drawn from a multi-variate Gaussian distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma} = v\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^T$. Thus $\boldsymbol{\Sigma}$ acts as a model of the true population covariance $\boldsymbol{C}$. The matrix $\boldsymbol{H}$ represents the signal considered present in the data and so is modelled as being due to a small number, k, of orthogonal signal components $\boldsymbol{u}_i$, $i = 1, \ldots, k$. Consequently we set,

$$\boldsymbol{H} = \boldsymbol{U}(\boldsymbol{L} - v\boldsymbol{I}_k)^{1/2}\boldsymbol{W} \ , \ \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_k \ , \ \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_k \ ,$$

where the columns of the orthonormal matrix $\boldsymbol{U}$ are formed from the vectors $\boldsymbol{u}_i$. The parameter v provides an estimator of the true population noise level σ^2. The diagonal matrix $\boldsymbol{L}$ has elements l_i, $i = 1, \ldots, k$, which represent estimators of the population covariance eigenvalues Λ_i. The orthonormal matrix $\boldsymbol{W}$ represents an irrelevant rotation within the subspace and is subsequently eliminated from the calculation. Model selection proceeds via the standard use of Bayes' theorem,

$$p(\boldsymbol{H}, \boldsymbol{m}, v|D) \ = \ \frac{p(D|\boldsymbol{H}, \boldsymbol{m}, v)p(\boldsymbol{H}, \boldsymbol{m}, v)}{p(D)} \ .$$

The signal dimensionality, k, is implicit in the matrix $\boldsymbol{H}$. With a non-informative prior, the mean $\boldsymbol{m}$ can be integrated out to yield the probability of observing the data set D given $\boldsymbol{H}$ and v (Minka, 2001a),

$$p(D|\boldsymbol{H},v) \;=\; N^{-d/2}(2\pi)^{-(N-1)d/2}|\boldsymbol{H}\boldsymbol{H}^T+v\boldsymbol{I}|^{-(N-1)/2}\exp\left(-\frac{N}{2}\mathrm{tr}((\boldsymbol{H}\boldsymbol{H}^T+v\boldsymbol{I})^{-1}\hat{\boldsymbol{C}})\right)\,.$$

Given a prior $p(\boldsymbol{U},\boldsymbol{W},\boldsymbol{L},v)$ the evidence for a signal dimensionality k is then,

$$p(D|k) \;=\; \int d\boldsymbol{U}d\boldsymbol{W}d\boldsymbol{L}dv\,p(D|\boldsymbol{U},\boldsymbol{W},\boldsymbol{L},v)p(\boldsymbol{U},\boldsymbol{W},\boldsymbol{L},v)\,.$$

The integration over the elements $l_i,\ i=1,\ldots,k$ is restricted to the region $l_i \geq 0 \forall i$. Similarly, the integration over $\boldsymbol{U}$ and $\boldsymbol{W}$ is over the entire space of $d\times k$ and $k\times k$ orthonormal matrices respectively. For the relevant integration over $\boldsymbol{U}$ this is equivalent to integration over the Stiefel manifold $V_k(\mathbb{R}^d)$ defined by the set of all orthonormal k-frames in $\mathbb{R}^d$ (James, 1954) .

Minka chooses a conjugate prior,

$$p(\boldsymbol{U},\boldsymbol{W},\boldsymbol{L},v) \;\propto\; |\boldsymbol{H}\boldsymbol{H}^T+v\boldsymbol{I}|^{-(\eta+2)/2}\exp(-\frac{\eta}{2}\mathrm{tr}((\boldsymbol{H}\boldsymbol{H}^T+v\boldsymbol{I})^{-1}))\,, \tag{4}$$

where the hyper-parameter η controls the sharpness of the prior. For a non-informative prior η should be small and ultimately we shall take $\eta\to 0^+$ in our resulting approximation to the evidence $p(D|k)$. With the prior given in Equation (4) the evidence is Minka (2000, 2001a),

$$\begin{aligned} p(D|k) \;&=\; \frac{\mathcal{N}_k(d)}{\mathrm{Area}(V_k(\mathbb{R}^d))}\int d\boldsymbol{U}d\boldsymbol{L}dv\,|\boldsymbol{H}\boldsymbol{H}^T+v\boldsymbol{I}|^{-(N+1+\eta)/2} \\ &\times\; \exp(-\frac{N}{2}\mathrm{tr}((\boldsymbol{H}\boldsymbol{H}^T+v\boldsymbol{I})^{-1}(\hat{\boldsymbol{C}}+N^{-1}\eta\boldsymbol{I})))\,, \end{aligned} \tag{5}$$

with,

$$\mathcal{N}_k(d) \;=\; \frac{N^{-d/2}(2\pi)^{-(N-1)d/2}}{\Gamma\left((\frac{1}{2}\eta+1)(d-k)-1\right)}\,(\eta(d-k)/2)^{(\frac{1}{2}\eta+1)(d-k)-1}\,\frac{1}{\Gamma(\eta/2)^k}\,(\eta/2)^{\eta k/2}\,,$$

and here $1/\mathrm{Area}(V_k(\mathbb{R}^d))$ is the reciprocal of the area of the Stiefel manifold $V_k(\mathbb{R}^d)$ (James, 1954),

$$\frac{1}{\mathrm{Area}(V_k(\mathbb{R}^d))} \;=\; 2^{-k}\prod_{i=1}^{k}\Gamma((d-i+1)/2)\pi^{-(d-i+1)/2}\,.$$

The dependence of $\mathcal{N}_k(d)$ upon k is relatively weak compared to other factors contributing to $\ln p(D|k)$, and so Minka drops $\mathcal{N}_k(d)$ from further consideration in approximating $p(D|k)$. As with the maximum likelihood case (Tipping and Bishop, 1999a), for a fixed choice, k, of the number of principal components, the maximum posterior estimators for $\{\boldsymbol{u}_i\}_{i=1}^k$ are known to be the eigenvectors of $\hat{\boldsymbol{C}}$ corresponding to the k largest eigenvalues of $\hat{\boldsymbol{C}}$. Minka approximates the evidence $p(D|k)$ in Equation (5) using a Laplace approximation, expanding about the maximum posterior solution. The stationary point values of v and $\{l_i\}_{i=1}^k$ are denoted by $\hat{v}$ and $\{\hat{l}_i\}_{i=1}^k$ respectively, and are given by (on taking $\eta\to 0$),

$$\hat{l}_i \;=\; \frac{N\lambda_i}{N-1} \;\simeq\; \lambda_i \quad,\quad \hat{v} \;=\; \frac{N\sum_{j=k+1}^{d}\lambda_j}{(N+1)(d-k)-2}\,. \tag{6}$$

Within this approximation $\hat{l}_i$ provides a point estimate of the i^{th} population covariance eigenvalue Λ_i. For $\alpha < 1$, as we have already commented in the previous section, λ_i can be highly biased and consequently a poor point estimate of Λ_i. Continuing with the Laplace approximation and setting $m = dk - k(k+1)/2$, Minka finds (again after taking $\eta \to 0$),

$$p(D|k) \simeq \frac{1}{\text{Area}(V_k(\mathbb{R}^d))} \left(\prod_{j=1}^{k} \lambda_j \right)^{-N/2} \hat{v}^{-N(d-k)/2} (2\pi)^{(m+k)/2} |\boldsymbol{A}_Z|^{-1/2} N^{-k/2} , \tag{7}$$

where,

$$|\boldsymbol{A}_Z| = \prod_{i=1}^{k} \prod_{j=i+1}^{d} (\hat{\Lambda}_j^{-1} - \hat{\Lambda}_i^{-1})(\lambda_i - \lambda_j) N .$$

The estimator $\hat{\Lambda}_i$ is given by $\hat{\Lambda}_i = \hat{l}_i \simeq \lambda_i$ for $i \leq k$ and $\hat{\Lambda}_i = \hat{v}$ for $i > k$.

Figure 2 shows simulation estimates of the performance of a model selection criterion based upon the evidence given by Equation (7). We have sampled data vectors $\boldsymbol{\xi}_\mu$ from a population covariance $\boldsymbol{C}$ containing three signal components. The noise level has been set to $\sigma^2 = 1$ and the signal strengths are $A_1^2 = 30, A_2^2 = 20, A_3^2 = 10$. The simulation results are averages evaluated over 1000 simulated data sets. Plotted in Fig.2(a) is the probability of selecting the correct model dimension against d, for different fixed values of N. As expected the accuracy of the model selection decreases with increasing d, with greater accuracy for larger sample sizes N at a given value of d. Plotted in Fig.2(b) is the probability of selecting the correct model dimension against d, for different fixed values of α. Note that the smallest value studied, $\alpha = 0.2$, is still greater than the retarded learning transition point of the weakest signal component, which occurs at $\alpha = A_3^{-2} = 0.1$.

The accuracy of the model selection procedure can potentially be improved by noting that PCA can simply be considered as constructing a representation of a matrix, in this case the mean centred sample data matrix. As such the transpose of the representation of the mean centred data matrix is equally as valid, which can be evaluated as the eigen-decomposition of the transpose of the mean centred data matrix. Given that we then model the transposed data matrix using k, N-dimensional vectors rather than k, d-dimensional vectors, then with $N < d$ and thus effectively lower model complexity, we would expect model selection based upon using the transposed mean centred data matrix to display superior accuracy. This is borne out by simulation results for model selection accuracy when applied to the transposed centred data matrix that are also shown in Fig.2. In all cases shown in Fig.2 the accuracy of the model selection is greater when using the transpose of the mean centred data matrix. One should note from Fig.2a, that even with transposing the centred data matrix, the model selection accuracy decreases with increasing data dimensionality d, at fixed sample size N. Taking a data set with $\alpha < 1$ and transposing does not produce an effective value of α that is larger than one - if true this would suggest one could have arbitrarily large effective values of α (by taking $d \to \infty$ at fixed N) and consequently asymptotically perfect model selection even though, as has already been highlighted, the expected spectrum in this limit is indistinguishable from that obtained by sampling from a distribution with an isotropic population covariance matrix. Consequently the accuracy of model selection based upon the sample covariance eigenspectrum will always decrease with increasing d, at fixed N, due to the distorting effects of high data dimensionality. We can attempt to mitigate these effects by taking proper account of the high dimensional contributions to the model evidence. This we do in the next section.

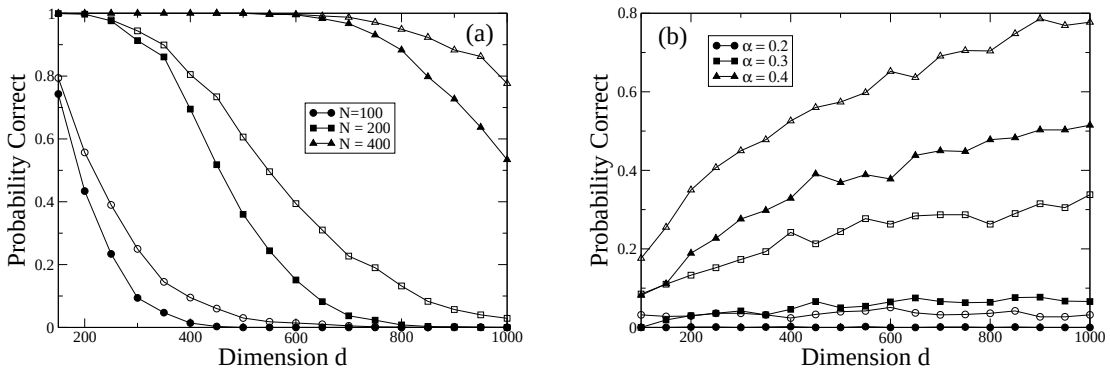

Figure 2: Probability of correct model selection using the method of Minka. The solid lines provide a guide to the eye. (a) & (b) Plots of model selection accuracy against data dimension d - (a) Fixed values of N, (b) Fixed values of α. The data is generated with a population covariance $\boldsymbol{C}$ containing three signal components -see main text for details. Solid symbols represent simulation results from the model selection procedure applied to the mean centred data matrix, whilst open symbols represent simulation results from the model selection procedure applied to the transpose of the mean centred data matrix.

4.2 Overlap Method

Although for $\alpha < 1$ the top k eigenvectors of $\hat{\boldsymbol{C}}$ are the maximum posterior choice of model principal components $\{\boldsymbol{u}_i\}_{i=1}^k$, for non-maximum posterior choices of $\boldsymbol{U}$ one still has a large rotational degeneracy of the k-frame within the d-dimensional space, which will make a large contribution to the integral in Equation (5). The integrand in Equation (5) can be written in terms of the overlaps $R_{ij} = \boldsymbol{u}_i \cdot \boldsymbol{v}_j$ between the model principal components $\boldsymbol{u}_i, i = 1,\ldots,k$, and the eigenvectors $\boldsymbol{v}_j$, $j = 1,\ldots,N-1$, of $\hat{\boldsymbol{C}}$ that correspond to the non-zero eigenvalues of $\hat{\boldsymbol{C}}$. One finds,

$$
\begin{aligned}
& |\boldsymbol{H}\boldsymbol{H}^T + \nu\boldsymbol{I}|^{-(N+1+\eta)/2} \exp(-\frac{N}{2}\mathrm{tr}((\boldsymbol{H}\boldsymbol{H}^T + \nu\boldsymbol{I})^{-1}(\hat{\boldsymbol{C}} + N^{-1}\eta\boldsymbol{I}))) \\
= \;& \exp\left[-\frac{N+1+\eta}{2}\left(\sum_{i=1}^{k} \ln l_i + (d-k)\ln\nu \right) - \frac{N}{2\nu}\sum_{j=1}^{N-1}\lambda_j \right. \\
+ \;& \left. \frac{N}{2}\sum_{i=1}^{k}(\nu^{-1} - l_i^{-1})\sum_{j=1}^{N-1}\lambda_j R_{ij}^2 \; - \; \frac{\eta d}{2\nu} \; + \; \frac{\eta}{2}\sum_{i=1}^{k}(\nu^{-1} - l_i^{-1}) \right] .
\end{aligned}
$$

This suggests performing the integration over $\{\boldsymbol{u}_i\}_{i=1}^k$ in terms of $\{R_{ij}\}$. The volume element that results from integrating over $\{\boldsymbol{u}_i\}_{i=1}^k$ at fixed $\{R_{ij}\}$ is $\det \boldsymbol{M}^{(d-N-1)/2} \times \mathrm{Area}(V_k(\mathbb{R}^{d-N+1}))$, where the matrix elements $M_{ii'} = \delta_{ii'} - \sum_j R_{ij}R_{i'j}$. For high dimensional spaces we might expect the vectors $\boldsymbol{u}_i, \boldsymbol{u}_{i'}$ to be orthogonal over any high-dimensional subspace, not just the entire d-dimensional space. Therefore we can approximate the matrix elements by $M_{ii'} = \delta_{ii'}(1 - \sum_j R_{ij}^2)$, and $\det \boldsymbol{M}$ is easily evaluated. With this approximation the evidence is,

$$p(D|k) \simeq \mathcal{N}_k(d)\frac{\text{Area}(V_k(\mathbb{R}^{d-N+1}))}{\text{Area}(V_k(\mathbb{R}^d))}\int\prod_{ij}dR_{ij}\int\prod_i dl_i\int dv$$
$$\times \quad \exp\left[\frac{d-N-1}{2}\sum_{i=1}^{k}\ln\left(1-\sum_{j=1}^{N-1}R_{ij}^2\right) - \frac{N+1+\eta}{2}\left(\sum_{i=1}^{k}\ln l_i + (d-k)\ln v\right)\right.$$
$$\left. - \frac{N}{2v}\sum_{j=1}^{N-1}\lambda_j + \frac{N}{2}\sum_{i=1}^{k}(v^{-1}-l_i^{-1})\sum_{j=1}^{N-1}\lambda_j R_{ij}^2 - \frac{\eta d}{2v} + \frac{\eta}{2}\sum_{i=1}^{k}(v^{-1}-l_i^{-1})\right] . \qquad (8)$$

Approximations to the model evidence can now be made by approximating this integration over the overlap variables $\{R_{ij}\}$, and consequently this approach is termed the "overlap" method. For large values of d and N we would expect the integral in Equation (8) to be dominated by the stationary points of the exponent and a Laplace approximation to the integral can be constructed. Denoting stationary point values by $\hat{v}, \hat{l}_i, \hat{R}_{ij}$, it is an easy matter to find that, on taking $\eta \to 0$, stationary points of Equation (8) satisfy for some j,

$$1-\hat{R}_{ij}^2 = \frac{(\hat{v}^{-1}-\hat{l}_i^{-1})N\lambda_j}{d-N-1} \quad , \quad \hat{R}_{ij'} = 0 \ , \ j' \neq j \, .$$

The dominant stationary point solution has the overlap between the i^{th} signal direction estimate, $\boldsymbol{u}_i$ and the i^{th} sample covariance eigenvector, $\boldsymbol{v}_i$, being non-zero, that is, $\hat{R}_{ii}^2 > 0$, $\hat{R}_{ii'}^2 = 0, \forall i \neq i'$,

For $j > k$ the dominant stationary point has $\hat{R}_{ij}^2 = 0$. Within this approximation the expectation value of R_{ij}^2 will be $O(N^{-1})$ due to small fluctuations about this stationary point. However, we have an extensive number, that is, proportional to N, of such overlap variables. Thus we expect $\sum_{j>k}R_{ij}^2 \sim 1$, and consequently the contribution from these small fluctuations cannot be ignored. The fluctuations in R_{ij}, for $j > k$, collectively affect the stationary point behaviour of the overlaps R_{ij} for $j \leq k$. To progress we integrate out the fluctuations by setting,

$$b_i = \sum_{j>k} R_{ij}^2 \, ,$$

and perform the integration over $\{R_{ij}\}_{j>k}$ by writing,

$$\int\prod_i\prod_{j>k}dR_{ij} = \int\prod_i\prod_{j>k}dR_{ij}\prod_i db_i\,\delta\left(b_i - \sum_{j>k}R_{ij}^2\right) .$$

Using the standard Fourier representation of a Dirac δ-function,

$$\delta(x) = \frac{1}{2\pi}\int_{-i\infty}^{i\infty}dp\,e^{px} \, ,$$

we obtain,

$$\int\prod_i\prod_{j>k}dR_{ij} = \frac{1}{(2\pi)^k}\int\prod_i db_i\,dp_i\prod_i\prod_{j>k}dR_{ij}\exp\left[\sum_i p_i\left(b_i - \sum_{j>k}R_{ij}^2\right)\right] , \qquad (9)$$

where the path of integration for p_i is between $-i\infty$ and $+i\infty$. Combining the integrand in Equation (9) with the integrand in Equation (8), the integration over $\{R_{ij}\}_{j>k}$ is Gaussian and so easily performed. We obtain,

$$\int dv \int \prod_{i=1}^{k} dl_i db_i dp_i \prod_i \prod_{j\leq k} dR_{ij} \exp\left(\sum_{i=1}^{k} p_i b_i \; + \; \frac{1}{2}(d-N-1)\sum_{i=1}^{k} \ln[1-\sum_{j=1}^{k} R_{ij}^2 - b_i] \right.$$
$$-\frac{1}{2}\sum_{i=1}^{k}\sum_{j>k} \ln[2p_i - N(v^{-1}-l_i^{-1})\lambda_j] \; - \; \frac{N+1}{2}\left[\sum_{i=1}^{k} \ln l_i \; + \; (d-k)\ln v\right]$$
$$\left. -\frac{N}{2}v^{-1}\sum_{j=1}^{N-1}\lambda_j \; + \; \frac{N}{2}\sum_{i=1}^{k}(v^{-1}-l_i^{-1})\sum_{j=1}^{k}\lambda_j R_{ij}^2 \right) . \tag{10}$$

With the path of integration for p_i being along the imaginary axis the remaining integrals in Equation (10) are approximated via steepest descent (Wong, 1989). For brevity we give only the solutions to the saddle point equations, with the caret again denoting saddle-point values of the corresponding integration variables,

$$\hat{v} \;=\; \frac{N}{(N+1)(d-k)}\left[\sum_{j=1}^{N-1}\lambda_j \; - \; \sum_{i=1}^{k}(1+N^{-1})\hat{l}_i\right] , \tag{11}$$

$$0 \;=\; \hat{l}_i^2\hat{v}^{-1}(1+N^{-1}) \; - \; \hat{l}_i(\lambda_i\hat{v}^{-1} - \alpha^{-1} + 1 + N^{-1}(k+3)) \; + \; \lambda_i \, , \tag{12}$$

$$\hat{R}_{ii}^2 \;=\; 1 \; - \; \frac{(d-N-1)}{N(\hat{v}^{-1}-\hat{l}_i^{-1})\lambda_i} \; - \; \frac{1}{N}\sum_{j>k}\frac{1}{(\hat{v}^{-1}-\hat{l}_i^{-1})(\lambda_i-\lambda_j)} \, , \tag{13}$$

$$\hat{R}_{ij}^2 \;=\; 0 \quad , j\neq i \, , \; j\leq k \, ,$$

$$\hat{p}_i \;=\; \frac{N}{2}(\hat{v}^{-1}-\hat{l}_i^{-1})\lambda_i \, , \tag{14}$$

$$\hat{b}_i \;=\; 1 \; - \; \hat{R}_{ii}^2 \; - \; \frac{(d-N-1)}{N(\hat{v}^{-1}-\hat{l}_i^{-1})\lambda_i} \, . \tag{15}$$

Again the saddle-point solution values $\hat{v}$ and $\hat{l}_i$ provide us with point estimates for the population noise level σ^2 and population signal eigenvalue Λ_i respectively. Equations (11) and (12) can be solved efficiently via an iterative process starting from an initial estimate of $\hat{v} = d^{-1}\sum_j \lambda_j$. Obtaining real-valued estimates, $\hat{l}_i$, for the population covariance eigenvalues is clearly dependent upon the quadratic equation in (12) having a non-negative discriminant. In practice, we have interpreted complex-valued estimates $\hat{l}_i$ for a particular choice of signal dimensionality k as meaning that the particular choice for k is not appropriate and should not be considered. From analysis of the asymptotic behaviour of the "overlap" approximation (see next section) we find that the discriminant of Equation (12) becomes negative for sample covariance eigenvalues λ_i which are below the edge of the Marčenko-Pastur bulk distribution given in Equation (2), that is, $\lambda_i < \lambda_{max} = \sigma^2(1+\alpha^{-\frac{1}{2}})^2$, so that indeed a negative discriminant is consistent with attempting to extract more signal components than can be genuinely distinguished from an isotropic population covariance. In other words complex solutions to Equation (12) suggest that the data do not support a model with that number, k, of signal components.

Once solutions for $\hat{v}$ and $\{\hat{l}_i\}_{i=1}^k$ have been obtained, values for $\hat{R}_{ii}^2, \hat{p}_i, \hat{b}_i$ follow from Equations (13), (14) and (15) respectively. Following Minka (2001a) and dropping the relatively weak k-dependence in $\mathcal{N}_k(d)$ we derive an approximation for the log-evidence as,

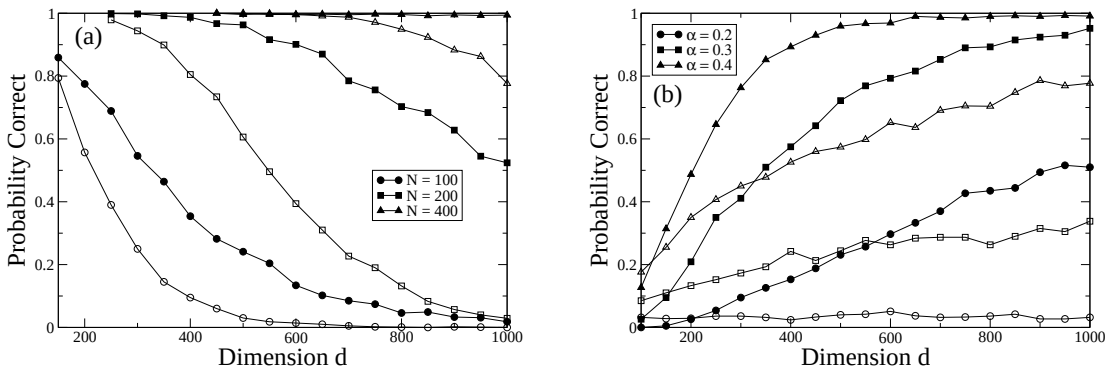

Figure 3: Plot of model selection accuracy for the "overlap" method. (a)Plot of model selection accuracy against data dimension d at fixed values of N. (b)Plot of model selection accuracy against data dimension for fixed values of α. For comparison open symbols represent simulation results from the model selection procedure of Minka applied to the transpose of the mean centred data matrix.

$$\begin{aligned}
\ln p(D|k) \;\simeq\; & \frac{N}{2}\sum_{i=1}^{k}(\hat{v}^{-1}-\hat{l}_i^{-1})\lambda_i \;-\; \frac{k}{2}(d-N-1) \;+\; k\frac{d-N-1}{2}\ln\left(\frac{d-N-1}{N}\right) \\
- \;& \frac{d-N-1}{2}\sum_{i=1}^{k}\ln((\hat{v}^{-1}-\hat{l}_i^{-1})\lambda_i) \;-\; \frac{k}{2}(N-k)\ln N \;-\; \frac{N-k}{2}\sum_{i=1}^{k}\ln(\hat{v}^{-1}-\hat{l}_i^{-1}) \\
- \;& \frac{1}{2}\sum_{i=1}^{k}\sum_{j>k}\ln(\lambda_i-\lambda_j) \;-\; \frac{N+1}{2}\sum_{i=1}^{k}\ln\hat{l}_i \;-\; \frac{N+1}{2}(d-k)\ln\hat{v} \;-\; \frac{N}{2}\hat{v}^{-1}\sum_{j=1}^{N-1}\lambda_j \\
+ \;& \frac{k}{2}(N-k-1)\ln 2\pi \;+\; \ln\left(\frac{\mathrm{Area}(V_k(\mathbb{R}^{d-N+1}))}{\mathrm{Area}(V_k(\mathbb{R}^{d}))}\right) \;-\; \frac{1}{2}\ln\det \boldsymbol{H}_s \\
+ \;& \frac{3k+k^2+1}{2}\ln 2\pi\,, \qquad (16)
\end{aligned}$$

where $\boldsymbol{H}_s$ is the Hessian of the exponent in the integrand evaluated at the saddle point. The last two terms in (16) come from integrating over the small fluctuations about the saddle point. Since the Hessian is of small dimension, and so not strongly dependent on N and d, we subsequently drop the last two terms from our approximation of the log-evidence. The "overlap" approximation to the log-evidence, given in Equation (16), can be used for model selection by selecting the value of k that has the highest value of $\ln p(D|k)$.

Figure 3 shows simulation estimates of the accuracy of the "overlap" model selection criterion given in Equation (16). Fig.3(a) shows the probability of selecting the correct model dimension against d, for different fixed values of N. Plotted in Fig.3(b) is the probability of selecting the correct model dimension against d, for different fixed values of α. Sample sizes and model pa-

rameter values are identical to those in Figure 2. Also reproduced (open symbols) in Fig.3(a) and Fig.3(b) are the simulation estimates of model selection accuracy for Minka's approximation to the model evidence applied to the transposed mean centred data. From Fig.3(a) it is clear that the "overlap" model selection criterion only suffers from degradation in performance at significantly higher values of dimension d compared to the approximation to the evidence in Equation (7). Similarly, Fig.3(b) demonstrates the superior model selection accuracy of the "overlap" method for increasing dimensionality d, at fixed values of α.

5. Asymptotic Analysis

The "overlap" approximation to the model evidence has been developed by applying a steepest descent approximation to the Bayesian evidence that has been re-formulated in terms of integration over variables that remain finite in number in the distinguished asymptotic limit $d, N \to \infty$, at fixed α. The "overlap" approximation essentially contains the leading order term of an asymptotic expansion of the evidence in that distinguished limit. It would be expected that the approximation to the model evidence would therefore become increasingly accurate in this limit. Note that this is very different from the traditional large sample limit $N \to \infty$ at fixed d, for which Minka's approximation to the Bayesian evidence will become increasingly accurate. It has been argued that since for many real high-dimensional data sets $\alpha \ll 1$, one would expect that approximations to the model evidence that are accurate in the distinguished limit will have superior model selection accuracy at finite values of d, N. The simulation results presented in Fig.3 would appear to confirm this. However, more concrete understanding of the accuracy of the "overlap" method in the distinguished asymptotic limit is required. A theoretical analysis of model selection accuracy in this limit would provide us with a firmer comparison of Minka's original Laplace approximation and the "overlap" method, in addition to the comparison provided by simulation study in Section 4.2. A number of quantities such as the eigenvalue spectrum are self-averaging in the asymptotic limit, that is, have vanishing sampling variation, so that for large data dimensions, d, the value for a single data set, $\{\boldsymbol{\xi}_\mu\}$, is well approximated by the ensemble average over data sets. Studying the ensemble expectation, in the asymptotic limit of $d \to \infty$ at fixed α, of the "overlap" approximation to the model evidence provides us with insight into its accuracy as a model selection procedure for high dimensional data.

From Equation (11) it is evident that $\hat{v} = d^{-1} \sum_{j=1}^{N-1} \lambda_j + O(N^{-1})$ as $N \to \infty$. Consequently, due to the self-averaging nature of the sample covariance eigenvalue spectrum, we have that $\hat{v} \to \mathrm{E}_{\boldsymbol{\xi}}(\lambda)$ as $N \to \infty$, where we have used $\mathrm{E}_{\boldsymbol{\xi}}(\cdot)$ to denote expectation over the ensemble of sample data sets. We already commented in Section 3 that $\mathrm{E}_{\boldsymbol{\xi}}(\lambda) = \sigma^2$ in the asymptotic limit $N \to \infty$ at fixed α, and so $\hat{v}$ provides an asymptotically unbiased estimate of the population noise level. Estimates of the population signal eigenvalues are given by $\{\hat{l}_i\}_{i=1}^k$, and in the distinguished asymptotic limit solutions to Equation (12) for $\hat{l}_i$ are given by,

$$\hat{l}_i = \frac{\hat{v}}{2}\left[(1+\lambda_i\hat{v}^{-1}-\alpha^{-1}) \pm \sqrt{(1+\lambda_i\hat{v}^{-1}-\alpha^{-1})^2 - 4\lambda_i\hat{v}^{-1}}\right] . \tag{17}$$

If we consider a "spiked" population covariance model of the form in Equation (1) the population covariance eigenvalues correspond to signal eigenvalues $\Lambda_i = \sigma^2(1+A_i)$, $i \leq S$ and noise eigenvalues $\Lambda_i = \sigma^2$, $i > S$. The resulting expected sample covariance eigenspectrum is given in Equation (3). Taking those sample eigenvalues which are separated from the bulk and also those at the upper bulk edge and substituting into Equation (17) we obtain on setting $\hat{v} = \sigma^2$ (on taking the positive

solution branch),

$$\begin{aligned}\hat{l}_i &= \sigma^2(1+A_i) \quad ,\text{for } \lambda_i = \sigma^2(1+A_i)(1+(1/\alpha A_i))\,,\\ \hat{l}_i &= \sigma^2(1+\alpha^{-\frac{1}{2}}) \quad ,\text{for } \lambda_i = \sigma^2(1+\alpha^{-\frac{1}{2}})^2\,.\end{aligned}$$

For sample covariance eigenvalues that are below the edge of the Marčenko-Pastur bulk distribution, that is, $\lambda_i < \sigma^2(1+\alpha^{-\frac{1}{2}})^2$, we obtain only complex solutions from Equation (17). Conversely, when $\lambda_i = \sigma^2(1+A_i)(1+(1/\alpha A_i))$, that is, when the sample covariance spectrum displays eigenvalues which are distinct from the bulk of the distribution, the estimator $\hat{l}_i = \sigma^2(1+A_i) = \Lambda_i$ and so gives an asymptotically unbiased estimate of the population signal eigenvalue Λ_i.

What is the asymptotic behaviour of the log-evidence? Inspecting Equation (16) we can see that, potentially, we need to evaluate $O(N^{-1})$ contributions to $\mathrm{E}_{\boldsymbol{\xi}}(\hat{\nu})$. However, it is easily shown that $O(N^{-1})$ contributions to $\mathrm{E}_{\boldsymbol{\xi}}(\hat{\nu})$ cancel out when evaluating $\mathrm{E}_{\boldsymbol{\xi}}(\ln p(D|k))$, and so we do not pursue them further here. We can evaluate the ensemble average $\mathrm{E}_{\boldsymbol{\xi}}(\sum_{j>k}\ln(\lambda_i-\lambda_j))$ through use of the replica trick (see Appendix A). Specifically we have for $\alpha > A_i^{-2}$,

$$\lim_{N,d\to\infty} N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right) \;=\; \ln\sigma^2 \;-\; (\alpha^{-1}-1)\ln(1+A_i) \;+\; \alpha^{-1}\ln A_i \;+\; \frac{1}{\alpha A_i}\,, \qquad (18)$$

whilst for $\alpha < A_i^{-2}$ we have,

$$\lim_{N,d\to\infty} N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right) \;=\; \ln\sigma^2 \;-\; (\alpha^{-1}-1)\ln(1+\alpha^{-\frac{1}{2}}) \;+\; \alpha^{-1}\ln\alpha^{-\frac{1}{2}} \;+\; \alpha^{-\frac{1}{2}}\,. \quad (19)$$

The asymptotic behaviour of the ratio $\mathrm{Area}(V_k(\mathbb{R}^{d-N+1}))/\mathrm{Area}(V_k(\mathbb{R}^d))$ is easily evaluated to give,

$$\ln\left(\frac{\mathrm{Area}(V_k(\mathbb{R}^{d-N+1}))}{\mathrm{Area}(V_k(\mathbb{R}^d))}\right) \;=\; \frac{Nk}{2}\left[-\ln\pi \;+\; \ln\frac{d}{2} \;-\; (\alpha^{-1}-1)\ln(1-\alpha) \;-1\right] \;+\; O(\ln N)\,. \quad (20)$$

Substituting Equations (18),(19),(20) and the asymptotic values for $\hat{\nu}$ and $\hat{l}_i$ into Equation (16), we obtain after some straight-forward algebra,

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\xi}}(\ln p(D|k)) &= \frac{N}{2}\sum_{i=1}^{k}\Theta(\alpha-A_i^{-2})\left[A_i-\frac{1}{\alpha A_i} \;+\; (\alpha^{-1}-1)\ln\left(\frac{1+A_i}{1+(1/\alpha A_i)}\right) \;+\; \alpha^{-1}\ln\left(\frac{1}{\alpha A_i^2}\right)\right]\\ &- \frac{Nd}{2}\ln\sigma^2 \;-\; \frac{Nd}{2} \;+\; O(\ln N)\,. \qquad (21)\end{aligned}$$

If we set $x = \alpha^{-1}$ we can write the summand in Equation (21) as $\Theta(\alpha - A_i^{-2})\,f(x,A_i)$ where,

$$f(x,A) = A \;+\; x\ln x \;+\; (x-1)\ln(1+A) \;-\; 2x\ln A \;-\; (x-1)\ln(1+xA^{-1}) \;-\; xA^{-1}\,.$$

We find that,

$$f(A^2,A) \;=\; 0 \quad , \quad \left.\frac{\partial f}{\partial x}\right|_{x=A^2} \;=\; 0 \quad , \quad \frac{\partial^2 f}{\partial x^2} \;>\; 0 \;\text{ for } x \;<\; A^2\,,$$

and so for $\alpha > A_i^{-2}$ the summand in Equation (21) is positive. Consequently if $\alpha > A_i^{-2}$, so that the sample covariance eigenvalue spectrum reflects the presence of the signal $\boldsymbol{B}_i$, then the addition of

i^{th} principal component results in an increase in the asymptotic approximation to the log-evidence. Conversely if $\alpha < A_i^{-2}$ there is no change in the asymptotic approximation to the log-evidence on including the i^{th} principal component. This is a satisfying result since we have already commented in Section 3 that for $\alpha < A_i^{-2}$ the sample covariance eigenspectrum is asymptotically indistinguishable from that produced from a population model with $A_i \equiv 0$, and so therefore from a Bayesian model selection perspective all population models with $A_i < \alpha^{-\frac{1}{2}}$ are equally likely (provide an equally accurate description of the observed data). Ultimately this is due to the fact that we are considering models with a finite number, k, of signal components, and so in the asymptotic limit we are considering a vanishingly small proportion of sample covariance eigenvalues as representing signal components. With the non-zero sample covariance eigenvalues giving a dense covering of the range $[\lambda_{min}, \lambda_{max}]$ in the asymptotic limit, the largest few sample covariance eigenvalues, which are not distinct from the Marčenko-Pastur bulk distribution given in Equation (2) will be aggregated at the upper edge of the bulk, where they do not lead to any change in the log-evidence. For finite sample sizes we would expect the higher order terms in the expansion of the log-evidence to lead to a decrease in the log-evidence on inclusion of principal components that correspond to sample covariance eigenvalues that are below the bulk edge. However, in the asymptotic limit we can apply an Occam's Razor like argument and only select those principal components that increase the log-evidence. The limiting model selection estimate, $\hat{S}$, for the true signal dimensionality, S, then simply corresponds to counting the number of sample covariance eigenvalues that are beyond the upper edge of the Marčenko-Pastur bulk distribution. That is,

$$\hat{S} \;=\; \sum_{j=1}^{d} \Theta(\lambda_j \;-\; \lambda_{max}) \,.$$

The asymptotic analysis of the "overlap" method reveals that unbiased estimates of the population signal eigenvalues can be recovered and that, asymptotically, model selection based upon the "overlap" approximation to the log-evidence performs optimally. From Fig.2b it would appear that, at least for larger values of α, model selection based upon Minka's approximation to the log-evidence also approaches 100% accuracy as $d \to \infty$. Is it possible that the two different approximations to the log-evidence asymptotically have the same model selection performance? Starting from Minka's approximation to the Bayesian evidence $p(D|k)$ given in Equation (7) we have,

$$\begin{aligned} \ln p(D|k) \;\;\simeq\;\; & -\ln \text{Area}(V_k(\mathbb{R}^d)) \;-\; \frac{N}{2}\sum_{i=1}^{k} \ln \lambda_i \;-\; \frac{N}{2}(d-k)\ln \hat{\nu} \;+\; \frac{m+k}{2}\ln 2\pi \\ & -\frac{1}{2}\sum_{i=1}^{k}\sum_{j=i+1}^{d}\Big[\ln(\hat{\Lambda}_j^{-1} - \hat{\Lambda}_i^{-1}) \;+\; \ln(\lambda_i - \lambda_j) \;+\; \ln N\Big] \;-\; \frac{k}{2}\ln N \,, \end{aligned} \tag{22}$$

where $\hat{\Lambda}_i = N\lambda_i/(N-1)$ for $i \leq k$ and $\hat{\Lambda}_i = \hat{\nu}$ for $i > k$, with $\hat{\nu}$ defined in Equation (6). In this instance $O(N^{-1})$ contributions to $\text{E}_{\boldsymbol{\xi}}(\hat{\nu})$ do make a contribution to the leading order asymptotic term in $\text{E}_{\boldsymbol{\xi}}(\ln p(D|k))$. From the definition of the point estimate $\hat{\nu}$ in (6) we find,

$$\text{E}_{\boldsymbol{\xi}}(\hat{\nu}) \;=\; (1 + N^{-1}(k\alpha - 1))\text{E}_{\boldsymbol{\xi}}(d^{-1}\text{tr}\hat{\boldsymbol{C}}) \;-\; \frac{\alpha}{N}\sum_{j=1}^{k}\text{E}_{\boldsymbol{\xi}}(\lambda_j) \;+\; O(N^{-2}) \,.$$

For the "spiked" covariance model of Equation (1) this can then be refined to,

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\xi}}(\hat{v}) &= \sigma^2 + \frac{\alpha\sigma^2}{N}\sum_{j=1}^{S} A_j + \frac{\sigma^2}{N}(k\alpha - 1) \\ &\quad - \frac{\alpha\sigma^2}{N}\sum_{i=1}^{k}\left[\Theta(\alpha - A_i^{-2})(1+A_i)(1+(1/\alpha A_i)) + \Theta(A_i^{-2}-\alpha)(1+\alpha^{-\frac{1}{2}})^2\right] \\ &\quad + O(N^{-2})\,.\end{aligned}$$

Retaining only k-dependent terms, the leading order asymptotic contribution to $\mathrm{E}_{\boldsymbol{\xi}}(\ln p(D|k))$ can be obtained within this approximation as,

$$\mathrm{E}_{\boldsymbol{\xi}}(\ln p(D|k)) = \frac{N}{2}\sum_{i=1}^{k}\left[\Theta(\alpha - A_i^{-2})\, f_M(x,A_i) + \Theta(A_i^{-2}-\alpha)\, f_M(x,\alpha^{-\frac{1}{2}})\right] + O(\ln N)\,,$$

where the subscript M on the function $f_M(x,A)$ is used to denote the asymptotic incremental change to the log-evidence obtained from Minka's approximation given in Equation (22), and again $x = \alpha^{-1}$. Specifically $f_M(x,A)$ is given as,

$$f_M(x,A) = A + x\ln x + (x-1)\ln(1+A) - 2x\ln A - x\ln\left[1 + xA^{-1} + xA^{-2}\right]\,. \qquad (23)$$

The transition point at which a signal component is strong enough to be distinguishable from the Marčenko-Pastur bulk distribution in Equation (2) is given by a signal strength $A = \alpha^{-\frac{1}{2}}$. If we put $A = y\alpha^{-\frac{1}{2}} = y\sqrt{x}$, then y directly measures the signal strength relative to that at which it is first detectable. We can then write Equation (23) as,

$$f_M(x, A = y\sqrt{x}) = (x-1)\ln(1+y\sqrt{x}) - x\ln\left(1 + y\sqrt{x} + y^2\right) + y\sqrt{x}\,.$$

A plot of $f_M(x = \alpha^{-1}, A = y\alpha^{-\frac{1}{2}})$ against y for different fixed values of α is shown in Figure 4. From Fig.4 we can see that at the transition point, $y = 1$, f_M is negative, and so selection of the i^{th} principal component will result in a reduction of the log-evidence, even if the signal strength A_i is sufficiently strong enough for the i^{th} sample covariance eigenvalue to be distinct from the Marčenko-Pastur bulk distribution. Thus, even though a detectable signal is present model selection based upon Equation (22) would not include that signal component. For the largest value of α shown f_M does not become positive until approximately $y > 1.8$. Therefore, even for $\alpha = 0.9$, not until the signal strength A_i is 1.8 times stronger than it need be for detection will the i^{th} signal component be correctly selected whilst using Minka's approximation to the log-evidence in Equation (22). For smaller values of α even stronger signal strengths are required, for example, $y > 2.0$ at $\alpha = 0.1$. For the simulations results shown in Fig.2b it is only at the largest value of α shown that we have $f_M > 0$ for all three signal components, and thus that all three signal components are guaranteed to be detectable in the asymptotic limit.

6. Comparison with Frequentist Approaches

In the distinguished asymptotic limit $N, d \to \infty$ the model selection process based upon the "overlap-method"approximation to the log-evidence simplifies (after applying a Occam's Razor like argument) to retaining those principal components whose corresponding eigenvalues are greater than

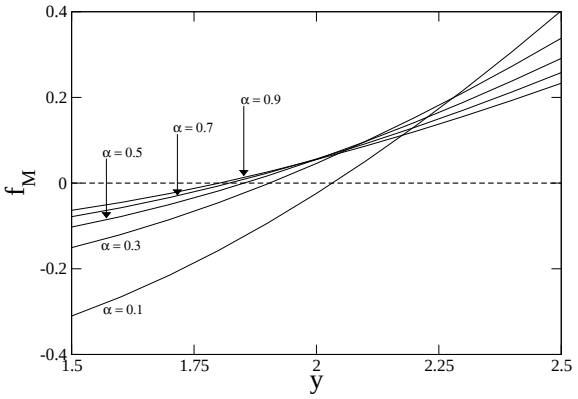

Figure 4: Plot of the function $f_M(x=\alpha^{-1}, A=y\alpha^{-\frac{1}{2}})$ against y for different values of α. $N f_M/2$ represents the incremental change (to leading order) in the log-evidence on retaining a principal component corresponding to a signal component of strength $A=y\alpha^{-\frac{1}{2}}$. The horizontal dashed line denotes the zero level for f_M.

the upper spectral edge, $\lambda_{max}=\sigma^2(1+\alpha^{-\frac{1}{2}})^2$, of the bulk eigenvalue distribution. Whilst this result appears intuitive from the viewpoint of the behaviour of eigenspectra of large sample covariance matrices presented in Section 3, we have also shown that not all approximations to the Bayesian evidence reduce in the asymptotic limit to this optimal choice for model selection. How then does the "overlap" method for model selection compare to other approaches, for example more traditional non-Bayesian approaches for dimensionality selection in PCA? In the asymptotic limit $N\rightarrow\infty$, where we have an infinite amount of data, we would naively expect frequentist and correctly formulated Bayesian approaches to model selection to give similar answers.

One of the most commonly applied techniques for dimensionality selection for PCA is to select sample covariance eigenvalues (and corresponding eigenvectors) that account for a fixed percentage of the total variance, for example, 90%. Typically this may only be the top two or three eigenvalues. Alternative methods consist of producing a 'scree plot', that is, plot of eigenvalue against rank, and attempting to detect by eye an 'elbow' in the plot where there is a significant change in scale of the sample covariance eigenvalues, supposedly reflecting the change from signal dominated eigenvalues to noise dominated eigenvalues. However, with sample covariance eigenvalues potentially being highly biased even when the population covariance is isotropic this is not always a reliable or easily implemented technique.

Hypothesis tests have been developed to detect departure from sphericity of the population covariance, based upon using $\mathrm{tr}\hat{\boldsymbol{C}}$ as the test statistic (John, 1971; Nagao, 1973). This approach has been modified by Ledoit and Wolf (2002) to account for smaller sample sizes but is still essentially only appropriate for $\alpha>1$. The effect of smaller values of α can be accounted for since the asymptotic form of the expected spectrum is given by the Marčenko-Pastur distribution (2) when $\boldsymbol{C}=\sigma^2\boldsymbol{I}$. Wachter has used this by producing Q-Q plots of the sample covariance eigenvalue quantiles against the Marčenko-Pastur distribution quantiles (Wachter, 1976). Sample covariance eigenvalues above

the 45 degree line in these Wachter plots indicate potentially signal containing principal components. At finite values of d a more principled, but non-Bayesian, approach would be to perform a series of iterative hypothesis tests whereby the null-hypothesis H_0 is that of a model containing k signal components. Comparison of the $(k+1)^{\text{th}}$ sample covariance eigenvalue, λ_{k+1}, against the sampling distribution of λ_{k+1} under H_0 would allow for potential rejection of the null-hypothesis and inclusion of the $(k+1)^{\text{th}}$ principal component as representing genuine signal in the data. After setting a rate at which one wishes to control the Type-I error, for example, $\gamma = 0.05$, testing of the $(k+1)^{\text{th}}, (k+2)^{\text{th}}, \ldots$ principal components proceed via,

$H_0 \; : \; \boldsymbol{C} \equiv \hat{\sigma}^2 \boldsymbol{I} \;\; , \;\; \hat{\sigma}^2 \; = \; d^{-1} \sum_{j=1}^{d} \lambda_j \;\; , \;\; k = 0$
while $p(\lambda > \lambda_{k+1} | k, d, N) < \gamma$
$k \rightarrow k+1$
$\hat{\Lambda}_k \; = \; \lambda_k$
$\hat{\sigma}^2 \; = \; (d-k)^{-1} \sum_{j=k+1}^{d} \lambda_j$
$H_0 \; : \; \boldsymbol{C} \equiv \text{diag}(\hat{\Lambda}_1, \ldots, \hat{\Lambda}_k, \hat{\sigma}^2, \ldots, \hat{\sigma}^2)$
end while

To implement this testing procedure we need the cumulative sampling distribution $p(\lambda > \lambda_{k+1} | k, d, N)$ of the $(k+1)^{\text{th}}$ sample covariance eigenvalue under the null hypothesis of $\boldsymbol{C}$ containing k signal components - that is the probability, when the population covariance contains only k signal components, of the $(k+1)^{\text{th}}$ sample covariance eigenvalue being larger than the eigenvalue λ_{k+1} observed in the real sample data. Johnstone (2001) has derived the sampling distribution for $k = 0$ by extending the analysis of Tracy and Widom (1996) on the Gaussian Orthogonal Ensemble (GOE) of random matrices. We can define location and scale constants,

$$\mu_{Nd} \; = \; N^{-1} \left(\sqrt{N-1} \, + \, \sqrt{d} \right)^2 \, ,$$

and

$$\sigma_{Nd} \; = \; N^{-1} \left(\sqrt{N-1} \, + \, \sqrt{d} \right) \left(\frac{1}{\sqrt{N-1}} \, + \, \frac{1}{\sqrt{d}} \right)^{\frac{1}{3}} \, .$$

Then for data drawn from an isotropic population covariance, $\boldsymbol{C} \; = \; \sigma^2 \boldsymbol{I}$, the largest sample covariance eigenvalue λ_1 (suitably centred and scaled) converges in distribution to the Tracy-Widom distribution W_1. Specifically one has,

$$\frac{(\lambda_1 / \sigma^2) - \mu_{Nd}}{\sigma_{Nd}} \xrightarrow{\mathcal{D}} W_1 \sim F_1 \, ,$$

where,

$$F_1(s) \; = \; \exp \left\{ -\frac{1}{2} \int_s^{\infty} q(x) + (x-s) q^2(x) dx \right\} \, ,$$

with $q(x)$ being the solution to the Painlevé II differential equation that is asymptotically equivalent to the Airy function $\text{Ai}(x)$,

$$\begin{aligned} \frac{d^2 q(x)}{dx^2} \; &= \; x q(x) \, + \, 2 q^3(x) \, , \\ q(x) \; &\sim \; \text{Ai}(x) \;\; , x \rightarrow \infty \, . \end{aligned}$$

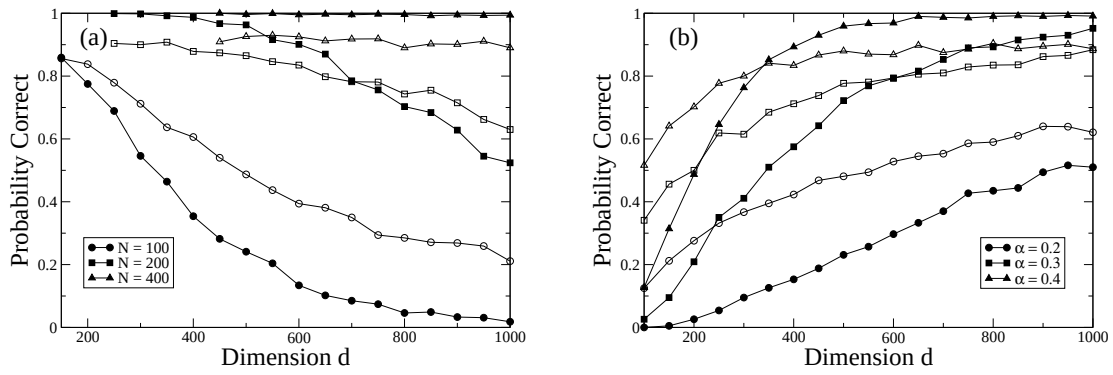

Figure 5: Comparison of the model selection accuracy for the "overlap" method (solid black symbols) with a null hypothesis test based upon the Tracy-Widom distribution for the largest eigenvalue of a sample covariance matrix (open symbols). a)Plot of model selection accuracy against data dimension d at fixed values of N. (b)Plot of model selection accuracy against data dimension for fixed values of α.

Note that the centering constant $\mu_{Nd} \to \lambda_{max}$, up to an irrelevant factor of σ^2. That is, the edge of the Marčenko-Pastur distribution, as $N, d \to \infty$ at fixed α. More recent analysis of the distribution of λ_1 when the data is complex and contains signal has been performed by Baik et al. (2005). The authors provide conjectures for the behaviour of the sampling distribution of λ_1 when the data is real, based upon their analysis of the complex case, but this still does not provide a means of calculating the sampling distribution $p(\lambda_{k+1}|k,d,N)$ for $k > 0$. Instead Johnstone (2001) derives the inequality $p(\lambda > \lambda_{k+1}|k,d,N) < p(\lambda > \lambda_1|0,d-k,N)$, with the latter distribution being given in terms of the Tracy-Widom distribution. Consequently, at finite N this provides a conservative hypothesis test since use of $p(\lambda_1|0,d-k,N)$ yields an over-estimate of the tail area of the sampling distribution $p(\lambda_{k+1}|k,d,N)$, and therefore an over-estimate of the Type-I error rate. With the variance σ^2_{Nd} tending to zero as $N \to \infty$, then in the asymptotic limit $N \to \infty$ the series of hypothesis tests given above corresponds simply to determining how many sample covariance eigenvalues λ_i are above λ_{max} - the edge of the Marčenko-Pastur bulk distribution - and so, as naively expected, is in agreement with the behaviour of the "overlap" method in the distinguished asymptotic limit.

Although in the distinguished asymptotic limit the Bayesian and frequentist approaches to model selection agree, it is interesting to compare model selection accuracies for finite values of N and d. For real data sets the sampling distribution of the individually ranked eigenvalues will have an effect upon the performance of the hypothesis test approach and likewise accuracy of point estimates for model parameters will impact upon the performance of the Bayesian methods. Figure 5 shows the model selection accuracy for the "overlap" method compared to that for the null hypothesis test outlined above that is based upon the Tracy-Widom distribution. Fig.5(a) shows the probability of selecting the correct model dimension against d, for different fixed values of N. Plotted in Fig.5(b) is the probability of selecting the correct model dimension against d, for different fixed values of α. Sample sizes and model parameter values are identical to those in Figure 2

and Figure 3. We have controlled the Type-I error at the 5% level ($\gamma = 0.05$) with the value of the abscissa for the 95^{th} centile of the Tracy-Widom distribution taken from Johnstone (2001). Within Fig.5(b) we would expect all model selection accuracies to converge to 1 as $d, N \to \infty$ since all the signal strengths have been chosen to be above their respective retarded learning transition points and therefore the sample eigenvalues corresponding to signal directions are all distinguishable from the Marčenko-Pastur bulk in this limit. However, for finite d and N Fig.5(a) and (b) reveal that overall the hypothesis testing approach has a superior model selection accuracy when both N and α are relatively small. By definition, the hypothesis test only considers a sample covariance eigenvalue to represent a signal if it exceeds that expected from the null model by more than reasonable sampling variation. As sampling variation will be greater at smaller values of N we might expect the hypothesis testing approach to be more sensitive for model selection than the "overlap" approach within this regime, particularly since higher order terms in the asymptotic expansion of the Bayesian evidence, that have not been incorporated into the "overlap" evidence approximation, will be more significant for smaller values of d and N. For larger values of d and N, the conservative nature of any hypothesis testing approach may adversely affect its model selection accuracy in comparison to a Bayesian evidence based approach.

7. Discussion & Conclusions

For calculations within high-dimensional inference problems we have argued that, rather than using results obtained by considering the traditional large sample limit $N \to \infty$, better approximations may be obtained by considering them to be close to the asymptotic value obtained in some distinguished limit, even though the sample size N may naively be considered large enough for routine application of the Laplace approximation to be accurate. What constitutes a large sample size, N, should clearly be defined with respect to the data dimensionality d. For PCA the appropriate distinguished asymptotic limit is $d, N \to \infty$, with $\alpha = N/d$ fixed, though for other models different distinguished limits may need to be considered in order to observe meaningful non-trivial behaviour that is distinct from the large sample limit, $N \to \infty$. For example, statistical physics studies of independent component analysis (ICA) suggests that $d \to \infty$ with $N = \alpha d^{\frac{3}{2}}$ at fixed α would be the appropriate distinguished limit to consider (Urbanczik, 2003). However, irrespective of the particular distinguished limit considered when developing an asymptotic approximation, one needs to be careful to keep track of the increasing number of contributions as $d \to \infty$, and potentially large contributions resulting from the rotational degeneracy of the integrand in the formulation of the Bayesian evidence.

The effect of high data dimensionality on model selection accuracy when $\alpha < 1$ is apparent from the simulation results shown in Fig.2a and Fig.3a. Ultimately this is due to the biased sample covariance eigenvalues and the poor accuracy of the sample covariance eigenvectors in representing the signal directions when $\alpha < 1$. The high-dimensional nature of the data leads to high-dimensional integral formulations of the Bayesian evidence. Approximation of the evidence has to be done carefully. Within the "overlap" method, inclusion of large contributions to the evidence from rotational degeneracy of the model $k-$frame and extensive Gaussian fluctuations leads to improved model selection accuracy. The observation that reformulating the integrand can lead to improved Laplace estimates of marginal distributions is not necessarily a new one (MacKay, 1998). For high-dimensional data the reformulation is essential, and for the "overlap" method reformulation of the evidence calculation in terms of a finite number of variables has ultimately led to an integrand that is better approximated by a single Gaussian, via a steepest descent calculation. There may exist

potentially superior estimation schemes, based upon a Gaussian parametrization of the integrand, that perform well when the integrand is essentially unimodal, for example expectation propagation based schemes (Minka, 2001b) or variational approximation similar to that employed by Bishop (1999b) for model selection within Bayesian PCA, although it should be noted that Bishop (1999b) does not impose an orthogonal constraint upon the low dimensional decomposition of the population covariance. However, it is the fact that one has to reformulate the evidence calculation for a Gaussian approximation to be accurate that is our main finding here, not the particular choice of approximation scheme that one employs once the reformulation has been made. Of greater interest perhaps is the fact that we have been able to demonstrate the asymptotic equivalence of the Bayesian evidence based model selection criterion and the frequentist hypothesis testing approach to model selection. Furthermore, analysis of the asymptotic behaviour of the "overlap" approximation to the log-evidence reveals that the estimators of the population signal eigenvalues are unbiased, at least for the "spiked" covariance models considered here.

The influence of high data dimensionality on estimates of model parameters can be explicitly demonstrated by re-visiting Minka's original Laplace approximation to the evidence. Although Minka's derivation provides a poorer approximation to the model evidence, in the distinguished limit $N, d \to \infty$ at fixed α, in comparison to the "overlap" approximation, it is still the correct leading order approximation in the asymptotic limit $N \to \infty$ at arbitrary fixed values of d. Therefore it contains information about how the model evidence behaves for large values of N and d. This suggests that Minka's Laplace approximation to the model evidence could be re-used to develop improved point estimates of population covariance eigenvalues $\{\Lambda_i\}$. One proceeds by noting that the eigenvectors of $\hat{\boldsymbol{C}}$ are the maximum posterior estimates of $\boldsymbol{U}$ for arbitrary choices of $\{l_i\}$ and v, since projection of the sample data onto the sample covariance eigenvectors retains the greatest variance. We can simply re-use Minka's approach to perform the Gaussian integration over $\boldsymbol{U}$ about this maximum posterior point, yielding $p(D|\{l_i\}, v)$ which can then be optimized with respect to $\{l_i\}$ and v. Specifically we have the following approximation to the log-evidence (taking $\eta \to 0$),

$$\begin{aligned} & -\frac{N+1}{2}\left(\sum_{i=1}^{k}\ln l_i \,+\, (d-k)\ln v\right) \,-\, \frac{N}{2v}\sum_{j=1}^{N-1}\lambda_j \,+\, \frac{N}{2}\sum_{i=1}^{k}(v^{-1}-l_i^{-1})\lambda_i \\ - \quad & \frac{1}{2}\ln|\boldsymbol{A}_Z| \,+\, \ln \mathcal{N}_k(d) \,-\, \ln \mathrm{Area}(V_k(\mathbb{R}^d)) \,+\, \frac{m+k}{2}\ln 2\pi \,-\, \frac{k}{2}\ln N\,, \end{aligned} \tag{24}$$

$$\ln|\boldsymbol{A}_Z| \,=\, m\ln N \,+\, \sum_{i=1}^{k}\left((d-k)\ln(v^{-1}-l_i^{-1}) \,+\, \sum_{j=i+1}^{k}\ln(l_j^{-1}-l_i^{-1}) \,+\, \sum_{j=i+1}^{d}\ln(\lambda_i-\lambda_j)\right).$$

It should be noted that the contribution from $\ln|\mathbf{A}_Z|$ is extensive in d and therefore affects the construction of point estimates for $\{l_i\}$ and v. Retaining only extensive terms in Equation (24) and locating stationary points with respect to l_i and v yields estimators $\hat{l}_i, \hat{v}$. In the asymptotic limit $N \to \infty$ these estimators are given by,

$$\begin{aligned} 0 \;&=\; \hat{v}^{-1}\hat{l}_i^2 \,-\, \hat{l}_i(1+\hat{v}^{-1}\lambda_i-\alpha^{-1}) \,+\, \lambda_i\,, \\ \hat{v} \;&=\; d^{-1}\sum_{j=1}^{N-1}\lambda_j\,. \end{aligned}$$

The equation above, determining the asymptotic behaviour of the estimator $\hat{l}_i$ is asymptotically identical to that given in Equation (12) for the "overlap" method and so, as already noted, gives asymptotically unbiased estimates for the population signal eigenvalue. Although this leading order approximation to the log-evidence can yield asymptotically unbiased estimators of the population covariance eigenvalues, it is still not an accurate estimation of the log-evidence and will still give inferior model selection performance in comparison to the "overlap" method. This is because higher order terms in the asymptotic expansion of the integral in Equation (5) will also be extensive in N, on taking the asymptotic limit $N, d \to \infty$ at fixed α. Ultimately this can seen from the "overlap" reformulation of the integral defining the evidence, which introduces higher than quadratic order terms in the extensive integration variables R_{ij} in the exponent of the integrand in Equation (8). These higher than quadratic order terms only arise for $\alpha < 1$ due to the contribution of the determinant $\det \boldsymbol{M}^{(d-N-1)/2}$ that results on changing integration variables from orthonormal vectors $\{\boldsymbol{u}_i\}_{i=1}^k$ of the model k-frame to the overlap variables R_{ij}. However, as we have demonstrated with the increased model selection accuracy of the "overlap" method, it is important to explicitly reformulate the integration in terms of variables that are finite in number even in the asymptotic limit $N, d \to \infty$ at fixed α.

For the simulations presented within this paper we have taken the signal dimensionality k to be finite and relatively small, for example, $k = 1, 2, 3$, so that $k \ll N < d$. This choice reflects the current interest in "spiked" covariance models and the generic challenge of identifying a fixed low-dimensional subspace as more and more features are considered. However, it is entirely feasible to imagine scenarios where the signal dimensionality is much larger than $k = 3$, and potentially even comparable to the sample size N. The derivation of the approximation to the log-evidence given in Equation (16) is valid for any finite value of k and thus can be used for model selection even for data sets where larger values of k are appropriate. Studying the accuracy of model selection for such data sets would prove more problematic. What would be the appropriate asymptotic limit to consider? If we consider a distinguished limit characterised by $N/d \to \alpha < 1$ and $k/N \to \beta < 1$ as $N, d, k \to \infty$, then any asymptotic analysis will need to take account of the effect a non-vanishing proportion of signal population eigenvalues has upon the distribution of sample covariance eigenvalues. The signal directions would no longer represent a small number of rank one perturbations of the identity matrix, with the consequence that the limiting sample covariance eigenvalue distribution would no longer correspond to the Marčenko-Pastur distribution given in Equation (2). Whilst tools exist to characterise the expected sample covariance eigenspectrum for an arbitrary population covariance eigenspectrum (Marčenko and Pastur, 1967; Wachter, 1978; Hoyle and Rattray, 2004b), obtaining closed form analytical results and proving the asymptotic correctness of the model selection for an arbitrary expected sample eigenspectrum would be difficult.

Finally, we should comment that we have illustrated ideas and concepts using model selection for PCA, in particular for $\alpha < 1$. Even today, with readily available compute power and sophisticated statistical learning algorithms, PCA is still a popular tool for dimensionality reduction or exploratory analysis. The application of PCA to extremely high-dimensional small sample size data sets has only increased the need for accurate model selection procedures. We also chose PCA as our exemplar because there already exists an elegant formulation of the Bayesian model selection problem (Minka, 2000, 2001a), and an approximation to the model evidence obtained by routine application of the Laplace approximation had already been developed. However, we believe that many of the ideas presented here are valid more generally. A large contribution to the Bayesian evidence for PCA arises from the rotational degeneracy of the model likelihood, that is, that there are many

orientations of the k-frame formed by the model signal vectors, $\{\boldsymbol{u}_i\}_{i=1}^k$, that are equally capable of accounting for the observed data. This ultimately stems from the fact that we are attempting to make inferences about vectors in $\mathbb{R}^d$ whilst we only have N sample vectors from which to construct a basis for the space. Thus, the degeneracy of the model likelihood is due to a combination of small sample size, $N < d$, and that the likelihood is expressed in terms of projections of the sample data onto the model signal vectors. This is true irrespective of whether the signal vectors $\{\boldsymbol{u}_i\}_{i=1}^k$ are constrained to be orthogonal or not, and so we expect that the issues illustrated here with PCA will be equally applicable to a number of other dimensionality reduction algorithms.

Acknowledgments

The author would like to thank Dr. Magnus Rattray for beneficial discussions and comments on the manuscript.

Appendix A.

Evaluation over data sets of the expectation value, $\mathrm{E}_{\boldsymbol{\xi}}(\sum_{j>k}\ln(\lambda_i-\lambda_j))$ (for $i \leq k$), would appear to be problematic. Since we are interested in the leading order behaviour of this expectation value, that is, the scaling with N, we can change the summation over j to include only those eigenvalues in the bulk distribution given in Equation (2). Potentially the leading order term can then be evaluated via,

$$\lim_{N,d\to\infty} N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right) = \alpha^{-1}\int_{\lambda_{min}}^{\lambda_{max}} d\lambda \ln(\lambda_i-\lambda)\rho_{bulk}(\lambda)\,, \tag{25}$$

and where $\rho_{bulk}(\lambda)$ is the Marčenko-Pastur bulk distribution given in Equation (2). Even if some sample covariance eigenvalues λ_j lie outside the Marčenko-Pastur bulk for $j > k$ the asymptotic result given in (25) is still valid since $N^{-1}\ln(\lambda_i-\lambda) \sim O(N^{-1})$ for $\lambda_i > \lambda > \lambda_{max}$. Thus contributions to $N^{-1}\mathrm{E}_{\boldsymbol{\xi}}(\sum_{j>k}\ln(\lambda_i-\lambda_j))$ from a small number of sample eigenvalues outside of the bulk distribution are vanishingly small in the asymptotic limit. The direct evaluation of the integral in (25) is difficult, so we prefer to use an indirect method. Since we are restricting the summation over j to eigenvalues in the bulk then if we denote the interval $[\lambda_{min},\lambda_{max}] \equiv I_{bulk}$, we can write,

$$\lim_{N,d\to\infty} N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{\lambda_j\in I_{bulk}}\ln(\lambda_i-\lambda_j)\right) = \lim_{N,d\to\infty} N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\mathrm{tr}\ln(\lambda_i\boldsymbol{I} - N^{-1}\boldsymbol{G})\right)\,. \tag{26}$$

where $\boldsymbol{G}$ is the Gram matrix formed from N, d-dimensional samples drawn from a multi-variate zero-mean Gaussian distribution with population covariance $\boldsymbol{C} = \sigma^2\boldsymbol{I}$, that is, the matrix $\boldsymbol{G}$ has elements $G_{\mu\mu'} = \boldsymbol{\xi}_\mu^T\boldsymbol{\xi}_{\mu'}$. The expectation, over data sets $\{\boldsymbol{\xi}_\mu\}_{\mu=1}^N$, of $\mathrm{tr}\ln(\lambda_i\boldsymbol{I} - N^{-1}\boldsymbol{G})$ is performed with the aid of the replica trick, which uses the representation,

$$\ln y = \lim_{n\to 0}\frac{(y^n-1)}{n}\,.$$

The calculation proceeds in a straight-forward fashion. We only give brief details here and the reader is referred to more in-depth explanations, given elsewhere, of the use of the replica trick in statistical physics and machine learning (Mezard et al., 1987; Hertz et al., 1991; Engel and Van den

Broeck, 2001). We find that evaluation of Equation (26) is given by the extremal value (with respect to x and q_1) of,

$$-\left\{\ln x + \frac{q_1}{x} - \alpha^{-1}\ln(1-\sigma^2 x) - \frac{\alpha^{-1}\sigma^2 q_1}{1-\sigma^2 x} - \lambda_i(x+q_1) + 1\right\}. \tag{27}$$

Differentiating with respect to x and q_1 the expression in (27) is easily maximized to give (for $\lambda_i = \sigma^2(1+A_i)(1+\alpha^{-1}A_i^{-1})$),

$$\ln\sigma^2 - (\alpha^{-1}-1)\ln(1+A_i) + \alpha^{-1}\ln A_i + \frac{1}{\alpha A_i}. \tag{28}$$

Here we have assumed the sample covariance eigenvalue λ_i will correspond to that from a "spiked" population covariance and that we are above the retarded learning transition for the i^{th} signal component. For $\alpha < A_i^{-2}$ we are below the retarded learning transition for the i^{th} signal and we expect λ_i to be located approximately at the upper edge of the bulk distribution so that $\lambda_i \simeq \sigma^2(1+\alpha^{-\frac{1}{2}})^2$. We expect the summation in $N^{-1}\sum_{j>k}\ln(\lambda_i-\lambda_j)$ will still converge since it is restricted to $j > k \geq i$. Setting $\lambda_i = \lambda_{max}$ in the previous replica calculation still yields a well-behaved estimate for $\lim_{N\to\infty} N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right)$, namely,

$$\ln\sigma^2 - (\alpha^{-1}-1)\ln(1+\alpha^{-\frac{1}{2}}) + \alpha^{-1}\ln\alpha^{-\frac{1}{2}} + \alpha^{-\frac{1}{2}}. \tag{29}$$

Since $A_i = \alpha^{-\frac{1}{2}}$ is the limit at which λ_i is indistinguishable from the bulk distribution, that is, $\lambda_i \to \lambda_{max}$, it is unsurprising that Equation (29) is obtained as the limit of Equation (28) as $A_i \to \alpha^{-\frac{1}{2}}$.

Figure 6a compares the limiting theoretical estimates for $N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right)$, given in Equations (28) and (29) with simulation for different values of α. Different plotted symbols represent different values of i, with $i = 1,\ldots,5$ running from top to bottom respectively. For each series we have set $k = i$ in the evaluation of the simulation averages. This was considered to be better than artificially setting $k = 3$, the true signal dimensionality which in general would not be known. Although in some cases, for evaluation of the simulation averages, this will lead to summation over sample covariance eigenvalues that are outside of the Marčenko-Pastur bulk distribution these will make only $O(N^{-1})$ contributions to $N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right)$, and so the simulation averages still provide a relevant test of the theoretical estimate of the asymptotic limiting value $\lim_{N\to\infty} N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right)$. Asymptotically, in the limit $N\to\infty$ and for the population covariance signal strengths chosen, three sample covariance eigenvalues are expected to be separated from the bulk distribution over the entire range of α plotted. Consequently we expect simulation averages to have a distinctly different behaviour for $i \leq 3$ compared to $i > 3$. A common limiting value for $N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>k}\ln(\lambda_i-\lambda_j)\right)$ when $i > 3$ is apparent from Figure 6a. Figure 6b compares simulation averages with the theoretical estimates in Equations (28) and (29) for different signal strengths. In this case the population covariance contains a single signal component, of strength A, whilst we have fixed $\alpha = 0.1$. The sample covariance eigenspectrum is expected to display a transition at $A = 1.0/\sqrt{\alpha} \simeq 3.16$. This is clearly reflected in the behaviour of the simulation average. The convergence towards the limiting theoretical estimate is also apparent from the comparison of simulation averages for $d = 1000$ and $d = 2000$.

References

T.W. Anderson. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 34:122–148, 1963.

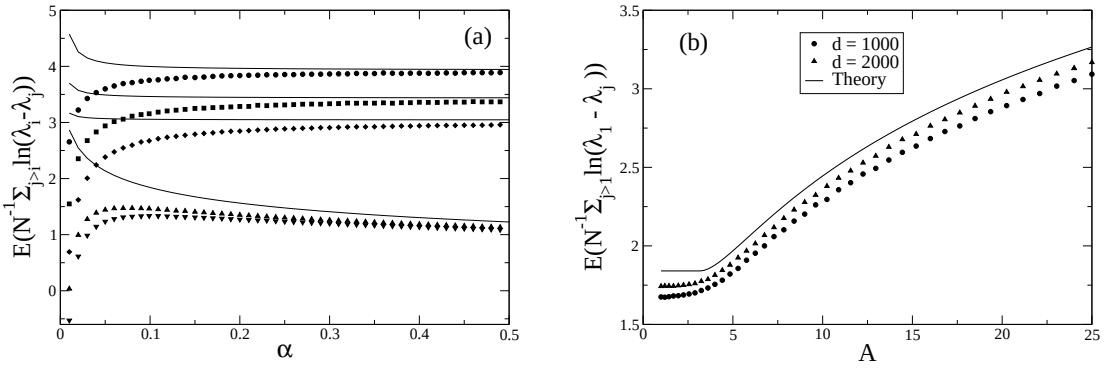

Figure 6: Comparison of simulation averages of $N^{-1}\mathrm{E}_{\boldsymbol{\xi}}\left(\sum_{j>i}\ln(\lambda_i-\lambda_j)\right)$ with the limiting theoretical estimates given in Equations (28) and (29). Figure a shows behaviour of the expectation value with α for $i = 1,\ldots,5$. Solid symbols show simulation averages whilst the solid lines show the corresponding theoretical estimates. We have set $d = 1000$ and $\sigma^2 = 1$. The population covariance contains three signal components with $A_1 = 50, A_2 = 30, A_3 = 20$. Figure b shows comparison of the theoretical result with simulation for different signal strengths, at two different values of the data dimensionality d. We have set $\alpha = 0.1, \sigma^2 = 1$. The population covariance contains a single signal component with signal strength A. For both Figure a and Figure b simulation averages are taken over 1000 matrices, and error bars of the simulation averages are smaller than the size of the plotted symbols.

J. Baik and J.W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97:1382–1408, 2006.

J. Baik, G. Ben Arous, and S. Peche. Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *Annals of Probability*, 33:1643–1697, 2005.

C.M. Bishop. Bayesian PCA. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems*, pages 382–388. MIT Press, 1999a.

C.M. Bishop. Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, pages 509–514. IEE, 1999b.

A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning*. CUP, Cambridge, 2001.

T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286: 531–537, 1999.

J.A. Hertz, A. Krogh, and R.G. Palmer. *Introduction to the Theory of Neural Computation (Santa Fe Institute Studies in the Sciences of Complexity)*. Addison-Wesley, Redwood City, CA, 1991.

D.C. Hoyle and M. Rattray. PCA learning for sparse high-dimensional data. *Europhysics Letters*, 62:117–123, 2003.

D.C. Hoyle and M. Rattray. Principal-component-analysis eigenvalue spectra from data with symmetry-breaking structure. *Physical Review E*, 69:026124, 2004a.

D.C. Hoyle and M. Rattray. Statistical mechanics of learning multiple orthogonal signals : asymptotic theory and fluctuation effects. *Physical Review E*, 75:016101, 2007.

D.C. Hoyle and M. Rattray. A statistical mechanics analysis of gram matrix eigenvalue spectra. In J. Shawe-Taylor and Y. Singer, editors, *Proceedings of COLT'04, Conference on Learning Theory, Banff, Canada, 2004. Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2004b.

A.T. James. Normal multivariate analysis and the orthogonal group. *Annals of Mathematical Statistics*, 25:40–75, 1954.

S. John. Some optimal multivariate tests. *Biometrika*, 58:123–127, 1971.

I.M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29:295–327, 2001.

I.M. Johnstone. High dimensional statistical inference and random matrices. In M. Sanz-Solé, J. Soria, J.L. Varona, and J. Verdera, editors, *Proceedings of International Congress of Mathematicians, Madrid, 2006*. European Mathematical Society Publishing House, 2006.

I.T. Joliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

D. Landgrebe. Hyperspectral image data analysis as a high dimensional signal processing problem. *IEEE Signal Processing Magazine*, 19:17–28, 2002.

O. Ledoit and M. Wolf. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Annals of Statistics*, 30:1081–1102, 2002.

D.J.C. MacKay. Choice of basis for Laplace approximation. *Machine Learning*, 33:77–86, 1998.

D.J.C MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sb*, 1:457–483, 1967.

M. Mezard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond*. World Scientific Publishing, Singapore, 1987.

T.P. Minka. Automatic choice of dimensionality for PCA. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *NIPS 13*, pages 598–604. MIT Press, 2001a.

T.P. Minka. Automatic choice of dimensionality for PCA. Technical Report TR-514, M.I.T. Media Laboratory Perceptual Computing Section, 2000. Available from http://vismod.media.mit.edu/tech-reports/TR-514-ABSTRACT.html.

T.P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, UAI-2001*, pages 362–369, 2001b.

H. Nagao. On some test criteria for covariance matrix. *Annals of Statistics*, 1:700–709, 1973.

A.L. Price, N.J. Patterson, R.M. Plenge, M.A. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38:904–909, 2006.

P. Reimann, C. Van den Broeck, and G.J. Bex. A gaussian scenario for unsupervised learning. *Journal of Physics A:Mathematical and General.*, 29:3521–3535, 1996.

S. Roweis. EM algorithms for PCA and SPCA. In M I. Jordan, M J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998.

B. Scholköpf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. *J. Royal Statistical Society B*, 61:611–622, 1999a.

M.E. Tipping and C.M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11:443–482, 1999b.

C.A. Tracy and H. Widom. On orthogonal and symplectic matrix ensembles. *Communications in Mathematical Physics*, 177:727–754, 1996.

R. Urbanczik. Statistical physics of independent component analysis. *Europhysics Letters*, 64: 564–570, 2003.

V. Šmídl and A. Quinn. On Bayesian principal component analysis. *Computational Statistics and Data Analysis*, 51:4101–4123, 2007.

K.W. Wachter. In David C. Hoaglin & Roy E. Welsch, editor, *Proceedings of the Ninth Interface Symposium Computer Science and Statistics*, page 299, Boston, 1976. Prindle, Weber and Schmidt.

K.W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *Annnals Probability*, 6:1–18, 1978.

R. Wong. *Asymptotic Approximations of Integrals*. Academic Press, Boston, MA, 1989.