RANK SELECTION IN NOISY PCA WITH SURE AND RANDOM MATRIX THEORY

M.O. Ulfarsson[†] and V.Solo[‡]

[†]University of Iceland, Dept. Electrical Eng., Reykjavik, ICELAND
[‡]University of New South Wales, School of Electrical Eng., Sydney, AUSTRALIA

ABSTRACT

Principal component analysis (PCA) is probably the best known method for dimensionality reduction. Perhaps the most important problem in PCA is to determine the number of principal components in a given data set, and in effect separate signal from noise in the data set. Many methods have been proposed to deal with this problem but almost all of them fail in the important practical case when the number of observations is comparable to the number of variables, i.e., the realm of Random Matrix Theory (RMT). In this paper, we propose to use Stein's Unbiased Risk Estimator (SURE) to estimate, with some assistance from RMT, the number of principal components. The method is applied on simulated data and compared to BIC and the Laplace method.

Index Terms— Principal component analysis, Stein's Unbiased Risk Estimator (SURE), Random matrix theory, model order selection.

1. INTRODUCTION

A central problem in the study of high dimensional data of T observations on M variables is dimensionality reduction. Its aim is to separate signal from the noise such that the signal is captured in a few leading functionals of the data and the noise in the rest. The noise can then be discarded leaving us with a smaller data set of higher signal content.

The best known method for this purpose is probably Principal Component Analysis (PCA) [1]. PCA is based on linear combinations of the data while newer methods such as kernel PCA use nonlinear functionals. However, PCA remains the most widely used dimensional reduction method. The applications for PCA in signal and image processing are wide ranging, e.g., data compression, denoising, and as a preprocessing step for various kinds of algorithms.

Although PCA is commonly viewed as an exploratory technique, there is in fact an underlying model behind PCA [2], which we call noisy PCA (nPCA), where the nPCs can be derived by the maximum likelihood method. This gives the researcher access to classical model selection methods based on information criterion such as AIC, BIC (see [3] and references therein). These methods were developed for the data rich case $T >> M$ so there is no guarantee that they will perform well on other cases.

Modern data sets are often very high dimensional with M on the order of hundreds or thousands and T and M of comparable sizes. Examples include: Meteorology and Oceanography [4], functional data analysis [5], and medical imaging [6]. The reference [7] develops a Bayesian model selection method, which we call the Laplace method, based on the Laplace approximation, and shows that it performs very well in cases where T and M are of comparable sizes. Furthermore, in simulations, it compares well against cross-validation, and other similar Bayesian methods [7]. Reference [8] employs the Laplace criterion for functional Magnetic Resonance Imaging (fMRI) data, in addition, it makes use of Random Matrix Theory (RMT) to modify it. Other methods include [9].

In this paper we propose to use Stein's Unbiased Risk Estimator (SURE) [10] to choose the number of nPCs. SURE was originally not designed for model selection, but following [11, 12] the second author realized [13] that it could be used as a general purpose tool for tuning parameter selection in non-linear ill-conditioned inverse problems.

The advantages of our SURE based selection method are: 1) it is computationally simple, i.e., does not require much more computation than that needed to obtain the PCs, 2) it has an unbiasedness property even for non-linear problems 3) it is exact, i.e., no approximations are needed. To implement SURE in practice it is necessary to estimate a noise variance and we develop a novel method based on RMT to do that.

This paper is organized as follows: In Section 2, we discuss nPCA. In Section 3, we derive SURE for PCA and discuss BIC and the Laplace method. Section 4 discusses estimation methods for the noise variance based on RMT. Section 5 presents simulation results. Finally, in Section 6, conclusions are drawn.

For easy reference, we list notations used in this paper. An estimate of θ is denoted by $\hat{\theta}$; $\gamma = T/M$ is the ratio of the number of observation to the number of variables; $I(x) = 1$, if $x \in S$, zero otherwise; a.s. means almost surely; $\|x\|^2 = \sum_{i=1}^{n} x_i^2$; E denotes the expectation operator.

2. NOISY PCA

The nPCA model is given by

$$\begin{aligned} y_t &= \mu_t + \epsilon_t \\ &= m + Gu_t + \epsilon_t, \quad t = 1, ..., T \end{aligned} \tag{1}$$

where y_t is a $M \times 1$ data vector, m is the mean vector, $G = (g_1, g_2, ..., g_r)$ is an $M \times r$ loading matrix, $u_t \sim N(0, I_r)$ is a r vector of nPCs, $\epsilon_t \sim N(0, \sigma^2 I_M)$ is white noise, and ϵ_t, u_t are mutually independent. The problem at hand is to estimate $\theta = (m, G, \sigma^2)$.

The log-likelihood is given by

$$l_\theta(y) = -\frac{T}{2}\text{tr}(S_y \Omega^{-1}) - \frac{T}{2}\log|\Omega|$$

where $\Omega = GG^T + \sigma^2 I_M$ and $S_y = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})(y_t - \bar{y})^T$. The maximum likelihood estimators are [2, 14]

$$\begin{aligned} \hat{m} &= \bar{y} \\ \hat{G} &= P(L - \hat{\sigma}_r^2 I_r)^{1/2} R \\ \hat{\sigma}_r^2 &= \frac{1}{M-r}\sum_{j=r+1}^{M} l_j. \end{aligned} \tag{2}$$

Here R is an arbitrary orthonormal rotation matrix, $L = \text{diag}(l_1, ..., l_r)$, where $l_1 > l_2 >, ..., > l_r$, contains the r largest eigenvalues of the data covariance matrix S_y , where the columns of $P = [p_1, ..., p_r]$ are the corresponding eigenvectors, so $S_y P = PL$. The reference [14] proved that the ML estimates are the global optimizers of the likelihood, this was rediscovered by [15].

For given data and the ML estimate of θ, the estimated nPCs and their corresponding variances are given by

$$\begin{aligned} \hat{u}_t &= E_{\hat{\theta}}(u_t|y_t) = W^{-1}\hat{G}^T(y_t - \bar{y}) \\ V &= \text{var}_{\hat{\theta}}(u_t|y_t) = \hat{\sigma}_r^2 W^{-1} \\ W &= \hat{G}^T\hat{G} + \hat{\sigma}_r^2 I_r. \end{aligned}$$

The estimate for μ_t in Equation (1) is given by

$$\hat{\mu}_t = \hat{G}\hat{u}_t = \bar{y} + \sum_{j=1}^{r} p_j \frac{l_j - \hat{\sigma}_r^2}{l_j} p_j^T (y_t - \bar{y}). \tag{3}$$

This is a nonlinear function in y_t.

3. RANK SELECTION

In this section, we first review the Laplace and the BIC methods and then introduce our SURE method. We chose the Laplace method since in [7] it was easily the best performing method among a number of methods designed for the low sample case. BIC is chosen as a surrogate for all methods that assume a large sample framework.

3.1. The Laplace and the BIC Methods

The Laplace method is derived from a Bayesian framework and is based on maximizing the evidence that the nPCA model consists of r PCs. The reference [7] approximated the evidence using the Laplace method yielding

$$\begin{aligned} -\log p(y|r) &= -l_{\hat{\theta}}(y) - \log p(P) - \frac{d+r}{2}\log 2\pi \\ &+ \frac{1}{2}\log|A_z| + \frac{r}{2}\log T \end{aligned}$$

where $d = Mr - r(r+1)/2$ and

$$\begin{aligned} p(P) &= 2^{-r}\prod_{i=1}^{r}\Gamma((M-i+1)/2)\pi^{-(M-i+1)/2} \\ |A_z| &= \prod_{i=1}^{r}\prod_{j=i+1}^{M} T(\tilde{l}_j^{-1} - \tilde{l}_i^{-1})(l_i - l_j) \end{aligned}$$

where $p(P)$ is a noninformative prior distribution for P, $\tilde{l}_j$ is equal to l_j when $j \leq r$ and equal to $\hat{\sigma}_r^2$ when $j > r$, and

$$-l_{\hat{\theta}}(y) = \frac{T}{2}\sum_{j=1}^{r}\log l_j + \frac{T}{2}(M-r)\log\hat{\sigma}_r^2.$$

The r that minimizes $-\log p(y|r)$ is picked as the number of nPCs. The BIC criterion can be thought as an approximation of the Laplace criterion by dropping all terms in it that do not grow with T [16] giving

$$\text{BIC}_r = -l_{\hat{\theta}}(y) + \frac{1}{2}(d+r)\log T.$$

The r that minimizes the BIC is chosen as the number of nPCs.

3.2. SURE

The proposed SURE based method is based on the following considerations. Ideally, we would like to choose the value of r that minimizes the risk

$$R_r = \frac{1}{T}\sum_{t=1}^{T} E\|\mu_t - \hat{\mu}_t\|^2$$

where μ_t is the true signal and $\hat{\mu}_t$ is an estimate of μ_t for a known value of r. We generally do not know the true signal μ_t so we cannot compute the risk. But the idea is to try to find a computable unbiased estimator of it and minimize that instead. Indeed, remarkably Stein [10] showed how to construct such an estimator under Gaussian assumptions. SURE is given by

$$\hat{R}_r = \frac{1}{T}\sum_{t=1}^{T}\|n_t\|^2 + 2\sigma^2\frac{1}{T}\sum_{t=1}^{T}\text{tr}(\frac{\partial\hat{\mu}_t}{\partial y_t^T}) - M\sigma^2$$

where our estimator of μ_t is given by Equation (3), and $n_t = y_t - \hat{\mu}_t$. The idea behind SURE as a tuning parameter selector is that [13] since it is an unbiased estimator of the risk then on average one can hope that its minimizer is an unbiased estimator of the minimizer of the risk.

It can be shown that

$$\begin{aligned} \hat{R}_r &= (M-r)\hat{\sigma}_r^2 + \hat{\sigma}_r^4\sum_{j=1}^{r}\frac{1}{l_j} + 2\sigma^2 r \\ &- 2\sigma^2\hat{\sigma}_r^2\sum_{j=1}^{r}\frac{1}{l_j} + \frac{4\sigma^2\hat{\sigma}_r^2}{T}\sum_{j=1}^{r}\frac{1}{l_j} + C \qquad (4) \\ C &= \frac{4\sigma^2}{T}\sum_{j=1}^{r}\sum_{i=r+1}^{M}\frac{l_j - \hat{\sigma}_r^2}{l_j - l_i} + \frac{2\sigma^2}{T}r(r-1) \\ &- \frac{2\sigma^2}{T}(M-1)\sum_{j=1}^{r}(1 - \frac{\hat{\sigma}_r^2}{l_j}). \end{aligned}$$

Note that the noise variance σ^2 is assumed known in the SURE formula. A natural choice for it is $\hat{\sigma}_r^2$ but that does not work in this case. Finding a reliable estimator for σ^2 turns out to be a non-trivial issue which we now pursue.

4. ESTIMATION OF σ^2 VIA RMT

We seek an estimator of σ^2 that does not require a good estimate of r. First, write the loading matrix in terms of its SVD, i.e., $G = F\Lambda R^T$. Then spectral decompose the covariance, i.e., $\Omega = F(\Lambda - \sigma^2 I_r)F^T + \sigma^2 F_\perp F_\perp^T$ where $F_\perp$ is orthogonal to F and of rank $M - r$. This shows that the eigenvalue spectrum consists of r signal eigenvalues equal to $\lambda_j - \sigma^2, j = 1, ..., r$, and $M - r$ noise eigenvalues all equal to σ^2. A simple idea is to use a sample noise eigenvalue of S_y as an estimate for σ^2, but it does not work well. Our idea is to correct the sample eigenvalues by 'flattening' the Empirical Distribution Function (EDF) of the sample eigenvalues using results from RMT and use a corrected sample noise eigenvalue as an estimate for σ^2. Now we discuss RMT

4.1. Random Matrix Theory

RMT is defined by a scenario in which $T \to \infty$, $M \to \infty$ while $T/M = \gamma \neq 0, \gamma < \infty$. This differs from the classical PCA

asymptotic where M is fixed and $T \to \infty$ [17]. In this RMT case the eigenvalues of the sample covariance matrix do not converge in probability to the true values. Rather the empirical distribution converges to a limit called the Marchenko-Pastur (MP) distribution and is described in a seminal paper [18].

Theorem 1 *Given a $T \times M$ data matrix Y with independent zero mean and unit variance entries. If $T, M \to \infty$, such that $T/M \to \gamma \geq 1$, and $\gamma < \infty$. Then the EDF*

$$\hat{F}_\gamma(x) = \frac{1}{M} \sum_{i=1}^{M} I(l_{M-i+1} \leq x) \tag{5}$$

of the eigenvalues associated with the covariance matrix S_y converges a.s. to the MP distribution

$$\hat{F}_\gamma(x) \to F_\gamma(x).$$

The associated MP density is given by

$$f_\gamma(x) = F'_\gamma(x) = \frac{\gamma}{2\pi x}\sqrt{(b-x)(x-a)}, \quad a \leq x \leq b$$

where $a = (1-\gamma^{-1/2})^2$ and $b = (1+\gamma^{-1/2})^2$.

Although this result is stated for the case where T and M go to infinity in fixed ratio there is empirical evidence [19, 20] that it is a good approximation for very low values of T and M, e.g., $T = M = 10$. For the case of non-unit noise variance σ^2 the distribution simply scales up and is given by $F_\gamma(\sigma^2 x)$ [21].

The case of nPCA where there are few leading signal eigenvalues and many equal valued noise eigenvalues is called the spiked model [19], i.e., noise model spiked with few significant eigenvalues. It has long been known that the MP result still holds [18] for this case, except that there may be some eigenvalues outside the MP support. Two recent papers [22, 23] give theoretical discussion of the asymptotic behavior of those eigenvalues.

4.2. Noise Variance σ^2 Estimation Method

The spiked model and the scaling property of the MP distribution lead to the following idea for estimating the noise variance. Given a sample covariance matrix S_y with eigenvalues $l_1 > l_2 > , ..., > l_M$.

1. Compute the corrected eigenvalues

$$\tilde{l}_j^{(1)} = \frac{l_j}{F_\gamma^{-1}(\hat{F}_\gamma(l_j))} = \frac{l_j}{F_\gamma^{-1}(\frac{M-j+1}{M})}, \quad j = 1, ..., M.$$

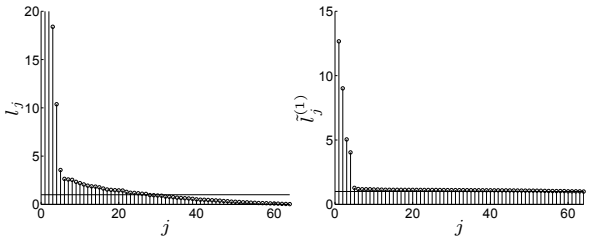
   where F_γ^{-1} is the quantile function associated with F_γ, and $\hat{F}_\gamma$ is the EDF given in Equation (5). For sample noise eigenvalues we expect that $\hat{F}_\gamma(l_j) \approx F_\gamma(\frac{l_j}{\sigma^2})$ so $\tilde{l}_j^{(1)} \approx \sigma^2$.

2. A rough estimate of σ^2 is given by one of the corrected sample noise eigenvalues, e.g., $\tilde{\sigma}^2 =$ 25th percentile of $\tilde{l}^{(1)}$.

3. Normalize the eigenvalues $\tilde{l}_j = \frac{l_j}{\tilde{\sigma}^2}, \quad j = 1, ..., M.$

4. Get a crude estimate of the number of signal eigenvalues using the upper support limit of the MP density

$$r = \text{argmin}(\tilde{l}_j^{(1)} - b),\ \tilde{l}_j^{(1)} - b > 0.$$



(a) $T = 96$ (b) $T = 96$

Fig. 1. Simulation: Scree Plots

5. Construct the EDF for the sample noise eigenvalues

$$\tilde{F}_\gamma(x) = \frac{1}{M-r} \sum_{i=r+1}^{M} I(l_{M-i+1} \leq x).$$

6. Recompute the corrected eigenvalue

$$\tilde{l}_j^{(2)} = \frac{l_j}{F_\gamma^{-1}(\tilde{F}_\gamma(l_j))} = \frac{l_j}{F_\gamma^{-1}(\frac{M-j+1}{M-r})}, j = r+1, ..., M.$$

7. The final estimate of the noise variance is given by one of the corrected sample noise eigenvalues, e.g., $\hat{\sigma}^2 =$ 25th percentile of $\tilde{l}^{(2)}$.

Notice that this noise variance estimation method also includes a crude estimate for r. This method is far from competitive with SURE but helps provide a good estimate for σ^2. In steps 2 and 7 we choose the 25th percentile of the corrected eigenvalues as an estimate for the noise variance. There is nothing special about the 25th percentile, we could have chosen the 30th percentile or the median instead, like Fig. 1(b) below illustrates.

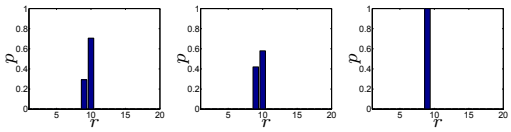
5. SIMULATION

In this section ,we present a simulation study, where we compare the SURE with the Laplace method, and BIC. The BIC and Laplace methods were implemented using formulas from Section 3. We simulated the data according to Equation (1), with the following parameters: $M = 64, T = [96, 128]$, $\lambda = l - \sigma^2 = [(r+1)^2, r^2, ..., 3^2, 2]$,$r = [5, 10, 15, 30]$, and $\sigma^2 = 1$.

The loading matrix G was simulated by generating $M \times r$ matrix of unit variance Gaussian random variables. It was then orthonormalized ($R = I_r$). All simulations were repeated $N_{rep} = 1500$ times, and it was recorded how many times each method choose the correct dimensionality. This number is binomially distributed random variable with parameters N_{rep} and p which is the classification probability.

Fig. 1(a) shows a sample Scree plot (sample eigenvalues plotted in decreasing order) from one of the replicates, the noise variance level σ^2 is indicated by a horizontal line. Fig. 1(b) shows the corresponding corrected Scree plot. The noise eigenvalues fall nicely on the horizontal noise variance line, so almost any of them could be used as an estimate for σ^2.

Table 1 shows the percentage of correct selection of PCs for SURE, Laplace, and BIC methods. The bold face entries indicate which method performs best according to the 95% significant level. It can be seen that SURE performs best in almost all of the cases, sometimes by a wide margin. BIC does not perform well in any of the cases presented, which is not surprising since BIC is based on



(a) (b) (c)

Fig. 2. Simulation: Histogram of number of PCs, where $r = 10, T = 96$. First column is SURE, second is Laplace, and third is BIC.

Table 1. The percentage of correct selection. Bold face entries represent the best performing method.

	$T = 96$			$T = 128$		
r	SURE	Laplace	BIC	SURE	Laplace	BIC
5	**0.671**	**0.661**	0	**0.886**	**0.899**	0
10	**0.718**	0.571	0	**0.901**	0.883	0.005
15	**0.775**	0.498	0.010	**0.930**	0.840	0.022
30	**0.825**	0.353	0.185	**0.956**	0.833	0.299

an asymptotic argument which does not hold. Finally, Fig. 2 shows a histogram of the number of PC chosen by the considered methods for $r = 10$, and $T = 96$.

6. CONCLUSION

In this paper, we have applied the nonlinear SURE technique to the problem of rank selection in nPCA where data and variable dimension are in the realm of RMT. This scenario causes significant problems for most other selection methods. For practical use it is necessary to estimate the noise variance and we have developed a reliable estimator based on RMT. In simulations, we have shown that BIC (and methods like it) fail badly and that our new method outperforms the Laplace method.

7. REFERENCES

[1] I. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY: Springer, 2002.

[2] D. Lawley, "A modified method of estimation in factor analysis and some large sample results," in *Uppsala symposium on psychological factor analysis*, Uppsala, Sweden, 1953, pp. 35–42.

[3] P. Stoica and Y. Selen, "Model-order selection. A review of information criterion rules," *IEEE Signal Proc. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[4] R. Preisendorfer, *Principal Component Analysis in Meteorlogy and Oceangraphy*. Amsterdam, Holland: Elsevier, 1988.

[5] J. Ramsay and B. Silverman, *Functional Data Analysis*, 1st ed. New York, NY: Springer, 1997.

[6] M. Ulfarsson and V. Solo, "Smooth principal component analysis with application to functional magnetic resonance imaging," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06)*, vol. 2, Toulouse, France, May 2006, pp. II–993 – II–996.

[7] T. Minka, "Automatic choice of dimensionality for PCA," in *Advances in Neural Information Processing Systems (NIPS'00)*, 2000, pp. 598–604.

[8] C. Beckmann and S. Smith, "Probabilistic independent component analysis for functional magnetic resonance imaging," *IEEE Trans. Med. Imag*, vol. 23, no. 2, 2004.

[9] A. Seghouane and A. Cichocki, "Bayesian estimation of the number of principal components," *Signal Processing*, vol. 87, pp. 562–568, 2007.

[10] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135–1151, 1981.

[11] D. Donoho and I. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[12] M. Hudson, "Maximum likelihood restoration and choice of smoothing parameter in deconvolution of image data subject to Poisson noise," *Comput Stat Data Anal*, vol. 26, no. 4, pp. 393–410, 1998.

[13] V. Solo, "A sure-fired way to choose smoothing parameters in ill-conditioned inverse problems," in *Proc. IEEE International Conference on Image Processing (ICIP'96)*, vol. 3, Lausanne, Switzerland, 1996, pp. 89–92.

[14] C. Theobald, "An inequality with application to multivariate analysis," *Biometrika*, vol. 62, no. 2, pp. 461–466, 1975.

[15] M. Tipping and C. Bishop, "Probabilistic principal component analysis," *J. Royal Stat. Soc., Series B*, vol. 61, no. 3, pp. 611–622, 1999.

[16] R. Kass and A. Raftery, "Bayes factors," *J. Amer. Stat. Assoc.*, vol. 90, no. 430, pp. 773–795, 1995.

[17] T. Anderson, "Estimating linear statistical relationships," *Ann. Statist.*, vol. 12, no. 1, pp. 1–45, 1984.

[18] V. Marcenko and L. Pastur, "Distribution of eigenvalues of some sets of random matrices," *Math. USSR-Sb.*, vol. 1, pp. 507–536, 1967.

[19] I. Johnstone, "On the distribution of the largest eigenvalue in principal component analysis," *Annals Stat.*, vol. 29, no. 2, pp. 295–327, 2001.

[20] R. Everson and S. Roberts, "Inferring the eigenvalues of covariance matrices from limited, noisy data," *IEEE Trans. Signal Proc.*, vol. 48, no. 7, pp. 2083–2091, 2000.

[21] J. Silverstein and P. Combettes, "Signal detection via spectral theory of large dimensional random matrices," *IEEE Trans. Signal Proc.*, vol. 40, no. 8, 1992.

[22] J. Baik and J. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *J. Multivariate Anal.*, vol. 97, pp. 1382–1408, 2006.

[23] D. Paul, "Asymptotics of the leading sample eigenvalues for a spiked covariance model," Department of Statistic, Stanford University, Tech. Rep., 2004.