

Fast Variational PCA for Functional Analysis of Dynamic Image Sequences

Václav Šmídl, Anthony Quinn

Dept. of Electronics and Electrical Engineering,
Trinity College Dublin,
IRELAND

[smidl, aquinn]@mee.tcd.ie

Abstract

Principal Component Analysis (PCA) is a well-known algorithm used in many areas of science. It is usually taken as the golden standard of dimensionality reduction. However, PCA usually does not provide information about uncertainty of its results, thus preventing further investigation of model structure. Full Bayesian treatment is not feasible. Recently, Variational PCA (VPCA) was proposed as an approximate Bayesian solution of the problem. In this text we summarise the iterative solution to the PCA problem arising from a variational approach. A new model with orthogonality restrictions is constructed in order to overcome its limitations. Notably, a highly efficient computational algorithm of variational PCA is revealed. It is applied in the functional analysis of medical imaging problem yielding solution in a fraction of the time needed by the conventional technique.

1. Introduction

Principal Component Analysis (PCA) is one of the classical data analysis tools. In digital signal processing it is used for data compression, de-noising, pattern recognition, shape analysis and spectral analysis to mention a few examples. Overview of its use in different scientific areas see [8].

It is often used as a black box for its mature implementation and ease of use. However, one important question that is not satisfactorily answered by classical theory is number of relevant principal component. Number of criterion is available [2] based on various assumption and approximations. Some applications, like de-noising, are quite insensitive to selected order, but in many others it is an important part of the problem. Typical example is spectral analysis or functional analysis of dynamic image data.

PCA is known to be a maximum likelihood estimation method for special case of model known as factor analysis model. Recently, the problem was addressed using Bayesian methodology [3, 11]. Variational PCA (VPCA) seems to be a reasonable compromise of speed and accuracy on small scale problems, however, with increasing dimensionality problems with numerical stability and conver-

gence arise.

In this paper we introduce modification of the original model and provide its variational estimation algorithms. The results of a new algorithm, Orthogonal Variational PCA (OVPCA), correspond to those of VPCA but are obtained in significantly shorter time and with ease of use compared to the classical PCA.

2. Mathematical Model

Functional analysis of dynamic image sequences is an analytical method used in medical imaging for estimation of the temporal activity of internal body organs. Its main application area is in nuclear medicine. A sequence of p images is taken within a given time range. The images are supposed to reflect the activity of a relatively small number, r , of underlying physiological organs.

Under assumptions that the shape and distribution of activity inside a single organ does not change, we can consider every measured image to be a linear combination of *factor images* reflecting organ shapes weighted by their activity at given time t . The evolution of activity of each factor image in time is called *factor curve*, illustrated in Figure 1. The images are stored in one column, of length n , where n is a number of pixels in each image. The whole sequence is then modelled as

$$D = AX + E \quad (1)$$

where $D \in \mathbb{R}^{n \times p}$ is matrix of observed image sequence, $A \in \mathbb{R}^{n \times r}$ is a matrix of r underlying factor images, and $X \in \mathbb{R}^{r \times p}$ is a matrix factor curves, $E \in \mathbb{R}^{p \times n}$, denotes matrix of realizations of zero mean noise with Gaussian distribution

$$f(E) = \mathcal{N}(\mathbf{0}_{p,n}, \sigma I_p \otimes I_n) \quad (2)$$

$$f(D) = \mathcal{N}(AX, \sigma I_p \otimes I_n) \quad (3)$$

where $\mathbf{0}_{p,n} \in \mathbb{R}^{p \times n}$ with all elements equal to 0, $I_p \in \mathbb{R}^{p \times p}$ is identity matrix and scalar σ models variance of the noise. The model (1) is closely related to model known as *factor analysis*, [2] which is defined vector-wise

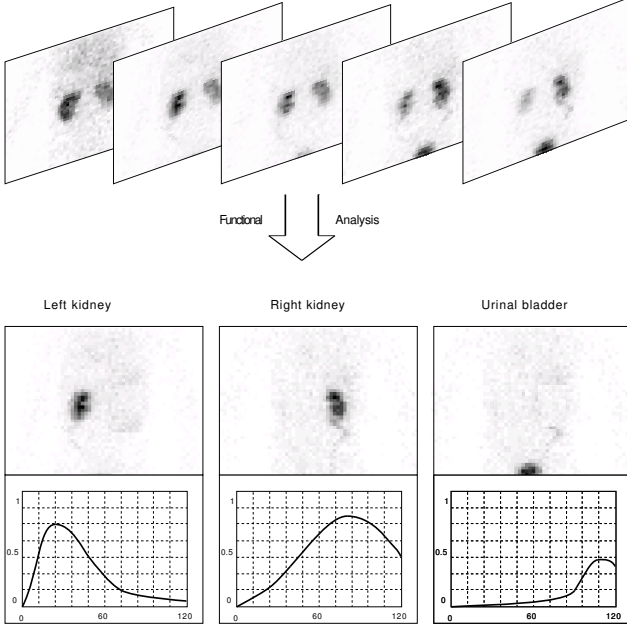


Figure 1. Illustration of Functional Analysis of Dynamic Image Data

$$d_i = \mathbf{A}\mathbf{x}_i + \mu + e_i \quad (4)$$

$$f(e_i) = \mathcal{N}(\mathbf{0}_{p,1}, \Psi) \quad (5)$$

Hence, we refer to our model as *restricted factor analysis* model with restrictions $\mu = \mathbf{0}_{p,1}$ and $\Psi = I_p$.

The model is doubly degenerate, firstly, via additivity of the signal and noise (1) and secondly, via multiplication of parameters A, X which can be replaced by

$$BY = (AT)(T^{-1}X) = AX \quad (6)$$

with equal probability of data for any invertible matrix $T \in \mathbb{R}^{r \times r}$. T is called a *rotation matrix* in the factor analysis literature. The problem of rotational ambiguity can be addressed by imposing extra restrictions to matrices A and X and derive specialised algorithms, which is often the case in factor analysis [2].

3. Maximum likelihood estimation

The maximum likelihood estimation of parameters A and X is easily found using singular value decomposition [6], of data matrix D

$$D = USV' \quad (7)$$

where $U \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$ are orthogonal matrices $U'U = I_p$, $V'V = I_p$ of so-called singular vectors and $S \in \mathbb{R}^{p \times p}$ is a diagonal matrix of ordered singular values ordered $s_1 \geq \dots \geq s_r \geq \dots \geq s_p$. It is easy to show that

the maximum of the likelihood (3) conditioned by knowledge of r is reached for

$$AX = U_r S_r V_r' \quad (8)$$

where $S_r \in \mathbb{R}^{r \times r}$ is a diagonal matrix of the r most significant singular values, and matrices $U_r \in \mathbb{R}^{n \times r}$, $V_r \in \mathbb{R}^{r \times p}$ are composed by corresponding singular vectors.

The rotational indeterminacy (6) means that the maximum likelihood solution for parameters A and X is an infinite set indexed by each possible rotation matrix, T_r

$$S_{ML} = \{A, X : A = U_r S_r T_r^{-1}, X = T_r V_r', \quad (9) \\ \forall T_r \in \mathbb{R}^{r \times r} : \exists T_r^{-1} \in \mathbb{R}^{r \times r}\}.$$

For elements $T_r = \pm I_r$, the columns of matrix X form an orthogonal basis. Because V_r is a matrix of eigenvectors of $D'D$, it is equivalent to principal component analysis (PCA) with mean value μ set to zero.

The physical nature of the problem imposes extra conditions on A and X , namely positivity of all elements. It has been shown that rotation matrix T that transforms orthogonal solution to the required form can be found [10]. It is therefore advantageous to confine attention to estimation of orthogonal solution only, i.e. a subset of (9).

Note that the solution (8) is conditioned by knowledge of the number of components, r , but having evaluated the whole matrix U we have solution for any rank r . Various techniques for selection of optimal rank have been developed, for summary see [2, 8].

4. Bayesian solution

Recently, two approximate Bayesian solutions to a model called probabilistic PCA (PPCA) were published, using variational Bayes [3] and Laplace [11] approximations respectively. The model is almost identical to (3) up to a mean value, μ . However, the difference between the solution for PPCA and for (3) are trivial.

4.1. Variational Bayes

A variational approximation can be applied to evaluation of a marginal distribution if the likelihood function is dependent on at least two parameters. This method is known also as ensemble learning or minimisation of free entropy [12]. It is summarised by the following theorem.

Theorem 4.1 (Variational Bayes) Let $f(a, b, D)$ be a joint pdf of known D and random variables a, b . Let $q(a, b|D)$ be an approximate pdf restricted to the set of conditionally independent distributions on a, b

$$q(a, b|D) = q_a(a|D) q_b(b|D).$$

Then, minimum of KL distance $KL(q(a, b|D) || f(a, b, D))$ is reached for

$$q_a(a|D) \propto \exp(\mathcal{E}_{q_b}\{\ln(f(a, b, D))\}), \\ q_b(b|D) \propto \exp(\mathcal{E}_{q_a}\{\ln(f(a, b, D))\})$$

where \mathcal{E}_q denotes expected value with respect to distribution $q(\cdot|D)$, [3].

The above theorem is available powerful tool for approximation of joint pdfs. It provides functional optimisation of parameter pdf, i.e. it yields the form of its posterior. However, moments of the resulting distribution are mutually dependent, thus forming a set of implicit equations.

4.2. Variational PCA

In order to apply Bayesian solution the model (3) has to be complemented by a selected prior distribution on the parameters. It was chosen

$$\begin{aligned} f_0(X) &= \mathcal{N}(0, I_p \otimes I_n), \\ f_0(A) &= \mathcal{N}(0, \text{diag}(\alpha) \otimes I_n), \\ f_0(\sigma) &= \Gamma(c_0, d_0), \\ f_0(\alpha) &= \prod_{i=1}^{p-1} \Gamma(a_0, b_0), \end{aligned}$$

where α is vector of hyper-parameters $\alpha_i, i = 1 \dots p-1$.

Application of the theorem 4.1 gives following results

$$\begin{aligned} f(X) &= \mathcal{N}(\bar{X}, \Sigma_X \otimes I_{(N \times N)}) \quad (10) \\ f(A) &= \mathcal{N}(\bar{A}, I_{(p \times p)} \otimes \Sigma_A) \\ f(\sigma) &= \Gamma(c, d) \\ f(\alpha) &= \prod_{i=1}^R \Gamma(a, b_i) \end{aligned}$$

with moments given by a set of implicit equations

$$\bar{X} = \frac{c}{d} \Sigma_X' \bar{A}' D \quad (11)$$

$$\Sigma_X = \left(I_r + \frac{c}{d} \left(p \Sigma_A + \bar{A}' \bar{A} \right) \right)^{-1} \quad (12)$$

$$\bar{A} = \frac{c}{d} \Sigma_A \bar{X} D' \quad (13)$$

$$\Sigma_A = \left(\text{diag}\left(\frac{a}{b}\right) + \frac{c}{d} \left(N \Sigma_X + \bar{X}' \bar{X} \right) \right)^{-1} \quad (14)$$

$$a = a_0 + 0.5p \quad (15)$$

$$b = \text{diag}\left(b_0 + \left(p \Sigma_A + \bar{A}' \bar{A}\right)\right) \quad (16)$$

$$c = c_0 + 0.5np \quad (17)$$

$$\begin{aligned} d &= d_0 + \frac{1}{2} \text{trace}(DD') - 2 \text{trace}(D' \bar{A} \bar{X}) + \\ &+ \text{trace}\left(\left(p \Sigma_A + \bar{A}' \bar{A}\right) \left(N \Sigma_X + \bar{X}' \bar{X}\right)\right). \end{aligned} \quad (18)$$

The above implicit set of equation can be solved by various methods. The solution suggested in [3] is an iterative algorithm with random initial conditions. New values of estimated parameters are obtained by evaluation of equations (11-18) for previous values of parameters. Optimal rank, r , is determined by a number of hyper-parameters α_i that remains higher than zero after all iterations. Technically, as there is a prior on α we need to set up some threshold above prior value a_0/b_0 .

The rotational ambiguity (6) is mentioned in the original paper but little attention is paid to its influence on the iterative algorithm. It is worth to notice, that maximum of the model (3) is the infinite set (9). The chosen prior distributions (10) reduce this ambiguity but the joint likelihood on the set (9) remains very flat. Chosen iterative algorithm thus can be very slow if starting point lies far from the optimal solution. A probability that it happen using random initialisation is rather high. The problem is magnified in higher dimensions.

5. Orthogonal Variational PCA

The above mentioned problems can be solved by restriction of the searched space and careful choice of the initial conditions. The existence of methods for rotation of orthogonal solution to match the additional restrictions allows us to restrict the space of parameters to orthogonal decompositions.

5.1. Orthogonal model

The ambiguous parameterisation of mean value by product AX in the model (3) is replaced by svd decomposition of the mean value,

$$f(D|A, L, X) = \mathcal{N}(ALX, \sigma I_p \otimes I_n) \quad (19)$$

where matrices $A \in \mathbb{R}^{n \times p}$ and $X \in \mathbb{R}^{p \times p}$ are now restricted to be orthogonal, $A'A = I_p$, $XX' = I_p$. Matrix L is diagonal with non-negative elements, $l_i = L_{i,i}, i = 1, \dots, p$. The rank of mean value is no longer given by the size of matrices A and X but by the number of nonzero elements l .

Similar model was used for the solution based on Laplace approximation [11]. In this paper we provide the solution based on the variational approach.

5.2. Variational Bayes solution

Solution of the new model (??) follows methodology of variational estimates described in section 4.1. The first step of any Bayesian method is to set up prior distributions. Matrices A, X are restricted to be orthogonal, i.e. living on Stiefel manifold with finite volume and thus we can choose prior distribution to be uniform. This space has finite volume and thus uniform distribution has a normalising constant given by [9]. However, its exact value does not influence further evaluations and we do not state it explicitly. Prior information for the remaining parameters is chosen as

$$\begin{aligned} f(l_i) &= \mathcal{N}(0, c_0^{-1}) \\ f(\sigma) &= \Gamma(a_0, b_0) \end{aligned} \quad (20)$$

where parameters a_0, b_0, c_0 are chosen as very small values. Non-informative prior on those variables provides almost identical results.

Direct application of the theorem 4.1 gives us the following results

$$f(A) = \mathcal{M}(\overline{LX}D'\overline{\sigma}) \quad (21)$$

$$f(X) = \mathcal{M}(D'\overline{AL}\overline{\sigma}) \quad (22)$$

$$f(l) = \mathcal{N}(\bar{l}, \Sigma_L) \quad (23)$$

$$f(\sigma) = \Gamma(a, b). \quad (24)$$

We can see that minimum of KL distance to joint likelihood is reached for posterior distribution of matrices A and X in the von Mises-Fisher type, \mathcal{M} , which is a form of normal distribution with orthogonal restrictions.

Moments for variables l and σ are quite easy to evaluate, $\overline{\sigma}$ denotes mean value of variable σ (same for other variables) and Σ_L denotes variance of L

$$\begin{aligned} \bar{l} &= \frac{\overline{\sigma}}{\overline{\sigma} + c} \text{diag}(\overline{X}D'\overline{A}) \\ \Sigma_L &= (\overline{\sigma} + c)^{-1} I_r \\ \overline{\sigma} &= \frac{a}{b} \\ a &= a_0 + \frac{pN}{2}. \\ b &= b_0 + \frac{1}{2} \left[\text{tr}(D'D) - 2\text{tr}(\overline{X}'\overline{L}\overline{A}'D) + \text{tr}(\overline{L}\overline{L} + \Sigma_L) \right] \end{aligned} \quad (25)$$

However, moments of von Mises-Fisher distribution are more complicated. Fortunately, we need to evaluate only the first moment, (mean values \overline{A} and \overline{X}), for which an analytical solution is known [5, 9] but hard to evaluate, as discussed in Section 5.3. However, analytical results for von-Mises-Fisher distribution provide an important insight into the problem.

Theorem 5.1 (Moments of von Mises-Fisher distribution)
Let $X \sim \mathcal{M}(F)$ be distributed as von Mises-Fisher with given parameter F , and $F = \tilde{U}\tilde{S}\tilde{V}$ be svd decomposition of matrix F . Then mean value of X ,

$$\overline{X} = \tilde{U}\tilde{S}\tilde{V} \quad (26)$$

where \tilde{S} is a diagonal matrix with elements

$$s_i = \mathcal{L}(p, \tilde{S}) = \frac{\partial}{\partial s_i} \log \left({}_0F_1 \left(\frac{1}{2}p, \frac{1}{4}\tilde{S}^2 \right) \right) \quad (27)$$

where ${}_0F_1$ denotes the generalised hypergeometric function [7]. This result is due to [9].

This theorem has very useful consequences for the whole iteration algorithm. We adopt the iterative scheme of VPCA and assign the new estimates as values of functions (25) based on the previous estimates. Let us assume that parameter estimates $\overline{A}^{(old)}$ and $\overline{X}^{(old)}$ are in the form

$$\begin{aligned} \overline{A}^{(old)} &= US_A^{(old)} \\ \overline{X}^{(old)} &= S_X^{(old)}V \end{aligned} \quad (28)$$

where orthogonal matrices U and V are singular vectors of the data matrix $D = US_DV$. Matrices S_A , S_X and S_D are diagonal. Using equations (21), (22) we get

$$\begin{aligned} D\overline{X}'^{(old)}\overline{L}^{(old)}\overline{\sigma}^{(old)} &= US_DVV'S_X^{(old)}\overline{L}^{(old)}\overline{\sigma}^{(old)} \\ \overline{\sigma}^{(old)}\overline{L}^{(old)}\overline{A}'^{(old)}D &= \overline{\sigma}^{(old)}\overline{L}^{(old)}S_A^{(old)}U'US_DV \end{aligned}$$

which together with the theorem 5.1 gives new estimates

$$\begin{aligned} \overline{A}^{(new)} &= U\mathcal{L}(p, S_DS_X^{(old)}\overline{L}^{(old)}\overline{\sigma}^{(old)}) \\ \overline{X}^{(new)} &= \mathcal{L}(n, \overline{\sigma}^{(old)}\overline{L}^{(old)}S_A^{(old)}S_D)V \end{aligned} \quad (29)$$

which preserves the assumed form (28). This allows us to replace iterative algorithm of \overline{A} , \overline{X} by iterations of S_A , S_X .

This result has very advantageous consequences

- It reduces number of iteratively estimated parameters from $[p(3p + n + 2) + 2]$ of VPCA to $[3p + 2]$ of OVPCA, thus making it completely independent of n which is usually much greater than p .
- Taking advantage of (28) we can start at singular value decomposition of the data, i.e. classical PCA. $\overline{A}_0 = U$, $\overline{X}_0 = V$, $\overline{L}_0 = S$, $S_A = I_p$, $S_X = I_p$. The most sensitive parameter to choice of initial value is the variance $\overline{\sigma}$. The safest way is to choose the smallest eigenvalue, if it is distinct enough from zero. Otherwise heuristic suggestions like mean of last few eigenvalues can be used.
- The convergence of the method is very fast and robust. Stopping rule on increments of parameter \bar{l} could be set close to machine precision.
- The method does not involve any numerically sensitive operations like inversions.
- The method was derived with priors (20) however, they do not have any regularising effects. Values $a_0 = b_0 = c_0 = 0$ can be safely used.

The most problematic point remains operation (27) because evaluation of hypergeometric function of matrix argument is far from trivial [7].

5.3. Approximations of hypergeometric function

The original approximation given in [9] is based on assumption of all singular values being large. It has proven to be fast but inaccurate. That is why we have developed our own approximation. The evaluation is complicated by the fact that function (27) is a function of hypergeometric function ${}_0F_1(p, \tilde{S})$ of the matrix argument which is expressible in terms of zonal polynomials, [7]. No analytical simplification is known to us.

On the other hand, the logarithm of the same hypergeometric function, ${}_0F_1(p, s)$, of a scalar argument can be expressed as a ratio of Bessel functions, \mathcal{B} ,

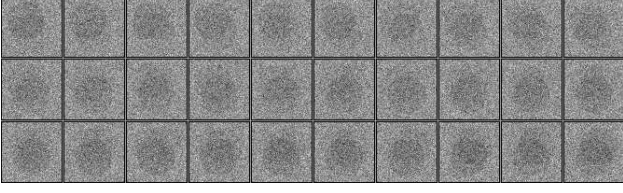


Figure 2. Dynamic image sequence for synthetic data, SNR=1.

$$d_i = \frac{\mathcal{B}(\frac{1}{2}p, s_i)}{\mathcal{B}(\frac{1}{2}p - 1, s_i)}. \quad (30)$$

Moreover, the formula (30) could be transformed to a continuous fraction expansion given in [1]. The evaluation of this p.f.e. converge very fast for $p > s_i$ however when $s_i \gg p$ (say $s_i > 10p$) the convergence is quite slow. It is faster to replace it by Taylor expansion at $s_i \rightarrow \infty$

$$d_i = 1 - \left(\frac{p-1}{2s_i} \right) \exp\left(-\frac{p-3}{4s_i}\right) + o(5)$$

which gives excellent approximation in such a case.

This motivates our choice to approximate function (27) as differential of a scalar variable

$$\mathbf{d}_i = \mathcal{L}(p, S) \approx d_i = \frac{\partial}{\partial s_i} \log \left({}_0F_1 \left(\frac{1}{2}p, \frac{1}{4}s_i^2 \right) \right) \quad (31)$$

where s_i is i -th singular value of the data matrix D .

Recently, a very good approximation of ${}_0F_1$ of matrix argument was developed, [4]. It is based on the Laplace approximation in the saddle point. The authors provide two solutions calibrated and non-calibrated. For our purpose the calibrated one gives better results. Results based on the saddle point approximation are almost identical with our approximation but the convergence is slower and its evaluation is more computationally demanding.

6. Application

Both variational method described above were applied to the estimation of the number of relevant underlying images in functional analysis of dynamic image data. Synthetic data were generated by the model (1) with values $n = 4096$, $r = 3$, $p = 60$ and various levels of signal-to-noise ratio. The original factor images, A_0 , and factor curves, X , are displayed in Figure 3 (left). An image sequence with SNR=1 generated from this model is displayed in Figure 2.

Results of estimation based on the sequence are displayed in Figure 3 (right) for comparison. The results were rotated to the original space using

$$X_{disp} = X_0 \overline{X}' \left(\overline{X} \overline{X}' \right)^{-1} \overline{X}$$

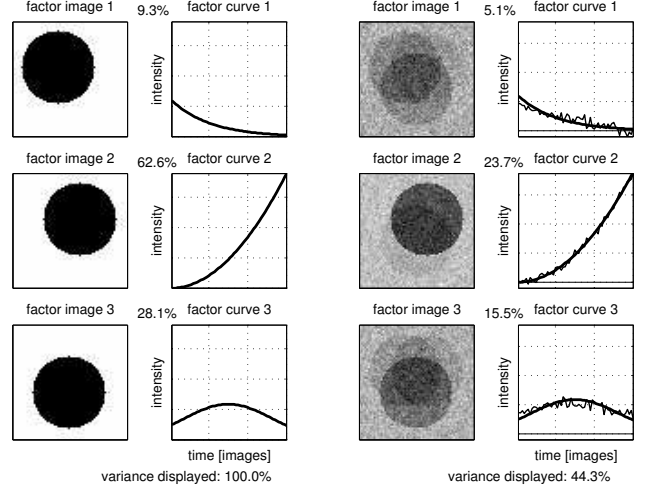


Figure 3. Illustration of results of functional analysis of dynamic image sequence. Simulated phantom (left) and estimated factors (right). Estimated factors were rotated to the space of simulated data for comparison.

SNR	order \hat{r}		CPU time [s]	
	VPCA	OVPCA	VPCA	OVPCA
2000	9	4	21243	1
100	3	3	6299	1
20	3	3	2496	1
5	3	3	891	1
1	3	3	207	2

Table 1. Comparison of VPCA and OVPCA in terms of speed and

where X_0 is the original simulation and X_{disp} is result displayed in the Figure. The work was performed on a 1GHz Pentium III machine. Comparison of typical execution times of both methods is listed in the Table 1.

Both methods were effective in estimating the number, \hat{r} , of underlying factor images. The significant differences in CPU time can be explained by the dependence of VPCA on n and on indeterminacy (6) of its model, as discussed in Section 4.2. For the low intensity noise (SNR=2000) the expected precision of estimates c/d is very high. Thus the shift of parameters in one iteration is tiny. It explains extremely high computational burden for low intensity noise.

Figure 4 shows the typical evolution of parameters determining number of underlying factors (L for OVPCA and α for VPCA). Evolution of L is very fast with reliable convergence. Evolution of α is significantly slower and less reliable. Moreover it seems to drift to higher values. These artefacts can be explained by as consequence of model indeterminacy (6). Chosen prior distributions of VPCA are designed to deal with this problem, but they are not strong enough to prevent transitions to other allowed rotations, T .

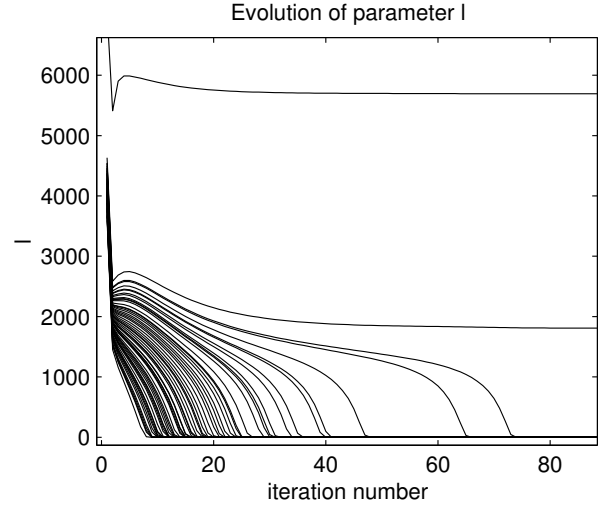
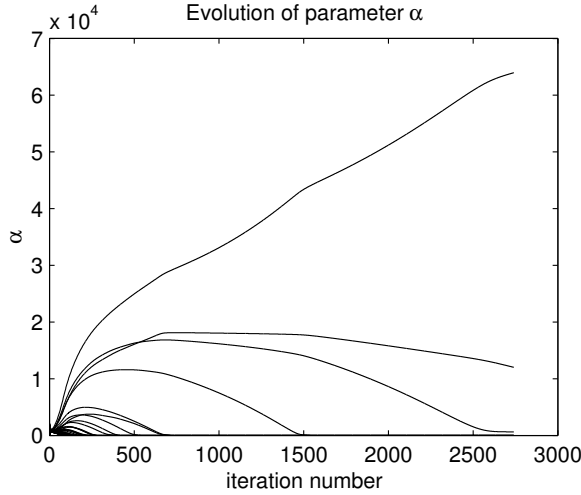


Figure 4. Comparison of evolution of hyper-parameters α of VPCA (left) and parameters l of OVPCA (right). These parameters determine estimated number of underlying factors, in each method, allowing to be compared.

7. Conclusion

In this paper we have introduced a method for approximate Bayesian solution of the restricted factor analysis model. The model was re-parameterised in order to assure uniqueness of solution. The resulting estimation method has advantages that are beneficial both for existing solutions and for further development of dimensionality reduction and rank deficiency problems in general.

The resulting point estimates of OVPCA are strongly related to results of PCA (8)

PCA	OVPCA
$AX = U_r S_r V_r$	$AL\bar{X} = U\Lambda V$
	$\Lambda = S_A \bar{L} S_X$

In other words, orientational part of the solution is identical, but estimates of singular values differ. Estimates obtained using OVPCA already reveals estimated dimensionality. Hence, OVPCA could be used in any application where PCA is successfully used, e.g. data compression, de-noising, pattern recognition, shape analysis and spectral analysis. It provides the same level of

numerical stability as no inversions or numerically sensitive operations except *svd* decomposition are used.

robustness no regularisers or tuning parameters are required.

ease of use the method does not require any extra knowledge than that one used for PCA.

Moreover, the solution of the model is Bayesian, i.e. posterior distribution of parameters are provided. This allows testing of various hypothesis and further investigation of the model.

7.1. Acknowledgement

We would like to thank Prof. Butler and Prof. Wood for electronic version of their article. This work was supported by EU Grant, IST-12058.

References

- [1] M. Abramowitz and I. Stegun. *Handbook of mathematical functions*. Dover Publications, Inc., New York, 1972.
- [2] T. W. Anderson. *An introduction to multivariate statistical analysis*. John Wiley and Sons, 1971.
- [3] C. M. Bishop. Variational principal components. In *Proceedings of Ninth International Conference on Artificial Neural Networks*, 1999.
- [4] R. W. Butler and T. A. Wood. Laplace approximation of bessel function of matrix argument. *Journal of Computational and Applied Mathematics*, 2003.
- [5] T. D. Downs. Orientational statistics. *Biometrika*, 59:665–676, 1972.
- [6] G. Golub and C. VanLoan. *Matrix Computations*. The John Hopkins University Press, Baltimore – London, 1989.
- [7] A. T. James. Distribution of matrix variates and latent roots derived from normal samples. *Annals of Mathematical Statistics*, 35:475–501, 1964.
- [8] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.
- [9] C. G. Khatri and K. V. Mardia. The von Mises-Fisher distribution in orientation statistics. *Journal of Royal Statistical Society B*, 39:95–106, 1977.
- [10] H. v. M. Šámal, M. Kárný, E. Maříková, and Z. Dienstbier. Rotation to simple structure in factor analysis of dynamic radionuclide studies. *Physics in Medicine and Biology*, 32(3):371–382, 1987.
- [11] T. P. Minka. Automatic choice of dimensionality for PCA. Technical report, MIT, 2000.
- [12] J. W. Miskin. *Ensemble Learning for Independent Component Analysis*. PhD thesis, University of Cambridge, 2000.