Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Václav Šmídl, Anthony Quinn[1]

# Variational Bayes for Orthogonal Probablistic Principal Component Analysis

[1]Trinity College Dublin

2125 10th February 2005

**1ET 100 750 401, GACR 102/03/0049**

ÚTIA AVČR, P.O.Box 18, 182 08 Prague,
Czech Republic
Fax: (+420)286890378, http://www.utia.cas.cz, E-mail: utia@utia.cas.cz

# 1 Introduction

Principal Component Analysis (PCA) is one of the classical data analysis tools for dimensionality reduction. It is used in many application areas including data compression, de-noising, pattern recognition, shape analysis and spectral analysis. For an overview of its use, see (Jolliffe 2002).

PCA is often used as a black box numerical tool, because of its mature numerical implementation and ease of use. The correspondence between Principal Components (PCs) and the eigenvectors of a symmetric matrix makes it intuitively appealing. However, further investigation of PCs—namely the choice of an optimal number of relevant PCs and confidence intervals on the PCs—is complicated. Many approximate solutions are available, both formal and *ad hoc* (Anderson 1971; Jolliffe 2002). Simple *ad hoc* criteria are used in applications such as data compression and de-noising, where the number of PCs is restricted by an available bit-rate or computational cost, rather than by statistical relevance. However, formal solutions are required in applications where uncertainty of estimates is an important part of the problem. A typical example is spectral analysis (Kay 1993) or functional analysis of dynamic image data (Buvat et al. 1998).

Traditionally, probability distributions for PCs were derived using sampling theory (Jolliffe 2002). These results are mostly asymptotic. Recently, the problem was addressed using Bayesian methodology (Bishop 1999), by invoking the Factor Analysis (FA) model with isotropic Gaussian noise. However, the FA model does not impose restrictions of orthogonality, and so posterior results are not identical to PCA. Moreover, rotational ambiguity in the FA model presents a computational difficulty that must be overcome by means of regularizing priors.

In this paper, we review the original concept of PCA (Section 2) as well as probabilistic models which yield a Maximum Likelihood (ML) solution identical to that of PCA (Section 3). The ML solution does not provide an estimate of rank nor uncertainty bounds for the model estimates. Therefore, the problem is reformulated using the Bayesian methodology (Section 4). A variational approximation of the posterior distribution is investigated, and a numerically efficient algorithm—Orthogonal Variational PCA (OVPCA)—for estimation of an approximating posterior distribution is presented (Section 5). Further analysis yields both the distribution of rank, as well as uncertainty bounds on PCs (Section 6). The performance of the method is illustrated via a simple simulation study. A contemporary application in medical imaging is also presented. The numerical efficiency of the OVPCA technique depends on the approach taken to evaluating the involved hypergeometric function of matrix argument, ${}_0F_1$. A novel approximation of this function—yielding results of acceptable accuracy and computational cost—is presented in Appendix B.

Throughout the paper, we will use the following notational conventions:

| | |
|---|---|
| $\Re$ | set of real numbers. |
| $A \in \Re^{n\times m}$ | matrix of dimensions $n \times m$, generally denoted by a capital letter. |
| $\boldsymbol{a}_i$ | $i$th column of matrix $A$, $i = 1 \dots m$, using bold-face letter. |
| $a_{i,j}$, $a_{D,i,j}$ | $(i,j)$th element of matrix $A$, $A_D$, respectively, $i = 1 \dots n$, $j = 1 \dots m$. |
| $a_i$, $a_{D,i}$ | $i$th element of vector $\boldsymbol{a}$, $\boldsymbol{a}_D$, respectively. |
| $A_{;r}$, $A_{D;r}$ | operator selecting the first $r$ columns of matrix $A$, $A_D$, respectively. |
| $A_{;r,r}$, $A_{D;r,r}$ | operator selecting the $r\times r$ upper-left sub-block of matrix $A$, $A_D$, respectively. |
| $\boldsymbol{a}_{;r}$, $\boldsymbol{a}_{D;r}$ | operator extracting upper length-$r$ sub-vector of vector $\boldsymbol{a}$, $\boldsymbol{a}_D$, respectively. |
| $A_{(r)} \in \Re^{n\times m}$ | subscript $_{(r)}$ denotes matrix $A$ with restricted rank, $\mathrm{rank}\,(A) = r \leq \min(n,m)$. |
| $I_r \in \Re^{r\times r}$ | square identity matrix . |
| $\mathbf{1}_{p,q}$, $\mathbf{0}_{p,q}$ | matrix of size $p \times q$ with all elements equal to one, zero, respectively. |
| $\mathrm{diag}\,(\cdot)$ | two distinct meanings, clearly distinguished by the context:<br>(i) $\boldsymbol{a} = \mathrm{diag}\,(A)$, $A \in \Re^{n\times m}$, then $\boldsymbol{a} = [a_{1,1}, \dots, a_{q,q}]'$, $q = \min(n,m)$<br>(ii) $A = \mathrm{diag}\,(\boldsymbol{a})$, $\boldsymbol{a} \in \Re^q$, then $a_{i,j} = \begin{cases} a_i & \text{if i=j} \\ 0 & \text{if i}\neq\text{j} \end{cases}$, $i,j = 1,\dots,q$. |
| $\mathrm{tr}\,(A)$ | trace of matrix $A$. |

| | |
|---|---|
| $A = U_A L_A V_A'$ | Singular Value Decomposition (SVD) of matrix $A \in \Re^{n\times m}$, where $U_A \in \Re^{n\times q}$, $L_A \in \Re^{q\times q}$, $V_A \in \Re^{m\times q}$, $q = \min(n, m)$. Therefore, in this paper, the SVD is expressed in 'economic' form, i.e. in terms of the only guaranteed-non-zero part of $L_A$, namely the upper-left $q \times q$ diagonal sub-matrix. |
| $[A \otimes B] \in \Re^{np\times mq}$ | Kronecker product of matrices $A \in \Re^{n\times m}$ and $B \in \Re^{p\times q}$, such that $$A \otimes B = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,m}B \\ \vdots & \ddots & \vdots \\ a_{n,1}B & \cdots & a_{n,m}B \end{bmatrix}.$$ |
| $[A \circ B] \in \Re^{n\times m}$ | Hadamard product of matrices $A \in \Re^{n\times m}$ and $B \in \Re^{n\times m}$, such that $$A \circ B = \begin{bmatrix} a_{1,1}b_{1,1} & \cdots & a_{1,m}b_{1,m} \\ \vdots & \ddots & \vdots \\ a_{n,1}b_{n,1} & \cdots & a_{n,m}b_{n,m} \end{bmatrix}.$$ |
| $\chi(\cdot)$ | the indicator (characteristic) function on the argument set. |
| $\mathrm{erf}(x)$ | error function: $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\, dt$. |
| ${}_0F_1(a, AA')$ | hypergeometric function, ${}_pF_q(\cdot)$, with $p = 0$, $q = 1$, scalar parameter $a$, and symmetric matrix parameter $AA'$. |
| $\Pr(\cdot)$ | probability of argument . |
| $f(x\|\theta)$ | probability density function (pdf) of continuous random variable $x$, conditioned by known $\theta$. |
| $\hat{\theta}$ | maximizer of $f(x\|\theta)$, with latter taken as a function of $\theta$ (the ML estimate). |
| $\mathsf{E}_x[\cdot]$ | expected value of argument with respect to pdf $f(x)$. |
| $\widehat{g(x)}$ | simplified notation for $\mathsf{E}_x[g(x)]$. |
| $\overline{x}, \underline{x}$ | upper bound, lower bound of random variable $x$. |
| $\mathcal{N}(\mu, s^2)$ | Normal distribution with mean value, $\mu$, and variance, $s^2$. |
| $t\mathcal{N}(\mu, s^2; (a, b])$ | Normal of type $\mathcal{N}(\mu, s^2)$, confined to support $(a, b]$. |
| $\mathcal{M}(F)$ | von-Mises-Fisher distribution with matrix parameter $F$. |
| $\mathcal{G}(a, b)$ | Gamma distribution with parameters $a$ and $b$. |
| $\mathcal{U}(\cdot)$, $\mathcal{U}((a, b])$ | Uniform distribution on the argument set, on interval $(a, b]$, respectively. |

## 2 Principal Component Analysis (PCA)

PCA is a widely used tool (Jolliffe 2002) for representation of data sets. Specifically, we consider a set of $n$ $p$-dimensional data vectors, $d_i$, from data space $\mathcal{D}$:

$$D = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n], \quad \boldsymbol{d}_i \in \mathcal{D} = \Re^p.$$

For simplicity, we assume that the sample mean vector $\langle \boldsymbol{d} \rangle_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{d}_i$ is zero. If $\langle \boldsymbol{d} \rangle_n \neq 0$, it can be subtracted from the raw data in a pre-processing step. Explicit modelling of the sample mean is discussed in Section 7.3. We also assume, without loss of generality, that $p \leq n$.

Let $\mathcal{P}_r$ be the orthogonal projection operator from $\mathcal{D}$ into the $r$-dimensional subspace, $\mathcal{A}_r$, with orthonormal basis $W_r = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_r] \in \Re^{p\times r}$:

$$\begin{aligned} \mathcal{P}_r : \quad \mathcal{D} &\rightarrow \mathcal{A}_r, \\ \boldsymbol{d}_i &\rightarrow \boldsymbol{m}_i. \end{aligned}$$

Then, $M_{(r)} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n]$ is the image of $D$ under $\mathcal{P}_r$, inspiring the following decomposition:

$$\begin{aligned} D &= M_{(r)} + E, &(1)\\ &= W_r\left(W_r' D\right) + E. &(2) \end{aligned}$$

Here, $E = [\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n]$ is the matrix of residuals, $\boldsymbol{e}_i = \boldsymbol{d}_i - \boldsymbol{m}_i$, $i = 1, \ldots, n$.

Consider a (possibly unique) space, $\mathcal{A}_r^*$, for which the variation of the projected image, $M_{(r)}^*$, is maximized:

$$\mathcal{A}_r^* = \arg\max_{\mathcal{A}_r} \left( \mathrm{tr} \left( M_{(r)} M_{(r)}' \right) \right). \tag{3}$$

This space is, of course, determined by $D$, or, more specifically, by the sample covariance matrix of $D$:

$$S = \frac{1}{n-1} \sum_{i=1}^{n} \boldsymbol{d}_i \boldsymbol{d}_i' = \frac{1}{n-1} DD'. \tag{4}$$

The eigendecomposition of $S \in \Re^{p \times p}$ is:

$$S = U \Lambda U'. \tag{5}$$

$U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p]$ is an orthogonal matrix of eigenvectors, such that $U'U = I_p$. $\Lambda = \mathrm{diag}\,(\boldsymbol{\lambda})$ is a diagonal matrix of eigenvalues, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_p]'$. For the purpose of this work, it is assumed, without loss of generality, that

$$\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0. \tag{6}$$

Then, decomposition (5) is unique—up to the sign of each eigenvector, $\boldsymbol{u}_i$—since $\boldsymbol{u}_i \boldsymbol{u}_i' = (-\boldsymbol{u}_i)(-\boldsymbol{u}_i)'$. Thus, there are $2^p$ possible decompositions (5) satisfying the stated constraints, all equal to within a sign. It is possible to show (Hotelling 1933) that for any positive $r \leq p$, the first $r$ eigenvectors,

$$U_{;r} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r], \tag{7}$$

form an orthonormal basis of the maximum variation space, $\mathcal{A}_r^*$ (3). Furthermore, (6) ensures that $\mathcal{A}_r^*$ is uniquely determined (the sign ambiguity notwithstanding). The method is known as *Principal Component Analysis* (PCA) of $D$ (Hotelling 1933):

$$\begin{aligned} D &= M_{(r)}^* + E^*, & (8) \\ &= U_{;r} \left( U_{;r}' D \right) + E^*, & (9) \end{aligned}$$

where $M_{(r)}^*$ and $E^*$ are the image and residual defined with respect to $\mathcal{A}_r^*$. The $r$ columns of $U_{;r}$ (7) are called the first $r$ *principal components (PCs)* of $D$.

### 2.1 Least Squares (LS) Interpretation

The chosen additive decomposition (8) was found by optimization of a property (3) of the projected image (i.e. of the 'signal' part of $D$, namely $M_{(r)}$). Alternatively, we can optimize the additive decomposition (1) in terms of the projection trajectories (i.e. the 'noise', $E$) by minimizing *its* variance. It is true that

$$M_{(r)}^* = \arg\min_{M_{(r)}} \left( \mathrm{tr} \left( EE' \right) \right),$$

where $M_{(r)}$ is defined in (1), (2), and $M_{(r)}^*$ is given by (8), (9). Thus, the LS criterion is equivalent to the maximum variation criterion (3), identifying the *same* subspace, $\mathcal{A}_r^*$, of $D$. This LS interpretation of PCA is the earliest (Pearson 1901).

## 3 Rank-Restricted Modelling

Guided by the interpretation above, we now analyze the additive decomposition model (1) as a rank-restricted signal and noise separation problem. The additive degeneracy in (1) is overcome (i.e. 'regularized') by modelling explicitly the properties of $M_{(r)}$ and $E$. Since the rank-restricted LS optimization of Section 2.1 is equivalent to Maximum Likelihood (ML) estimation under the uncorrelated Gaussian noise assumption (Kay 1993), we design the model so as to yield (8) as its ML estimate. Hence the following assumptions are appropriate:

**The Noise**

$E$ is a normally distributed random matrix:

$$E \sim f\left(E|\omega\right) = \mathcal{N}\left(\mathbf{0}_{p,n}, \omega^{-1} I_p \otimes I_n\right), \tag{10}$$

where scalar $\omega > 0$ denotes precision, and other symbols have their usual meaning (Section 1). (10) is known as the isotropic Gaussian noise model (Tipping and Bishop 1998b).

**The Signal**

Since the columns of $M_{(r)}$ exist in a lower dimensional space, $\mathcal{A}_r$, it follows that $\text{rank}\left(M_{(r)}\right) = r$. This fact permits $M_{(r)}$ to be expressed via the 'economic' Singular Value Decomposition (SVD) (Golub and VanLoan 1989):

$$M_{(r)} = A_r L_r X_r'. \tag{11}$$

Since the rank $r$ of the matrix $M_{(r)}$ is known, we can restrict matrices $A_r$ and $X_r$ to $\Re^{p\times r}$ and $\Re^{n\times r}$ respectively, with orthogonality restrictions $A_r'A_r = I_r$, $X_r'X_r = I_r$. Also $L_r = \text{diag}\left(\boldsymbol{l}_r\right) \in \Re^{r\times r}$ is a diagonal matrix of non-zero *singular values,* $\boldsymbol{l}_r = [l_1, \ldots, l_r]'$. Analogously to (6), we assume that

$$l_1 > l_2 > \ldots > l_r > 0. \tag{12}$$

The decomposition (11) is unique, up to the sign of the $r$ singular vectors, (i.e. there are $2^r$ possible decompositions (11) satisfying the stated constraints, all equal to within a sign ambiguity).

Model (1), extended by (10), (11), yields:

$$f\left(D|A_r, L_r, X_r, \omega, r\right) = \mathcal{N}\left(A_r L_r X_r',\ \omega^{-1} I_p \otimes I_n\right). \tag{13}$$

The ML solution for the model parameters, conditioned by known $r$, is given by

$$\left(\hat{A}_r, \hat{L}_r, \hat{X}_r, \hat{\omega}\right) = \arg \max_{A_r, L_r, X_r, \omega} f\left(D|A_r, L_r, X_r, \omega, r\right),$$

with assignments

$$\hat{A}_r = U_{D;r}, \quad \hat{L}_r = \text{diag}\left(\boldsymbol{l}_{D;r}\right), \quad \hat{X}_r = V_{D;r}, \quad \hat{\omega} = \frac{pn}{\sum_{i=r+1}^{p} l_{D,i}^2}. \tag{14}$$

Here, $U_{D;r}$, and $V_{D;r}$ are the first $r$ columns of the matrices $U_D$, and $V_D$ respectively, obtained from the SVD

$$D = U_D L_D V_D'. \tag{15}$$

Finally, $L_D = \text{diag}\left(\boldsymbol{l}_D\right)$, where $\boldsymbol{l}_D = [l_{D,1}, \ldots, l_{D,p}]'$ are the singular values of $D$.

**Remark 1 (Parametric Components)** *Since $\hat{A}_r$ (14) is numerically identical to Principal Components (PCs), $U_{D;r}$, we will therefore interpret the columns of $A_r$ (11) as* parametric components, *i.e. a probability modelling counterpart to classical PCs.*

**Remark 2** *From (15):*

$$DD' = U_D L_D V_D' V_D L_D U_D' = U_D L_D L_D U_D'.$$

*Hence, from (4), and (5), it follows that*

$$U_D = U, \quad L_D = (n-1)^{\frac{1}{2}} \Lambda^{\frac{1}{2}}. \tag{16}$$

*From (11), (14), the ML signal inference is*

$$\hat{M}_{(r)} = U_{D;r}\left(U_{D;r}' D\right) = M_{(r)}^*. \tag{17}$$

*Hence, the rank-r ML signal inference is, by design, equal to the LS (maximum variation) estimate, $M_{(r)}^*$ (8), (9).*

**Remark 3 (Factor Analysis (FA) model)** *PCA is usually interpreted as ML estimation of the FA model (Anderson 1971; Tipping and Bishop 1998b); i.e.:*

$$M_{(r)} = A_r X_r, \tag{18}$$

*in (1), with $A \in \Re^{p\times r}$, $X \in \Re^{r\times n}$ being arbitrary real matrices. The zero-mean assumption on the columns, $\boldsymbol{m}_i$, of $M_{(r)}$, is again adopted for clarity. The conditional distribution of data is now*

$$f\left(D|A_r, X_r, \omega, r\right) = \mathcal{N}\left(A_r X_r, \omega^{-1} I_p \otimes I_n\right). \tag{19}$$

*The ML estimate of $M_{(r)}$ is, once again, identical to that from (14), i.e. $\hat{M}_{(r)} = M^*_{(r)}$ (17).*

*However, ML estimates of $A_r$ and $X_r$ from (18) are* not *unique, because (18) exhibits multiplicative degeneracy in the sense that:*

$$\hat{M}_{(r)} = \hat{A}_r \hat{X}_r = \left(\hat{A}_r T_r\right)\left(T_r^{-1}\hat{X}_r\right) = \tilde{A}_r \tilde{X}_r, \tag{20}$$

*for any arbitrary invertible matrix, $T_r \in \Re^{r\times r}$. This is known as* rotational ambiguity *in the factor analysis literature (Anderson 1971). The ML solution of model (18) is therefore*

$$\begin{aligned} \hat{A}_r &= U_{D;r} T_r, \\ \hat{X}_r &= T_r^{-1} L_{D;r,r} V'_{D;r}. \end{aligned}$$

*Thus, the unique optimal subspace, $\mathcal{A}^*_r$ (3), is once again inferred, but the oriented orthonormal basis of $\mathcal{A}^*_r$ corresponding to PCA—i.e. $U_{;r} = U_{D;r}$ (7), (16)—is revealed only for the choice $T_r = I_r$.*

**Remark 4** *Published solutions to the FA model estimation problem (Tipping and Bishop 1998b; Anderson 1971), do not directly maximize the model (19), but complement (19) by a Gaussian prior on $X_r$, $f\left(X_r\right) = \mathcal{N}\left(\mathbf{0}_{r,n}, I_r \otimes I_n\right)$ and marginalize over $X_r$. The resulting maximum of the marginal likelihood, conditioned by $r$, is then reached for $\hat{\omega}$ given by (14) and*

$$\hat{A}_r = U_{D;r}\left(\Lambda_{;r,r} - \hat{\omega}^{-1} I_r\right)^{\frac{1}{2}} R_r. \tag{21}$$

*Here, $U_{D;r} = U_{;r}$, $\Lambda_{;r,r}$ are given by (5), (7), $\hat{\omega}$ by (14), and $R_r \in \Re^{r\times r}$ is any orthogonal (i.e. rotation) matrix. In this case, indeterminacy of the model is reduced from an arbitrary invertible matrix, $T_r$ (20), to an orthogonal matrix, $R_r$. This reduction is a consequence of restricting the model by the prior on $X_r$. Once again, the same optimal subspace, $\mathcal{A}^*_r$, is identified (3), and the ML solution of the FA model is any orthogonal set (21) spanning $\mathcal{A}^*_r$.*

### 3.1 Open problems

With respect to the solutions above, i.e. (8), (14) and (21), we note the following:

1. All are conditioned by knowledge of the dimension, $r$, of the signal space, $\mathcal{A}_r$. ML estimation of $r$ is unsuccessful, since its likelihood is strictly increasing with $r$, reaching its maximum at $\hat{r} = p$. This insensitivity of ML solutions to Ockham's Razor is widely known (Quinn 1998).

2. The ML solution yields point estimates of parameters. Distributions for quantities $A_r$, $L_r$, $X_r$, and $\omega$ (10), (11) are available, but only in conditional form. For example,

$$f\left(A_r|D, \hat{L}_r, \hat{X}_r, \hat{\omega}, r\right) \propto f\left(D|A_r, L_r = \hat{L}_r, X_r = \hat{X}_r, \omega = \hat{\omega}, r\right),$$

   with $\hat{L}_r, \hat{X}_r, \hat{\omega}$ given by (14). This approach, however, neglects the uncertainty in the conditioning parameters, $L_r$, $X_r$, $\omega$.

Both of these problems can be solved by extending the ML framework, presented above, using the Bayesian methodology.

# 4 Bayesian Treatment of the Rank-Restricted Model

Bayesian solutions of rank-restricted models have usually been based on the Factor Analysis model (18) (Press and Shigemasu 1989; Bishop 1999). However, the rotational ambiguity (20) presents a difficulty for estimation. Therefore, priors on $A_r$ and $X_r$ must be used to restrict the optimizers and reach a solution. Even then, the estimation process is difficult, and approximate iterative evaluation techniques such as MCMC (Rowe and Press 1998; Press and Shigemasu 1989) and Variational Bayes (Bishop 1999) are necessary. Furthermore, estimation in the context of these flat posterior densities negatively influences the speed of convergence.

We have already noted that the orthogonal rank-restricted model (13) forces $\mathcal{A}_r^*$ to be spanned, in the ML procedure (14), by an orientated, orthonormal basis, $U_{D;r}$ (7), (16), leaving only a countable sign-based ambiguity. We exploit this rotational selectivity in this paper and seek an approximate Bayesian solution for the orthogonal rank-restricted model. However, this advantage is gained at the expense of orthogonal restrictions which are generally difficult to handle. Specifically, parameters $A_r$ and $X_r$ (11) are now restricted to having orthonormal columns, i.e. $A_r'A_r = I_r$ and $X_r'X_r = I_r$ respectively. Intuitively, each column $\boldsymbol{a}_i, i = 1 \ldots r$, of $A_r$ belongs to the unit hyperball in $p$ dimensions, i.e. $\boldsymbol{a}_i \in \mathcal{H}_p$. Hence, $A_r \in \mathcal{H}_p^r$, the Cartesian product of $r$ $p$-dimensional unit hyperballs. However, the requirement of orthogonality—i.e. $\boldsymbol{a}_i'\boldsymbol{a}_j = 0,\ \forall i \neq j$—confines the space further. The orthonormally constrained subset, $\mathcal{S}_{p,r} \subset \mathcal{H}_p^r$ is known as the Stiefel manifold (Khatri and Mardia 1977). Therefore, both the prior and posterior distributions of $A_r$ have a support confined to $\mathcal{S}_{p,r}$.

The posterior distribution is obtained via Bayes' rule:

$$f\left(A_r, L_r, X_r, \omega | D, r\right) \propto f\left(D | A_r, L_r, X_r, \omega, r\right) f\left(A_r, L_r, X_r, \omega, r\right), \tag{22}$$

where, for the present, $r$ is assumed known *a priori*. Priors on parameters are chosen to be mutually independent and as non-committal as possible.

Orthogonally constrained parameters $A_r$ and $X_r$ are confined to $\mathcal{S}_{p,r}$ and $\mathcal{S}_{n,r}$, respectively. The finite area, $C\left(p, r\right)$, of $\mathcal{S}_{p,r}$ is given by (Khatri and Mardia 1977):

$$C\left(p, r\right) = \frac{2^r \pi^{\frac{1}{2}pr}}{\pi^{\frac{1}{4}r(r-1)} \prod_{j=1}^{r} \Gamma\left\{\frac{1}{2}\left(p - j + 1\right)\right\}}, \tag{23}$$

where $\Gamma\left(\cdot\right)$ is the Gamma function (Abramowitz and Stegun 1972). We choose the priors on $A_r$ and $X_r$ to be the least informative, i.e. uniform on $\mathcal{S}_{p,r}$ and $\mathcal{S}_{n,r}$ respectively (Jeffreys 1961):

$$f\left(A_r\right) = C\left(p, r\right)^{-1} \chi\left(\mathcal{S}_{p,r}\right), \tag{24}$$

$$f\left(X_r\right) = C\left(n, r\right)^{-1} \chi\left(\mathcal{S}_{n,r}\right). \tag{25}$$

There is no upper bound on $\omega > 0$ (10). An appropriate prior is therefore (the improper) Jeffreys' prior on scale parameters (Jeffreys 1961):

$$f\left(\omega\right) \propto \omega^{-1}. \tag{26}$$

## 4.1 Prior on $\boldsymbol{l}_r$

We assume, that the sum of squares of elements of $D$ is normalized:

$$\sum_{i=n}^{p} \sum_{j=1}^{n} d_{i,j}^2 = \mathrm{tr}\left(DD'\right) = 1. \tag{27}$$

This can easily be achieved in a pre-processing step. (27) can be expressed, using (15), as:

$$\mathrm{tr}\left(DD'\right) = \mathrm{tr}\left(U_D L_D L_D U_D'\right) = \sum_{i=1}^{p} l_{D,i}^2 = 1.$$

This implies the following constraint on $\boldsymbol{l}_r$:

$$\sum_{i=1}^{r} l_i^2 \leq \sum_{i=1}^{p} l_{D,i}^2 = 1. \tag{28}$$

This, together with (12), confines $\boldsymbol{l}_r$ to the space

$$\mathcal{L}_r = \left\{ \boldsymbol{l}_r \middle| l_1 > l_2 > \ldots > l_r > 0, \ \sum_{i=1}^{r} l_i^2 \leq 1 \right\}, \tag{29}$$

which is a segment of the unit hyperball, $\mathcal{H}_r$, of volume

$$h_r = \pi^{\frac{r}{2}} / \Gamma\left(\frac{r}{2}+1\right). \tag{30}$$

The positivity constraints in (29) restrict the volume of $\mathcal{L}_r$ to $h_r/2^r$, while hyperplanes $\{l_i = l_j, \ \forall i,j = 1 \ldots r\}$ partition this positive segment into $r!$ sub-segments, each of equal volume, and only one of which satisfies (12). Hence, the volume of the support (29) is

$$\mathcal{V}_r = h_r \frac{1}{2^r\,(r!)} = \frac{\pi^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}+1\right) 2^r\,(r!)}.$$

Therefore, we choose the prior distribution on $\boldsymbol{l}_r$ to be non-committal—i.e. uniform—on support (29):

$$f(\boldsymbol{l}_r) = \mathcal{U}(\mathcal{L}_r) = \mathcal{V}_r^{-1} \chi(\mathcal{L}_r). \tag{31}$$

### 4.2 The joint distribution

Multiplying (13) by (24), (25), (26), and (31), and using the chain rule of probability, we obtain the joint distribution:

$$\begin{aligned} f(D, A_r, L_r, X_r, \omega | r) \ &\propto \ \mathcal{N}\left(A_r L_r X_r, \ \omega^{-1} I_p \otimes I_n\right) \times \\ &\quad \omega^{-1} \mathcal{V}_r^{-1} C(p,r)^{-1} C(n,r)^{-1}, \end{aligned} \tag{32}$$

on support $\{A_r \in \mathcal{S}_{p,r}\} \times \{\boldsymbol{l}_r \in \mathcal{L}_r\} \times \{X_r \in \mathcal{S}_{n,r}\} \times \{\omega > 0\}$.

Exact posterior inference from (32) is not available. Since the Variational Bayes (VB) approximation method was successfully used for the factor analysis model (18) (Bishop 1999), it will now be invoked for the orthogonally restricted model (32).

## 5 Orthogonal Variational PCA (OVPCA)

### 5.1 Review of the Variational Bayes (VB) Framework

An intractable posterior pdf, $f(\theta|D)$, can be approximated via a tractable pdf, $\tilde{f}(\theta|D)$, which minimizes the Kullback-Leibler distance (KLD) between the latter and the former. The KLD is defined as

$$KL\left(\tilde{f}(\theta|D) \,||\, f(\theta|D)\right) \ = \ -\int_{\Theta} \tilde{f}(\theta|D) \ln \frac{f(\theta|D)}{\tilde{f}(\theta|D)} d\theta \geq 0, \tag{33}$$

where $\Theta$ denotes the support of $\theta$. Equality is reached in (33) iff $\tilde{f}(\theta|D) = f(\theta|D)$ (Kullback and Leibler 1951).

Given a multivariate parameterization, $\theta = [\alpha', \beta']'$, we can functionally constrain $\tilde{f}(\theta|D)$ to distributions exhibiting posterior independence between $\alpha$ and $\beta$. Then, the minimum of the KLD is found via functional optimization (Bishop 1999; Miskin 2000; Ghahramani and Beal 2000) using the Variational Bayes (VB) procedure. This is expressed by the following theorem.

**Theorem 1 (Variational Bayes (VB))** *Let $f(\alpha,\beta|D)$ be a joint posterior pdf of random variables $\alpha,\beta$, given $D$. Let $\tilde{f}(\alpha,\beta|D)$ be an approximate pdf restricted to the set of conditionally independent distributions on $\alpha,\beta$:*

$$\tilde{f}(\alpha,\beta|D) = f_\alpha(\alpha|D)\, f_\beta(\beta|D). \tag{34}$$

*Then, the minimum of the KL distance, $KL\left(\tilde{f}(\alpha,\beta|D)\,||f(\alpha,\beta|D)\right)$, is reached for*

$$f_\alpha(\alpha|D) \propto \exp\left(\mathsf{E}_\beta\left[\ln\left(f(\alpha,\beta,D)\right)\right]\right), \tag{35}$$
$$f_\beta(\beta|D) \propto \exp\left(\mathsf{E}_\alpha\left[\ln\left(f(\alpha,\beta,D)\right)\right]\right), \tag{36}$$

*where $\mathsf{E}_\alpha[\cdot]$ and $\mathsf{E}_\beta[\cdot]$ denote expectation with respect to $f_\alpha(\alpha|D)$ and $f_\beta(\beta|D)$ respectively.*

*Proof:* A simple exercise in probability calculus, using non-negativity of the KLD. ■

## 5.2 Orthogonal Variational PCA (OVPCA)

**Corollary 1** *Theorem 1, applied to model (32), yields the following approximate posterior distributions:*

$$f(A_r|D,r) = \mathcal{M}(F_A), \tag{37}$$
$$f(X_r|D,r) = \mathcal{M}(F_X), \tag{38}$$
$$f(\boldsymbol{l}_r|D,r) = t\mathcal{N}\left(\boldsymbol{m}, s^2 I_r; \tilde{\mathcal{L}}_r\right), \tag{39}$$
$$f(\omega|D,r) = \Gamma(a,b). \tag{40}$$

*Here, $\mathcal{M}(\cdot)$ denotes the von-Mises-Fisher matrix distribution (i.e. the matrix normal distribution restricted to the Stiefel manifold (Khatri and Mardia 1977): see Appendix A). Their matrix parameters are $F_A \in \Re^{p\times r}$ in (37), and $F_X \in \Re^{n\times r}$ in (38). $t\mathcal{N}(\cdot)$ is the truncated Normal distribution with support formally given by (29). The simplified support, $\tilde{\mathcal{L}}_r$, adopted in (39) is explained in Remark 5, to follow.*

*The data-dependent (and rank-dependent) parameters of (37)–(40) are:*

$$F_A = \widehat{\omega} D \widehat{X_r} \widehat{L_r}, \tag{41}$$
$$F_X = \widehat{\omega} D' \widehat{A_r} \widehat{L_r}, \tag{42}$$
$$\boldsymbol{m} = \mathrm{diag}\left(\widehat{X_r}' D' \widehat{A_r}\right), \tag{43}$$
$$s^2 = \widehat{\omega}^{-1}, \tag{44}$$
$$a = \frac{pn}{2}, \tag{45}$$
$$b = \frac{1}{2}\mathrm{tr}\left(DD' - 2D\widehat{X_r}\widehat{L_r}\widehat{A_r}'\right) + \frac{1}{2}\widehat{\boldsymbol{l}_r'\boldsymbol{l}_r}. \tag{46}$$

*These, therefore, are defined in terms of moments of distributions (37)–(40), namely $\widehat{A_r}$, $\widehat{X_r}$, $\widehat{\boldsymbol{l}_r}$, $\widehat{\boldsymbol{l}_r'\boldsymbol{l}_r}$, and $\widehat{\omega}$. The SVD of parameters $F_A$ (41) and $F_X$ (42) are*

$$F_A = U_{F_A} L_{F_A} V_{F_A}', \tag{47}$$
$$F_X = U_{F_X} L_{F_X} V_{F_X}', \tag{48}$$

*with $L_{F_X}$ and $L_{F_A}$ both in $\Re^{r\times r}$. Then:*

$$\widehat{A_r} = U_{F_A} G(p, L_{F_A}) V_{F_A}', \tag{49}$$
$$\widehat{X_r} = U_{F_X} G(n, L_{F_X}) V_{F_X}', \tag{50}$$
$$\widehat{\boldsymbol{l}_r} = \boldsymbol{m} + s\,\zeta(\boldsymbol{m}, s), \tag{51}$$
$$\widehat{\boldsymbol{l}_r'\boldsymbol{l}_r} = rs^2 + \boldsymbol{m}'\widehat{\boldsymbol{l}_r} - s\rho(\boldsymbol{m}, s), \tag{52}$$
$$\widehat{\omega} = \frac{a}{b}. \tag{53}$$

*Moments of $\mathcal{M}(\cdot)$ and $t\mathcal{N}(\cdot)$—from which (49)–(52) are derived—are reviewed in Appendices A and C respectively. Functions $G(\cdot,\cdot)$, $\zeta(\cdot,\cdot)$, and $\rho(\cdot,\cdot)$ are also defined there.*

*Proof:* An easy but lengthy exercise in probability calculus. ■

**Remark 5** *The exact distribution for $\boldsymbol{l}_r$ in the VB procedure is*

$$f\left(\boldsymbol{l}_r|D,r\right)=t\mathcal{N}\left(\boldsymbol{m},s^2I_r;\mathcal{L}_r\right), \tag{54}$$

*i.e. the Normal distribution confined to support $\mathcal{L}_r$ (29). However, the moments of (54) are difficult to evaluate. Hence, we approximate the support $\mathcal{L}_r$ by its envelope $\overline{\mathcal{L}}_r \approx \mathcal{L}_r$, as follows, $l_r$ is maximized if $l_1 = l_2 = \ldots = l_r$, $l_{r+1} = l_{r+2} = \ldots = l_p = 0$. In this case, $\sum_{i=1}^{p} l_i^2 = rl_r^2 < 1$ which defines an upper bound $l_r < \bar{\bar{l}}_r$ to be $\bar{\bar{l}}_r = r^{-\frac{1}{2}}$. Thus, (29) has a rectangular envelope:*

$$\overline{\mathcal{L}}_r=\left\{\boldsymbol{l}_r|0<l_i\leq\bar{\bar{l}}_i=i^{-\frac{1}{2}},\quad i=1,\ldots,r\right\}. \tag{55}$$

*(54) is then approximated by (39). The error of approximation is largest at the boundaries $l_i = l_j,\ i \neq j,\ i,j \in \{1 \ldots r\}$, and is negligible when no two $l_i$'s are equal.*

Note that equations (41–53) form a set of implicit equations for which a closed form solution is not available. Implicit equations in the VB context (Bishop 1999; Ghahramani and Beal 2000; Miskin 2000) are usually solved iteratively via an algorithm of the Expectation Maximization (EM) kind (Dempster et al. 1977; Rubin and Thayer 1982). Properties of the variational EM algorithm are therefore similar to those of EM (Beal and Ghahramani 2003). Closer analysis of equations (41)–(53) reveals that the variational EM algorithm for our model can be simplified, as follows.

**Proposition 1 (Orthogonal Variational PCA (OVPCA))** *Consider the special case where $\widehat{A_r}$ (49) and $\widehat{X_r}$ (50) are formed from scaled singular vectors of the data matrix, D (15):*

$$\widehat{A_r} = U_{D;r}K_A, \tag{56}$$
$$\widehat{X_r} = V_{D;r}K_X. \tag{57}$$

*and $K_A = \mathrm{diag}\left(\boldsymbol{k}_A\right) \in \Re^{r\times r}$, $K_X = \mathrm{diag}\left(\boldsymbol{k}_X\right) \in \Re^{r\times r}$ denote constants of proportionality which must be determined. Then, each iteration using equations (41)–(53) satisfies (56) and (57).*

*Proof:* Consider the $t$th iteration step, $t = 1, 2, \ldots$, where superscript $^{(t)}$ denotes the optimized parameter values in this step. Assume that estimates, $\widehat{A_r}^{(t-1)}$, $\widehat{X_r}^{(t-1)}$, at the end of the previous step[1], are of the form (56), (57); i.e.

$$\widehat{A}_r^{(t-1)} = U_{D;r}K_A^{(t-1)},\quad \widehat{X_r}^{(t-1)} = V_{D;r}K_X^{(t-1)}.$$

Hence, the von-Mises-Fisher parameters, $F_A$ and $F_X$, are updated, at iteration $t$, via (41) and (42) respectively, and using (15):

$$F_A^{(t)} = \widehat{\omega}^{(t-1)}\left(U_DL_DV_D'\right)V_{D;r}K_X^{(t-1)}\widehat{L_r}^{(t-1)} = \widehat{\omega}^{(t-1)}U_{D;r}L_{D;r,r}K_X^{(t-1)}\widehat{L_r}^{(t-1)}, \tag{58}$$
$$F_X^{(t)} = \widehat{\omega}^{(t-1)}\left(U_DL_DV_D'\right)'U_{D;r}K_A^{(t-1)}\widehat{L}_r^{(t-1)} = \widehat{\omega}^{(t-1)}V_{D;r}L_{D;r,r}K_A^{(t-1)}\widehat{L}_r^{(t-1)}. \tag{59}$$

These are in the SVD form of $F_A$ (47), and $F_X$ (48) respectively, with assignments:

$$U_{F_A}^{(t)} = U_{D;r},\quad L_{F_A}^{(t)} = \widehat{\omega}^{(t-1)}L_{D;r,r}K_X^{(t-1)}\widehat{L}_r^{(t-1)},\quad V_{F_A}^{(t)} = I_r, \tag{60}$$
$$U_{F_X}^{(t)} = V_{D;r},\quad L_{F_X}^{(t)} = \widehat{\omega}^{(t-1)}L_{D;r,r}K_A^{(t-1)}\widehat{L}_r^{(t-1)},\quad V_{F_X}^{(t)} = I_r. \tag{61}$$

Substituting (60) and (61) into (49) and (50) respectively:

$$\widehat{A}_r^{(t)} = U_{D;r}G\left(p,\widehat{\omega}^{(t-1)}L_{D;r,r}K_X^{(t-1)}\widehat{L}_r^{(t-1)}\right)I_r = U_{D;r}K_A^{(t)}, \tag{62}$$
$$\widehat{X}_r^{(t)} = V_{D;r}G\left(n,\widehat{\omega}^{(t-1)}L_{D;r,r}K_A^{(t-1)}\widehat{L}_r^{(t-1)}\right)I_r = V_{D;r}K_X^{(t)},$$

[1]Initial conditions, i.e. at $t = 0$, will be specified shortly.

since the function $G(\cdot,\cdot)$ with diagonal matrix argument returns also a diagonal matrix (A.9). Therefore, updated estimates, $\widehat{A}_r^{(t)}$ and $\widehat{X}_r^{(t)}$, remain of the same type as (56), (57) with assignments:

$$K_A^{(t)} = G\left(p, \widehat{\omega}^{(t-1)} L_{D;r,r} K_X^{(t-1)} \widehat{L}_r^{(t-1)}\right), \tag{63}$$

$$K_X^{(t)} = G\left(n, \widehat{\omega}^{(t-1)} L_{D;r,r} K_A^{(t-1)} \widehat{L}_r^{(t-1)}\right). \tag{64}$$

■

Note that, under Proposition 1, the optimal values of $\widehat{A_r}$ and $\widehat{X_r}$ are determined up to the constants of proportionality, $\boldsymbol{k}_A$ and $\boldsymbol{k}_X$. The iterative algorithm is then greatly simplified, since we need only iterate on the $2r$ degrees of freedom constituting $K_A$ and $K_X$ together, and not on $\widehat{A_r}$ and $\widehat{X_r}$ with $r\left(p+n-\frac{r-1}{2}\right)$ degrees of freedom. To achieve this, we must, however, satisfy the requirement of Proposition 1, namely we must initialize the iterative scheme to satisfy (56) and (57), using any diagonal matrices $K_A$ and $K_X$ with positive[2] diagonal elements. In fact, for $K_A^{(0)} = K_X^{(0)} = I_r$, (56) and (57) are the ML solutions (14), and so an ML-initialized iteration is proposed, leading finally to the Orthogonal Variational PCA (OVPCA) algorithm, which follows. Note that:

- initialization via the ML solution guarantees fast convergence to the unique solution, since (32) is likelihood-dominated by design.
- (41)–(53) now involve products of diagonal matrices. Hence, we need only evaluate diagonal elements, using identities of the kind

  $$\operatorname{diag}(K_A K_X) = \boldsymbol{k}_A \circ \boldsymbol{k}_X,$$

  where $\circ$ denotes Hadamard product. Equations (43), (46), (63), and (64) can now be reformulated in efficient diagonal form.

The OVPCA algorithm is as follows.

**Algorithm 1 (OVPCA)**

1. *Initialize estimates using ML solution (14), i.e.* $\boldsymbol{k}_A^{(0)} = \boldsymbol{k}_X^{(0)} = \mathbf{1}_{r,1}$, $\widehat{\boldsymbol{l}_r}^{(0)} = \boldsymbol{l}_{D;r}$, $\widehat{\omega}^{(0)} = \frac{pn}{\sum_{i=r+1}^{p} l_{D,i}^2}$.
2. *Evaluate until convergence is reached:*

$$\boldsymbol{k}_A^{(t)} = G\left(p, \widehat{\omega}^{(t-1)} \boldsymbol{l}_{D;r} \circ \boldsymbol{k}_X^{(t-1)} \circ \widehat{\boldsymbol{l}_r}^{(t-1)}\right), \tag{65}$$

$$\boldsymbol{k}_X^{(t)} = G\left(n, \widehat{\omega}^{(t-1)} \boldsymbol{l}_{D;r} \circ \boldsymbol{k}_A^{(t-1)} \circ \widehat{\boldsymbol{l}_r}^{(t-1)}\right), \tag{66}$$

$$\boldsymbol{m}^{(t)} = \boldsymbol{k}_X^{(t-1)} \circ \boldsymbol{l}_{D;r} \circ \boldsymbol{k}_A^{(t-1)}, \tag{67}$$

$$s^{(t)} = \left(\widehat{\omega}^{(t-1)}\right)^{-\frac{1}{2}}, \tag{68}$$

$$\widehat{\boldsymbol{l}_r}^{(t)} = \boldsymbol{m}^{(t-1)} + s^{(t-1)} \zeta\left(\boldsymbol{m}^{(t-1)}, s^{(t-1)}\right), \tag{69}$$

$$\widehat{\boldsymbol{l}_r' \boldsymbol{l}_r}^{(t)} = \left(\boldsymbol{m}^{(t-1)}\right)' \widehat{\boldsymbol{l}_r}^{(t-1)} + r\left(s^{(t-1)}\right)^2 - s^{(t-1)} \rho\left(\boldsymbol{m}^{(t-1)}, s^{(t-1)}\right)' \mathbf{1}_{1,r}, \tag{70}$$

$$\widehat{\omega}^{(t)} = pn\left[\boldsymbol{l}_D' \boldsymbol{l}_D - 2\left(\boldsymbol{k}_X^{(t-1)} \circ \widehat{\boldsymbol{l}_r}^{(t-1)} \circ \boldsymbol{k}_A^{(t-1)}\right)' \boldsymbol{l}_{D;r} + \widehat{\boldsymbol{l}_r' \boldsymbol{l}_r}^{(t-1)}\right]^{-1}. \tag{71}$$

*Note that, $G(\cdot,\cdot)$, $\zeta(\boldsymbol{m}, s)$, and $\rho(\boldsymbol{m}, s)$ are functions returning vectors of the same length as their vector arguments, as defined in Appendixes A and C, respectively.*

[2]Non-negativity is required by $G(\cdot,\cdot)$. Initialization with zeros will be discussed shortly.

**Remark 6 (Automatic Relevance Determination (ARD):)** *it is observed that estimates of $k_{A,i}$ (65) and $k_{X,i}$ (66) typically converge to zero for $i > r_u$, for some empirical upper bound, $r_u$. A similar property was used as a rank selection criterion in previously published Bayesian approaches. In those approaches, the model order was chosen as $\hat{r} = r_u$ (Bishop 1999).*

**Remark 7** *Equations (65)–(67) are (trivially) satisfied for*

$$\boldsymbol{k}_A = \boldsymbol{k}_X = \boldsymbol{m} = \mathbf{0}_{r,1}, \tag{72}$$

*independently of data, giving $\widehat{M_{(r)}} = \mathbf{0}_{p,n}$ (11). The only parameter then to be determined is $\omega$. Solution (72) is appropriate for data formed only by realizations of isotropic Gaussian noise without any signal (8), i.e. $r = 0$. This case will then be revealed by the ARD Property (Remark 6), i.e. $r_u$ will be found to equal zero. If the ARD Property yields a different estimate, i.e. $r_u \geq 1$, then (72) constitutes a local maximum (or singular point) of the VB approximation (37)–(40), of (32). Then, the global minimum of the KLD has to be found by evaluation of KLD for both cases[3], namely, the ML initialized case (Algorithm 1) and the zero-centred case (72). Further comment on evaluation of the KLD for this model (32) follows in Section 6.1.*

**Conjecture 1** *Apart from the zero-centred solution (72), there is only one solution of equations (41)–(53) for each of the $2^r$ cases of decomposition (11) (arising from those in (7)).*

**Remark 8** *If Conjecture 1 is true, then, under any random initialization, the iterative equations (41)–(53) also converge to the solution found under Proposition 1 (56)–(57).*

**Remark 9** *The columns of matrix parameter $A_r$ were named parametric components (Remark 1). Hence, their posterior expected value, $\widehat{A}_r$ (56) will be known as parametric principal components of $D$, in analogy to the ML definition (14) of the classical principal components (7). These two types of PCs' are collinear (56) under Proposition 1 and synonymous when $\boldsymbol{k}_A = \mathbf{1}_{r,1}$, which is the case for $\widehat{\omega} \to \infty$.*

Proposition 1 reveals the following interesting analytical insight:

- We noted $2^r$ cases of SVD decomposition (11), each determined by the signs of the singular vectors. Note, however, that Proposition 1 separates the posterior mean values, $\widehat{A_r}$ (56) and $\widehat{X_r}$ (57), into an orthogonal and proportionality part. Only the latter ($\boldsymbol{k}_A$ and $\boldsymbol{k}_X$ respectively) are estimated using the OVPCA algorithm (Algorithm 1). Since the elements of vector function $G(\cdot,\cdot)$ are confined to the interval $[0,1]$ (see Appendix A.2, Figure 3), estimated values of $\boldsymbol{k}_A$ (65) and $\boldsymbol{k}_X$ (66) are always positive. In other words, the VB solution is approximating only *one* of the possible $2^r$ modes. Symmetry ensures that the OVPCA solution is valid for any of them. This is important, as the multimodal distribution of $A_r$ is symmetric around the coordinate origin, which would consign the posterior mean to being $\widehat{A_r} = \mathbf{0}_{p,r}$. Note that this symmetry is also reflected in the VB equations (65)–(71) (Remark7).
- As a consequence of the ARD Property (Remark 6), the number of possible modes of the VB approximate posterior distribution is reduced to $2^{r_u}$.

# 6 Answering the Open Problems

The OVPCA solution presented above allows us to address the two 'open problems' related to the ML solution, listed in Section 3.1.

[3]Experiments suggests that if $r_u \geq 1$ its KLD distance is smaller than the one corresponding to (72).

### 6.1 Inference of Rank

In the foregoing, we assumed that the rank, $r$, of the model (11) was known *a priori*. If this is not the case, then inference of this parameter can be made using Bayes' rule:

$$f\left(r|D\right) \propto f\left(D|r\right) f\left(r\right), \tag{73}$$

where $f\left(r\right)$ denotes the prior on $r$, typically uniform on $1 \leq r \leq p$. The marginal data posterior, $f\left(D|r\right)$, can be approximated by a lower bound:

$$\ln f\left(D|r\right) \geq \ln f\left(D|r\right) - KL\left(\tilde{f}\left(\theta|D,r\right)||f\left(\theta|D,r\right)\right), \tag{74}$$

where $KL\left(\cdot\right) \geq 0$ can be minimized by the variational solution (Theorem 1). Hence, (74) is the greatest lower bound consistent with the conditionally independent approximation (34). Hence, we adopt the approximation

$$\begin{aligned} \ln f\left(D|r\right) &\approx \ln f\left(D|r\right) - KL\left(\tilde{f}\left(\theta|D,r\right)||f\left(\theta|D,r\right)\right) \\ &= \int_{\Theta} \tilde{f}\left(\theta|D,r\right)\left(\ln f\left(D,\theta|r\right) - \ln\left(\tilde{f}\left(\theta|D,r\right)\right)\right) d\theta. \end{aligned} \tag{75}$$

The parameters are $\theta = [A_r, L_r, X_r, \omega]$, and $f\left(D,\theta|r\right)$ is given by (32). The optimal approximation, $\tilde{f}\left(\theta|D,r\right)$, is the conditionally independent model, yielded by the VB framework (37)–(40):

$$\tilde{f}\left(A_r, L_r, X_r, \omega|D\right) = f\left(A_r|D,r\right) f\left(L_r|D,r\right) f\left(X_r|D,r\right) f\left(\omega|D,r\right) \tag{76}$$

Substituting (37)–(40) into (76), and (32) into (75), then (73) yields:

$$\begin{aligned} f\left(r|D\right) \propto \exp\Bigg\{ & -\frac{r}{2}\ln\pi + r\ln 2 + \ln\Gamma\left(\frac{r}{2}+1\right) + \ln\left(r!\right) \\ & + \frac{1}{2}s^{-2}\left(\boldsymbol{m}'\boldsymbol{m} - \widehat{\boldsymbol{l}_r}'\boldsymbol{m} - \boldsymbol{m}'\widehat{\boldsymbol{l}_r} + \widehat{\boldsymbol{l}_r'\boldsymbol{l}_r}\right) \\ & + \ln\,{}_0F_1\left(\frac{1}{2}p, \frac{1}{4}F_A F_A'\right) - \widehat{\omega}\left(\boldsymbol{k}_X \circ \widehat{\boldsymbol{l}_r} \circ \boldsymbol{k}_A\right)' \boldsymbol{l}_{D;r} \\ & + \ln\,{}_0F_1\left(\frac{1}{2}n, \frac{1}{4}F_X F_X'\right) - \widehat{\omega}\left(\boldsymbol{k}_X \circ \widehat{\boldsymbol{l}_r} \circ \boldsymbol{k}_A\right)' \boldsymbol{l}_{D;r} \\ & + \sum_{j=1}^{r} \ln\left[\operatorname{erf}\left(\left(s\sqrt{2}\right)^{-1}\left(\overline{\overline{l_j}} - m_j\right)\right) + \operatorname{erf}\left(\left(s\sqrt{2}\right)^{-1} m_j\right)\right] \\ & + r\ln\left(s\sqrt{\pi/2}\right) - (a+1)\ln b \Bigg\}. \end{aligned} \tag{77}$$

$\boldsymbol{k}_A$, $\boldsymbol{k}_X$, $\boldsymbol{m}$, $\widehat{\boldsymbol{l}_r}$, $\widehat{\boldsymbol{l}_r'\boldsymbol{l}_r}$, $s$ and $\widehat{\omega}$ are the converged solutions of the OVPCA algorithm (Algorithm 1), and $F_A$ (41) and $F_X$ (42) are functions of these. $\overline{\overline{l_j}}$ is given by (55).

We note the following:

- One of the main algorithmic advantages of PCA is that a single evaluation of all $p$ eigenvectors, i.e. $U = U_D$ (5), (15), (16) provides with ease the PCA solution for any rank $r < p$, via the simple extraction of the first $r$ columns, $U_{D;r}$ (7). The OVPCA algorithm also enjoys this property, thanks to the linear dependence of solution (56) on $U_D$. Furthermore, $V_D$ observes the same property (15). Therefore, in the OVPCA procedure, the optimal solution for a given rank is obtained by simple extraction of $U_{D;r}$ and $V_{D;r}$, followed by iterations involving only scaling coefficients, $\boldsymbol{k}_A$ and $\boldsymbol{k}_X$, for that rank. Hence, $p \times (p+n)$ values (those of $U_D$ and $V_D$) are determined rank-independently via the ML solution, and only $4r+3$ values (those of vectors $\boldsymbol{k}_A$, $\boldsymbol{k}_X$, $\boldsymbol{m}$, $\widehat{\boldsymbol{l}_r}$, and scalars $s$, $\widehat{\boldsymbol{l}_r'\boldsymbol{l}_r}$ and $\widehat{\omega}$ together) are involved in the rank-dependent iterations (65)–(71).

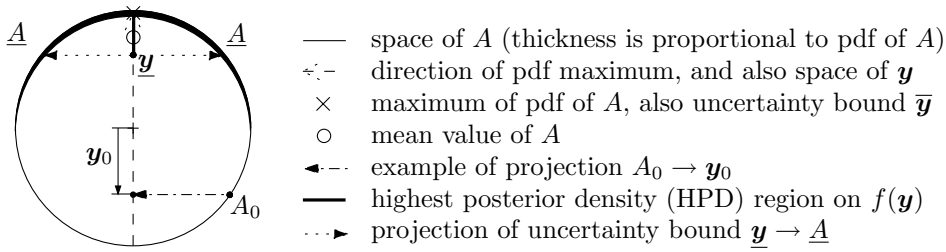



Figure 1: Illustration of properties of von-Mises-Fisher distribution $\mathcal{M}(F)$ of $A$, for $A, F \in \Re^{2\times 1}$.

- As a consequence of the ARD Property (Remark 6), values of all parameters $(A_r, L_r, X_r, \omega)$ inferred by the OVPCA algorithm are almost identical for $r \geq r_u$ (this will be demonstrated in simulation in Section 7). Therefore, it is reasonable to evaluate the OVPCA parameters for $r = p-1$ (we cannot use $r = p$ since then $\hat{\omega}_p$ is not valid (14)), and approximate $\widehat{A}_r \approx \widehat{A}_{p-1}$, $\forall r \geq r_u$, (and similarly for $L_r, X_r, \omega$). This approximation can significantly reduce the number of runs of the OVPCA algorithm required for evaluation of $f(r|D)$ (77).

- The explicit posterior distribution on $r$, i.e. (77), was not provided by previously published approaches (Bishop 1999; Ghahramani and Beal 2000). In its place, the ARD Property of *their* algorithms was used to infer rank. Since the OVPCA algorithm also possesses the ARD Property (Remark 6), it will be used for comparison with the formal Bayesian solution (77) in the simulation studies that follow (Section 7).

### 6.2 Moments of the Model Parameters

First moments of the model parameters—i.e. $\widehat{A}_r$, $\widehat{\boldsymbol{l}}_r$, $\widehat{X}_r$, $\widehat{\omega}$—are outputs of the OVPCA algorithm. The Bayesian framework also permits uncertainty measures on these parameters to be determined, notably, Highest Posterior Density (HPD) regions Bernardo and Smith (1997). HPD regions on the von-Mises-Fisher distribution (37), (38) are not known to us. Therefore, we derive them in Appendix A.3, using a Gaussian approximation. We use the approximate HPD regions to obtain uncertainty bounds on *parametric components*, as follows. From (A.4), (A.5), the pdf of $A_r \in \Re^{n\times r}$ is fully determined by the transformed variable $\boldsymbol{y}_A \in \Re^r$:

$$\boldsymbol{y}_A(A_r) \;=\; \operatorname{diag}\left(U'_{F_A} A_r V_{F_A}\right) = \operatorname{diag}\left(U'_{D;r} A_r\right), \tag{78}$$

using (61). Hence, uncertainty bounds for $A_r$ can be mapped to uncertainty bounds for $\boldsymbol{y}_A$ using (78), as shown in Appendix A.3. The idea is illustrated graphically for $p = 2$ and $r = 1$ in Figure 1. Using this mapping, lower and upper uncertainty bounds on $A_r$ can be defined as follows:

$$\underline{A_r} \;=\; \left\{A_r :\; \boldsymbol{y}_A(A_r) = \underline{\boldsymbol{y}}_A\right\}, \tag{79}$$

$$\overline{A_r} \;=\; \left\{A_r :\; \boldsymbol{y}_A(A_r) = \overline{\boldsymbol{y}}_A\right\}, \tag{80}$$

where $\boldsymbol{y}_A(\cdot)$ is given by (78), and $\underline{\boldsymbol{y}}_A$, $\overline{\boldsymbol{y}}_A$—given in Appendix A.3—are bounds to the approximating Gaussian distribution of $\boldsymbol{y}_A$ (A.15). Uncertainty bounds for $X_r$—again von-Mises-Fisher distributed (38)—are obtained in the same way.

Using Proposition 1, i.e. (56), transformation (78) of the parametric PCs $\widehat{A_r}$ (57) becomes

$$\boldsymbol{y}_A\left(\widehat{A_r}\right) = \operatorname{diag}\left(U'_{D;r} U_{D;r} K_A\right) = \boldsymbol{k}_A. \tag{81}$$

Uncertainty bounds on variables $\boldsymbol{l}_r$ and $\omega$ are relatively easy to obtain. The truncated Normal distribution of $\boldsymbol{l}_r$ (39) can be approximated by Normal distribution, using Maximum Entropy principle, as discussed in Appendix C. Parameter $\omega$ is Gamma distributed (40), and so uncertainty bounds are therefore available.

Table 1: Bayesian inference of singular values, $\boldsymbol{l}_r$, given different rank estimates: Bayesian selection ($r = 3$), and ARD Property ($r = 9$).

| | Bayesian selection $r = \max f(r\|D) = 3$ | | | ARD Property $r = 9$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | $l_1$ | $l_2$ | $l_3$ | $l_1$ | $l_2$ | $l_3$ | $l_4$ | $l_5 \dots l_9$ |
| simulated | 19.48 | 11.70 | 1.66 | 19.48 | 11.70 | 1.66 | 0 | 0 |
| upper bound | 19.50 | 12.00 | 1.87 | 19.50 | 12.00 | 1.87 | 0.21 | 0.21 |
| mean value | 19.29 | 11.79 | 1.66 | 19.28 | 11.79 | 1.66 | 0.08 | 0.08 |
| lower bound | 19.08 | 11.58 | 1.45 | 19.07 | 11.58 | 1.44 | 0 | 0 |

Table 2: Inference of $A_r$ via OVPCA (displayed using transformed variable $\boldsymbol{y}_{A;r}$).

| | Bayesian selection $r = \max f(r\|D) = 3$ | | | ARD Property $r = 9$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | $y_{A,1}$ | $y_{A,2}$ | $y_{A,3}$ | $y_{A,1}$ | $y_{A,2}$ | $y_{A,3}$ | $y_{A,4}$ | $y_{A,5} \cdots y_{A,9}$ |
| simulated | 0.9999 | 0.9995 | 0.9824 | 0.9999 | 0.9995 | 0.9824 | 0.4469 | $< 0.45$ |
| upper bound, $\overline{y_{A;i}}$ | 1.0000 | 1.0000 | 1.0037 | 1.0000 | 1.0000 | 1.0038 | 0.6325 | 0.6325 |
| mean value, $k_{A,1}$ | 0.9999 | 0.9997 | 0.9862 | 0.9999 | 0.9997 | 0.9861 | 0 | 0 |
| lower bound, $\underline{y_{A,i}}$ | 0.9997 | 0.9993 | 0.9688 | 0.9997 | 0.9993 | 0.9684 | $-0.6325$ | $-0.6325$ |

# 7 Illustrative examples

## 7.1 Simulated data study

A set of multidimensional data was simulated using model (13), with $p = 10$, $n = 200$, $r = 3$, and $\omega = 10$. True (i.e. the simulated) singular values are given in Table 1 together with approximating moments of their posterior distribution (39). Simulated values are clearly within the uncertainty bounds. We condition on two different cases of $r$, namely (i) maximum of the posterior distribution of $r$ (77), and (ii) highest possible rank $r = p-1 = 9$, which allows the ARD property (6) to be exploited. Note that values of $f(l_i|D, r = 3)$, and $f(l_i|D, r = 9)$ are almost identical for $i = 1, \dots, 3$, which can be exploited in the rapid evaluation of (77) as discussed in the second bullet point of Section 6.1.

VB approximating posterior distributions for orthogonal parameters $A_r$ (37) and $X_r$ (38) are presented in transformed variables, $\boldsymbol{y}_A$ and $\boldsymbol{y}_X$ (78), respectively. Moments of the transformed distributions are displayed in Table 2 and Table 3, together with the projection of the original simulated values in each case. Again, we condition on the two cases of $r$ used in Table 1. Projected true values of $A_r$ and $X_r$ are, in both cases, within the uncertainty bounds of the posteriors (A.15), $f(\boldsymbol{y}_A|D, r)$ and $f(\boldsymbol{y}_X|D, r)$, respectively.

Results of Bayesian rank selection (77) are displayed in Table 4. The ARD Property of the OVPCA algorithm (Remark 6) is apparent in the mean-value rows of the ARD sub-tables of Tables 2 and 3. In this simple simulation study, both methods—i.e. the Bayesian approach (77), and the ARD Property (Remark 6)—selected the true rank of the data.

## 7.2 Application in medical scintigraphic image analysis

PCA is regularly used as a dimensionality reduction step in the factor analysis of medical image sequences (Buvat et al. 1998). In this study, a scintigraphic dynamic image sequence of the kidneys is considered. It contains $p = 120$ images, each of size $64 \times 64$. These were preprocessed as follows:

Table 3: Inference of $X_r$ via OVPCA(displayed in transformed variable $\boldsymbol{y}_{X;r}$).

| | Bayesian selection $r = \max f(r\|D) = 3$ | | | ARD Property $r = 9$ | | | | |
|---|---|---|---|---|---|---|---|---|
| | $y_{X,1}$ | $y_{X,2}$ | $y_{X,3}$ | $y_{X,1}$ | $y_{X,2}$ | $y_{X,3}$ | $y_{X,4}$ | $y_{X,5} \cdots y_{X,9}$ |
| simulated | 0.9987 | 0.9973 | 0.8582 | 0.9987 | 0.9973 | 0.8582 | 0.0097 | $<$0.1706 |
| upper bound, $\overline{y_{X,i}}$ | 0.9990 | 0.9973 | 0.8923 | 0.9989 | 0.9972 | 0.8911 | 0.2000 | 0.2000 |
| mean value, $k_{X,i}$ | 0.9985 | 0.9961 | 0.8484 | 0.9985 | 0.9961 | 0.8469 | 0 | 0 |
| lower bound, $\underline{y_{X,i}}$ | 0.9981 | 0.9950 | 0.8046 | 0.9981 | 0.9950 | 0.8026 | -0.2000 | -0.2000 |

Table 4: Bayesian posterior distribution of rank $r$ for simulated data.

| | $r=1$ | $r=2$ | $r=3$ | $r=4$ | $r=5,\ldots,9$ |
|---|---|---|---|---|---|
| $f\,(r\|D)$ | 0 | 0 | 0.9821 | 0.0174 | $0\ldots0$ |

Table 5: Comparison of rank selection methods for scintigraphic image data.

| OVPCA $f\,(r\|D)$ | OVPCA ARD Property | Variational PCA (30 trials) rank | relative frequency |
|---|---|---|---|
| $\Pr\,(r=17\|D)=0.0004$ | | | |
| $\Pr\,(r=18\|D)=0.2761$ | $r_u=45$ | $r=25$ | 1/30 |
| $\Pr\,(r=19\|D)=0.7232$ | | $r=26$ | 16/30 |
| $\Pr\,(r=20\|D)=0.0002$ | | $r=27$ | 13/30 |

Note: where not listed, $f\,(r|D)<3\times10^{-7}$

- a rectangular area of $n=525$ pixels was chosen as the region of interest at the same location in each image.
- data were scaled by the correspondence analysis method (Benali et al. 1993). With this scaling, the noise on the scintigraphic data is approximately isotropic Gaussian (Benali et al. 1993), satisfying the model assumptions (10).

It is interesting to compare methods for selection of relevant principal components. The approximate posterior distribution of rank (77) and the ARD Property (Remark 6) infer significantly different optimal rank (Table 5). For comparison, we also inferred the rank of the data via (i) Variational PCA (VPCA) (Bishop 1999), and (ii) the *ad hoc* criterion of cumulative percentage of total variation (Jolliffe 2002) (Figure 2). Method (i) is initialized randomly, potentially yielding different results for each run, and so we performed 30 trials. Relative frequency of inferred rank using the ARD Property of this VPCA technique are included in Table 5. For method (ii), $r=5$ was chosen.

It is difficult to compare performance of the methods since the true dimensionality is not known. From a medical point of view, the number of physiological factors should be 4 or 5. This estimate is supported by the *ad hoc* criterion (Figure 2). From this perspective, the formal methods appear to overestimate significantly the number of relevant principal components (PCs). The reason for this can be understood by reconstructing the data using the number of PCs, $r$, recommended by each method (Table 6). Four consecutive frames of the actual scintigraphic data are displayed in the first row. Though the signal-to-noise ratio is poor, clearly functional variation is visible in the central part of the left kidney and in the upper part of the right kidney, which cannot be accounted for by noise. The same frames of the sequence, reconstructed from $r=5$ PCs (Table 6, second row), fail to capture this functional information. In contrast, the functional information *is* apparent on the sequence reconstructed using the Bayesian estimate—i.e. $r=19$ PCs—and, indeed, on sequences reconstructed using $r>19$ PCs, such as the $r=45$ choice suggested by the ARD Property of OVPCA (Table 6, last row).

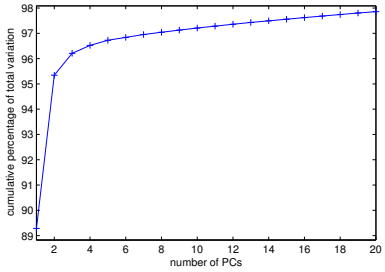
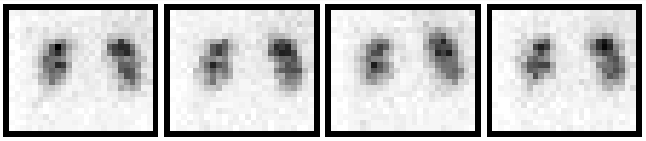
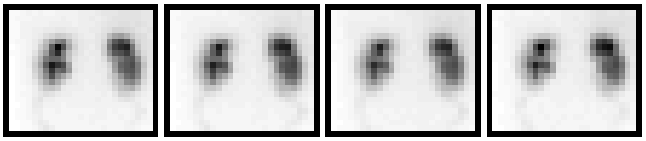
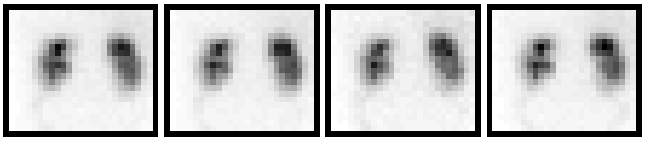
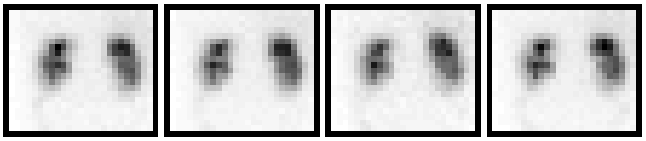



Figure 2: Cumulative percentage of total variation for scintigraphic data. For clarity, only the first 20 elements are shown out of a total of $p=120$.

Table 6: Reconstruction of scintigraphic data for different numbers of PCs

| number of PCs used | frames 48–51 of the dynamic image sequence |
|---|---|
| original images ($r = 120$) |  |
| *Ad hoc* criterion ($r = 5$) |  |
| Bayesian selection ($r = 19$) |  |
| ARD property ($r = 45$) |  |

### 7.3 Discussion of OVPCA and its Performance

The simulated data study in the last section suggests that the proposed posterior rank distribution (77) is capable of correct inference, if the data comply with the model (Table 4). Furthermore, all but one of the true simulated parameters are within the estimated uncertainty bounds (Tables 1–3). Note, however, that some true values are close to an uncertainty bound, suggesting that these inferred bounds are not overly conservative. The scintigraphic data study demonstrates the contrast between "formally modelled noise" and "unwanted structure artefacts". The OVPCA procedure identified signal structure which can not be considered as Gaussian-distributed noise. Notwithstanding this, some parametric PCs identified as signal elements will be considered as unnecessary (i.e. noise) from a medical point-of-view, and *ad hoc* criteria may be preferred. This could be accounted for in the automatic technique by a better model of the noise. We note, however, that OVPCA appears to yield lower estimates of rank than competing formal criteria (Table 5). This feature will be verified by further extensive comparative studies.

The full factor analysis model (Anderson 1971) explicitly includes a common non-zero mean value for the data columns, $\mathsf{E}[d_i] = \boldsymbol{\mu}$. This was not considered as a part of the orthogonal model (13) introduced in this paper. It is easy to introduce a common mean value for applications where it is regarded as important, e.g. PCA mixtures (Tipping and Bishop 1998a). The model (13) can be readily extended to contain the common mean value, as follows:

$$M_{(r)} = A_r L_r X_r + \boldsymbol{\mu} \mathbf{1}_{1,n},$$

with $\boldsymbol{\mu} \in \Re^{p \times 1}$. The Variational Bayes approximation (Theorem 1) for this model requires more algebra, but is straightforward. However, Proposition 1 is not then valid. Hence, the iterative algorithm arising in place of OVPCA (Algorithm 1) has the following disadvantages: (i) computational complexity is much higher; and (ii) posterior means are not collinear with the principal components of PCA.

We note the following features of the OVPCA algorithm:

- Convergence is very fast in comparison to other iterative methods (Bishop 1999). The stopping rule on increments of parameter estimates (e.g. $\widehat{\omega}$) can be set close to the machine precision.
- The method does not involve any numerically sensitive operations like inversions. Hence, no regularization terms are required.

The OVPCA model includes the assumption of isotropic Gaussian noise which might be relaxed in further work. The extension of the model (13) for non-isotropic Gaussian noise is straightforward. So is the application of the Variational Bayes estimation method in this case. However, the resulting posterior distributions are of the generalized Bingham type (Khatri and Mardia 1977), whose moments are not known to us. This suggests that efficient numerical evaluation—possible for OVPCA—cannot be achieved in the non-isotropic noise case. Furthermore, it implies that the parametric PCs of such an extended model are not collinear with PCs extracted by PCA (7). This latter point may explain the unavailability in the literature of non-asymptotic sampling distributions of principal components for non-Gaussian noise distributions (Jolliffe 2002).

# 8 Conclusion

A complete Bayesian framework for PCA has been proposed in this paper. The key step was to force orthogonality in the governing model (13), and in this way to ensure a decomposition with only a countable ($2^r$) sign-based ambiguity, and in turn, only a countable number of modes in the posterior distribution. The proposed Variational Bayes (VB) procedure was shown to approximate each mode of the posterior distribution, and yielded a von-Mises-Fisher distribution on each of the orthogonal parameters, $A_r$ and $X_r$. Using ideas from orthogonal statistics, we proved a proposition which established the mean of the posterior distribution of $A_r$ (called the parametric principal components) as collinear with the principal components (PCs) produced by PCA. Because of this collinearity, the resulting iterative algorithm—known as OVPCA—can be initialized easily by the PCs. The OVPCA algorithm therefore converges quickly and robustly to the VB approximation of the posterior distribution. Furthermore, a novel approximation of the associated hypergeometric function of matrix argument was developed to ensure numerical efficiency.

Among the key results are

- an approximate posterior distribution of rank $r$, corresponding to the number of relevant PCs in the data in classical PCA.
- uncertainty bounds on parametric components, $A_r$, can be interpreted as uncertainty bounds on PCs.

Under the assumption of isotropic Gaussian noise (10), the Bayesian procedure presented in this paper yields parametric PCs proportional to the classical PCs produced by PCA. This points to the optimality of PCA under these conditions. Currently, the fast identification of marginals via OVPCA also depends on the isotropic Gaussian noise assumption. Clearly, however, the VB framework offers a powerful tool for approximation of posterior distributions of parametric principal components for non-isotropic—and even some non-Gaussian (exponential family)—noise classes, without the need for asymptotic assumptions. Further work in orthogonal statistics and approximation will be necessary to achieve efficient algorithms in these cases.

# Appendix A: The von-Mises-Fisher Matrix Distribution

Moments of the von-Mises-Fisher matrix distribution are now considered. Proofs of all unproven results are available in (Khatri and Mardia 1977).

## A.1 Definition

The von-Mises-Fisher pdf of matrix random variable, $Z \in \Re^{p\times r}$, restricted to $Z'Z = I_r$, is given by:

$$f\left(Z|F\right) = \mathcal{M}\left(F\right) = \frac{1}{\kappa\left(p, F'F\right)} \exp\left(\mathrm{tr}\left(F'Z\right)\right), \tag{A.1}$$

$$\kappa\left(p, FF'\right) = {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}F'F\right) C\left(p, r\right), \tag{A.2}$$

where $F \in \Re^{p\times r}$ is a matrix parameter of the same dimensions as $Z$, and $\kappa(p, F'F)$ is the normalizing coefficient. ${}_0F_1(\cdot)$ denotes a hypergeometric function of matrix argument $F'F$ (James 1964), which is treated separately in Appendix B. $C(p,r)$ denotes the area of the relevant Stiefel manifold $\mathcal{S}_{p,r}$ (23). Without loss of generality it is assumed that $r \le p$.

(A.1) is a Gaussian distribution with restriction $Z'Z = I_r$, re-normalized on $\mathcal{S}_{p,r}$. It is governed by a single matrix parameter $F$. Consider the full[4] SVD decomposition,

$$F = U_F L_F V_F',$$

of the parameter $F$, where $U_F \in \Re^{p\times p}$, $L_F \in \Re^{p\times r}$, $V_F \in \Re^{r\times r}$. Then, the maximum of (A.1) is reached at

$$\hat{Z} = U_{F;r} V_F'. \tag{A.3}$$

Flatness of the distribution is controlled by $L_F$. When $\boldsymbol{l}_F = \mathbf{0}_{r,1}$, the distribution is uniform on $\mathcal{S}_{p,r}$ (Mardia and Jupp 2000). For $l_{F,i} \to \infty$, $\forall i = 1 \dots n$, the distribution converges to a Dirac delta function at $\hat{Z}$ (A.3).

## A.2 First Moment

Define the transformed variable

$$Y = U_F' Z V_F. \tag{A.4}$$

It can be shown that $\kappa(p, F'F) = \kappa\left(p, L_F^2\right)$. The pdf of $Y$ is then

$$f(Y|F) \;=\; \frac{1}{\kappa\left(p, L_F^2\right)} \exp\left(\mathrm{tr}\left(L_F Y\right)\right) = \frac{1}{\kappa\left(p, L_F^2\right)} \exp\left(\boldsymbol{l}_F' \boldsymbol{y}\right), \tag{A.5}$$

where $\boldsymbol{y} = \mathrm{diag}(Y)$. Hence

$$f(Y|F) \propto f(\boldsymbol{y}|\boldsymbol{l}_F). \tag{A.6}$$

The first moment of (A.5) is given by

$$\mathsf{E}[Y|L_F] \;=\; \Psi, \tag{A.7}$$

where $\Psi = \mathrm{diag}(\boldsymbol{\psi})$ is a diagonal matrix whose diagonal elements are:

$$\psi_i = \frac{\partial}{\partial l_{F,i}} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}L_F^2\right), \quad i = 1, \dots, r. \tag{A.8}$$

We will denote $\psi$ and $\Psi$ by

$$\boldsymbol{\psi} \;=\; G(p, \boldsymbol{l}_F), \tag{A.9}$$

$$\Psi \;=\; G(p, L_F), \tag{A.10}$$

where $G(p, L_F) = \mathrm{diag}(G(p, \boldsymbol{l}_F))$. The mean value of the original random variable, $Z$, is then (Downs 1972)

$$\mathsf{E}[Z] = U_{F;r} \Psi V_{F;r}' = U_{F;r} G(p, L_F) V_{F;r}'. \tag{A.11}$$

## A.3 Second Moment and Uncertainty Bounds

The second central moment of the transformed variable $\boldsymbol{y} = \mathrm{diag}(Y)$ (A.4) is given by

$$\mathsf{E}\left[\boldsymbol{y}\boldsymbol{y}' - \mathsf{E}[\boldsymbol{y}]\,\mathsf{E}[\boldsymbol{y}]'\right] = \Phi, \tag{A.12}$$

with elements,

$$\phi_{i,j} = \frac{\partial}{\partial l_{F,i} \partial l_{F,j}} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}L_F^2\right), \quad i, j = 1, \dots, r. \tag{A.13}$$

[4] as opposed to the economic SVD used in Section 1.

Transformation (A.4) is one-to-one, with unit Jacobian. Hence, uncertainty bounds on variables $Y$ and $Z$ can be mutually mapped using (A.4). However, the mapping $\boldsymbol{y} = \mathrm{diag}\,(Y)$ is many-to-one, and so also is $Z \rightarrow \boldsymbol{y}$. Conversion of second moments (and uncertainty bounds) of $\boldsymbol{y}$ to those of $Z$ (via (A.4), (A.5)) is therefore available in implicit form only. For example, the upper bound subspace within the support of $Z$ is expressible as follows:

$$\overline{Z} = \left\{ Z |\ \mathrm{diag}\left(U_F' Z V_F\right) = \overline{\boldsymbol{y}} \right\},$$

where $\overline{\boldsymbol{y}}$ is an appropriately chosen upper bound on $\boldsymbol{y}$. The lower bound, $Z$, is similarly constructed via a lower bound, $\underline{\boldsymbol{y}}$, on $\boldsymbol{y}$.

It remains, then, to choose appropriately the bounds $\underline{\boldsymbol{y}}$ and $\overline{\boldsymbol{y}}$ from (A.5). We use the first two moments, (A.7) and (A.12), to approximate (A.5) by a Gaussian. The Maximum Entropy (MaxEnt) principle (Jaynes 2003) ensures that uncertainty bounds for this MaxEnt Gaussian approximation of (A.5) enclose the uncertainty bounds of all distributions with the same first two moments. Highest Posterior Density (HPD) regions, and thus uncertainty bounds, for the Gaussian distribution with moments (A.8) and (A.13) are readily proposed. For example:

$$\Pr\left(-2\sqrt{\phi_i} < (y_i - \psi_i) \leq 2\sqrt{\phi_i}\right) \doteq 0.95. \tag{A.14}$$

Therefore, we choose

$$\overline{y_i} = \psi_i + 2\sqrt{\phi_i}, \tag{A.15}$$

$$\underline{y_i} = \psi_i - 2\sqrt{\phi_i}. \tag{A.16}$$

The required vector bounds are then constructed as $\overline{\boldsymbol{y}} = [\overline{y_1}, \ldots, \overline{y_r}]'$, and similarly for $\underline{\boldsymbol{y}}$. The geometric relationship between variables $Z$ and $\boldsymbol{y}$ is illustrated graphically for $p = 2$ and $n = r$ in Figure 1.

# Appendix B: Evaluation of Hypergeometric Functions

Numerical evaluation of the OVPCA algorithm requires evaluation of the following transformations of the hypergeometric function, ${}_0F_1$, of matrix argument: (i) its natural logarithm (ln), for Bayesian rank selection (77), and (ii) the first derivative of the ln, required for the first moment of the von-Mises-Fisher distribution (A.8). Analytical closed-form solutions are not known to us. Recently, a very good approximation of ${}_0F_1$ of matrix argument was developed (Butler and Wood 2003). It is based on the Laplace approximation at the saddle point. It yields reliable results for use in (i). Unfortunately, the first derivative of ln of this approximation for higher singular values, i.e. $l_i \gg 1$, are greater than one, thus placing the corresponding mean value $\mathsf{E}\left[y_i | \boldsymbol{l}_F\right]$ (A.8) outside of the unit circle, which is not permissible (Figure 1). Therefore, we now develop an approximation which overcomes this difficulty, by first considering the hypergeometric function ${}_0F_1$ of scalar argument.

## B.1 Hypergeometric function of scalar argument

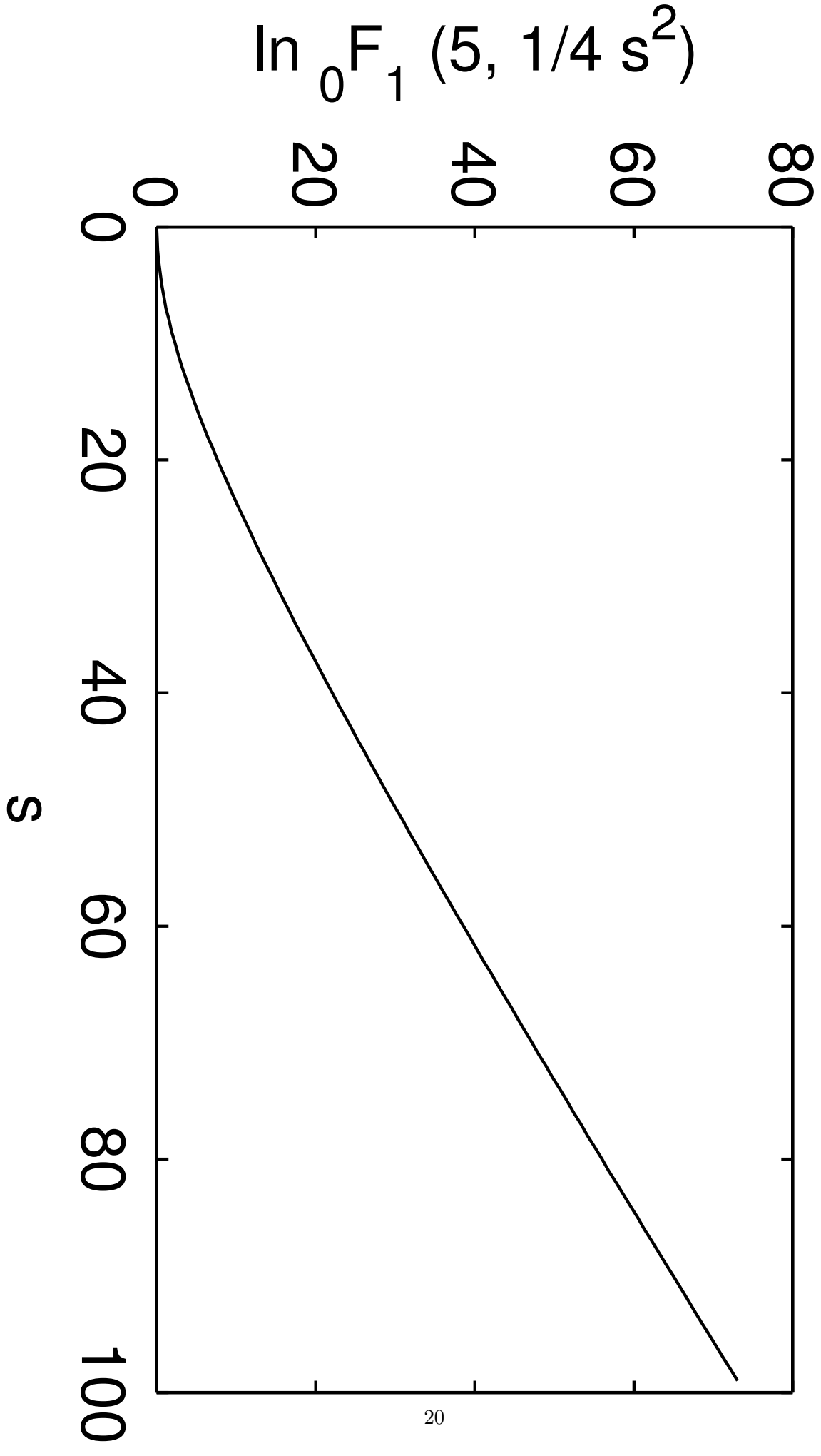
The natural logarithm (ln) of the hypergeometric function, ${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right)$, of a scalar argument can be expressed as

$$\ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = \ln \mathcal{B}\left(\frac{1}{2}p - 1, s\right) + \left(\frac{1}{2}p - 1\right)(\ln 2 - \ln(s)) + \ln \Gamma\left(\frac{1}{2}p\right), \tag{B.17}$$

where $\mathcal{B}$ denotes the modified Bessel function of the first kind (Abramowitz and Stegun 1972). (B.17) is plotted as a function of $s$ in Figure 3 (left), for $p = 5$. The first two derivatives of (B.17) are:

$$\frac{d}{ds} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = 2\frac{\mathcal{B}\left(\frac{1}{2}p, 2s\right)}{\mathcal{B}\left(\frac{p}{2} - 1, 2s\right)}, \tag{B.18}$$

$$\frac{d^2}{ds^2} \ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = 4\frac{\mathcal{B}\left(\frac{p}{2} + 1, 2s\right)}{\mathcal{B}\left(\frac{p}{2} - 1, 2s\right)} - 4\left[\frac{\mathcal{B}\left(\frac{1}{2}p, 2s\right)}{\mathcal{B}\left(\frac{p}{2} - 1, 2s\right)}\right]^2 + 2\frac{\mathcal{B}\left(\frac{1}{2}p, 2s\right)}{s\mathcal{B}\left(\frac{p}{2} - 1, 2s\right)}. \tag{B.19}$$


ln $_0F_1$ (5, 1/4 $s^2$)
0
20
40
60
80
0
20
40
60
80
100
s

The first derivative is illustrated in Figure 3 (right), for the same case, ($p = 5$). (B.18) can be expressed as a continuous fraction expansion (Abramowitz and Stegun 1972):

$$\frac{d}{ds}\ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = \frac{s_i}{\frac{p}{2}\left[1 + \frac{\frac{1}{4}s_i^2}{\left(\frac{p}{2}+1\right)\left(\frac{p}{2}+2\right)\left[1+\frac{\frac{1}{4}s_i^2}{\left(\frac{p}{2}+2\right)\left(\frac{p}{2}+3\right)+\left[1+\frac{\cdots}{\cdots}\right]}\right]}\right]}. \tag{B.20}$$

Furthermore, (B.19) can be expressed in terms of (B.18) and, therefore, (B.20). Expansion (B.20) converges very fast for $s < p$. However, when $s \gg p$ (say $s > 10p$) the convergence is quite slow. For large $s$, a more numerically efficient approximation is obtained via a Taylor expansion of (B.18) at $s \to \infty$:

$$\frac{d}{ds}\ln {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}s^2\right) = 1 - \left(\frac{p-1}{2s}\right)\exp\left(-\frac{p-3}{4s}\right) + o(5). \tag{B.21}$$

This gives an excellent approximation in the case $s \gg p$.

### B.2 Approximation of ${}_0F_1$ of matrix argument by ${}_0F_1$ of scalar arguments

Consider the special case of the von-Mises-Fisher matrix distribution (A.1) with $Z = [\boldsymbol{z}_1, \boldsymbol{z}_2] \in \Re^{p\times 2}$, and parameter $F = [\boldsymbol{f}_1, \boldsymbol{f}_2] \in \Re^{p\times 2}$, with added constraint that $\boldsymbol{f}_1, \boldsymbol{f}_2$ are mutually orthogonal: $\boldsymbol{f}_1'\boldsymbol{f}_2 = 0$. Then, the marginal distribution of $\boldsymbol{z}_1$ is Khatri and Mardia (1977):

$$f(\boldsymbol{z}_1|F) = \frac{{}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4}\left(I_p - \boldsymbol{z}_1\boldsymbol{z}_1'\right)\boldsymbol{f}_2\boldsymbol{f}_2'\right)}{{}_0F_1\left(\frac{1}{2}p, \frac{1}{4}FF'\right)C(p,1)}\exp\left(\mathrm{tr}\left(\boldsymbol{f}_1'\boldsymbol{z}_1\right)\right). \tag{B.22}$$

Note that the maximum (A.3) of the full pdf (A.1) occurs at:

$$\hat{\boldsymbol{z}}_1 = \arg\max_{\boldsymbol{z}_1} f(Z|F) = \boldsymbol{f}_1/\sqrt{\boldsymbol{f}_1'\boldsymbol{f}_1}.$$

This is orthogonal to $\boldsymbol{f}_2$, i.e. $\hat{\boldsymbol{z}}_1'\boldsymbol{f}_2 = 0$. Therefore, the contribution of the quadratic term in the argument of ${}_0F_1$ in the numerator of (B.22) would be negligible for values of $\boldsymbol{z}_1$ around $\hat{\boldsymbol{z}}_1$. Hence, we make the approximation

$${}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4}\left(I_p - \boldsymbol{z}_1\boldsymbol{z}_1'\right)\boldsymbol{f}_2\boldsymbol{f}_2'\right) \approx {}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4}\boldsymbol{f}_2\boldsymbol{f}_2'\right). \tag{B.23}$$

This will be satisfied when $f(\boldsymbol{z}_1|F)$ (B.22) is not diffuse, i.e. when $\boldsymbol{f}_1$ is large (see Section A.1). Under this approximation, the leading fraction in (B.22) is independent of $\boldsymbol{z}_1$, and thus acts as a normalizing coefficient. Distribution (B.22) is then of the von-Mises-Fisher type, namely $f(\boldsymbol{z}_1|F) \approx f(\boldsymbol{z}_1|\boldsymbol{f}_1) = \mathcal{M}(\boldsymbol{f}_1)$ (A.1). Comparing the normalizing coefficient in (B.22) with that in (A.2) yields

$${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}F'F\right) \approx {}_0F_1\left(\frac{1}{2}p, \frac{1}{4}\boldsymbol{f}_1\boldsymbol{f}_1'\right){}_0F_1\left(\frac{1}{2}(p-1), \frac{1}{4}\boldsymbol{f}_2\boldsymbol{f}_2'\right). \tag{B.24}$$

Extending (B.22) into higher dimension and using the chain rule of pdfs we obtain an approximation of the following type:

$${}_0F_1\left(\frac{1}{2}p, \frac{1}{4}L_F^2\right) \approx \prod_{i=1}^{p} {}_0F_1\left(\frac{1}{2}p - i + 1, \frac{1}{4}l_{F,i}^2\right). \tag{B.25}$$

## Appendix C: Remaining Distributions

### C.1 Truncated Normal Distribution

The truncated normal distribution of random variable $x$ is defined as normal—with functional form $\mathcal{N}\left(\mu, s^2\right)$—on a restricted support $a < x \leq b$. Its pdf is

$$f(x|\mu, s; (a,b]) = \frac{\sqrt{2}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{s}\right)^2\right)}{s\sqrt{\pi}\left(\mathrm{erf}(\beta) - \mathrm{erf}(\alpha)\right)}\chi((a,b]), \tag{C.26}$$

where $\alpha = \frac{a-\mu}{s\sqrt{2}}$, $\beta = \frac{b-\mu}{s\sqrt{2}}$. The first two moments of (C.26) are

$$\begin{aligned} \widehat{x} &= \mu - s\,\zeta(\mu, s), \\ \widehat{x^2} &= s^2 + \mu\widehat{x} - s\rho(\mu, s), \end{aligned}$$

which depend on the auxiliary functions

$$\zeta(\mu, s) = \frac{\sqrt{2}\left[\exp\left(-\beta^2\right) - \exp\left(-\alpha^2\right)\right]}{\sqrt{\pi}\left(\operatorname{erf}(\beta) - \operatorname{erf}(\alpha)\right)}, \tag{C.27}$$

$$\rho(\mu, s) = \frac{\sqrt{2}\left[b\exp\left(-\beta^2\right) - a\exp\left(-\alpha^2\right)\right]}{\sqrt{\pi}\left(\operatorname{erf}(\beta) - \operatorname{erf}(\alpha)\right)}. \tag{C.28}$$

(C.27) and (C.28) with vector arguments—e.g. $\rho(\boldsymbol{m}, s)$—are evaluated element-wise. HPD regions for this distribution can also be obtained. However, in our examples (Section 7), we use HPD regions calculated from the approximating MaxEnt (non-truncated) normal distribution, namely

$$\max\left(a, -2\sqrt{\widehat{x^2}}\right) < x - \widehat{x} < \min\left(b, 2\sqrt{\widehat{x^2}}\right). \tag{C.29}$$

The MaxEnt principle was already invoked in Appendix A.3.

### C.2 Gamma Distribution

The Gamma distribution has pdf

$$f(x|a, b) = \mathcal{G}(a, b) = \frac{b^a}{\Gamma(a)} x^{a-1}\exp(-bx)\,\chi([0, \infty)),$$

where $a > 0$, and $b > 0$, and $\Gamma(a)$ is the Gamma function (Abramowitz and Stegun 1972) evaluated at $a$. Its first moment, required in (53), is:

$$\widehat{x} = \frac{a}{b}.$$

## References


Abramowitz, M. and Stegun, I. (1972), *Handbook of Mathematical Functions*, New York: Dover Publications, Inc.

Anderson, T. W. (1971), *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons.

Beal, M. J. and Ghahramani, Z. (2003), "The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures," in *Bayesian Statistics 7*, ed. Bernardo, J. M. et. al., Oxford University Press.

Benali, H., Buvat, I., Frouin, F., Bazin, J. P., and Di Paola, R. (1993), "A statistical model for the determination of the optimal metric in factor analysis of medical image sequences (FAMIS)," *Physics in Medicine and Biology*, 38, 1065–1080.

Bernardo, J. and Smith, A. (1997), *Bayesian theory*, Chichester, New York, Brisbane, Toronto, Singapore: John Wiley & Sons, 2nd edition.

Bishop, C. M. (1999), "Variational Principal components," in *Proceedings of the Ninth International Conference on Artificial Neural Networks*, ICANN.

Butler, R. W. and Wood, T. A. (2003), "Laplace Approximation of Bessel Function of Matrix Argument," *Journal of Computational and Applied Mathematics*, 155, 359–382.

Buvat, I., Benali, H., and Di Paola, R. (1998), "Statistical distribution of factors and factor images in factor analysis of medical image sequences," *Physics in Medicine and Biology*, 43, 1695–1711.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, 39, 1–38.

Downs, T. D. (1972), "Orientational statistics," *Biometrica*, 59, 665–676.

Ghahramani, Z. and Beal, M. J. (2000), "Variational inference for Bayesian Mixtures of Factor analyzers," *Neural Information Processing Systems*, 12, 449–455.

Golub, G. and VanLoan, C. (1989), *Matrix Computations*, Baltimore, London: The John Hopkins University Press.

Hotelling, H. (1933), "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, 24, 417–441.

James, A. T. (1964), "Distribution of matrix variates and latent roots derived from normal samples," *Annals of Mathematical Statistics*, 35, 475–501.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press, 3rd ed.

Jolliffe, I. (2002), *Principal Component Analysis*, Springer-Verlag, 2nd ed.

Kay, S. M. (1993), *Fundamentals Of Statistical Signal Processing: Estimation Theory*, Prentice Hall.

Khatri, C. G. and Mardia, K. V. (1977), "The von Mises-Fisher Distribution in Orientation Statistics," *Journal of Royal Statistical Society B*, 39, 95–106.

Kullback, S. and Leibler, R. (1951), "On information and sufficiency," *Annals of Mathematical Statistics*, 22, 79–87.

Mardia, K. and Jupp, P. E. (2000), *Directional Statistics*, Chichester, England: John Wiley and Sons.

Miskin, J. W. (2000), "Ensemble Learning for Independent Component Analysis," Ph.D. thesis, University of Cambridge.

Pearson, K. (1901), "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science*, 2, 559–572.

Press, S. J. and Shigemasu, K. (1989), "Bayesian Inference in Factor Analysis," in *Contributions to Probability and Statistics*, ed. Glesser, L. J., Springer Verlag, New York, chap. 15.

Quinn, A. (1998), "Regularized Signal Identification using Bayesian Techniques," in *Signal Analysis and Prediction*, Birkhäuser Boston Inc.

Rowe, D. B. and Press, S. J. (1998), "Gibbs Sampling and Hill Climbing In Bayesian Factor Analysis," Tech. rep., University of California, Riverside.

Rubin, D. and Thayer, D. (1982), "EM algorithms for ML factor analysis," *Psychometrica*, 47, 69–76.

Tipping, M. E. and Bishop, C. M. (1998a), "Mixtures of Probabilistic Principal component analyzers," Tech. rep., Aston University.

— (1998b), "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, 61, 611–622.