Products of Gaussians and Probabilistic Minor Component Analysis

C.K.I. Williams
Division of Informatics, University of Edinburgh, Edinburgh EH1 2QL, UK
c.k.i.williams@ed.ac.uk http://anc.ed.ac.uk

F.V. Agakov
System Engineering Research Group, Chair of Manufacturing Technology
Friedrich-Alexander-University Erlangen-Nuremberg, 91058 Erlangen, Germany
F.Agakov@lft.uni-erlangen.de

July 2, 2001

Abstract

Recently Hinton introduced the Products of Experts (PoE) architecture for density estimation, where individual expert probabilities are multiplied and re-normalized. We consider products of Gaussian "pancakes" equally elongated in all directions except one, and prove that the maximum likelihood solution for the model gives rise to a minor component analysis solution. We also discuss the covariance structure of sums and products of Gaussian pancakes or 1-factor probabilistic principal component analysis (PPCA) models.

1 Introduction

Recently Hinton (1999) introduced a new Product of Experts (PoE) model for combining expert probabilities, where probability $p(\mathsf{x}|\Theta)$ is computed as a normalized multiplication of probabilities $p_i(\mathsf{x}|\Theta_i)$ of individual experts with parameters Θ_i:

$$p(\mathsf{x}|\Theta) = \frac{\prod_{i=1}^{m} p_i(\mathsf{x}|\Theta_i)}{\int_{\mathsf{x}'} \prod_{i=1}^{m} p_i(\mathsf{x}'|\Theta_i)\mathrm{d}\mathsf{x}'}. \tag{1}$$

Under the PoE paradigm experts may assign high probabilities to irrelevant regions of the data space, as long as these probabilities are small under other experts (Hinton, 1999).

Here we will consider a product of constrained Gaussians in the d-dimensional data space, whose probability contours resemble d-dimensional Gaussian "pancakes" (GP), contracted in one dimension and equally elongated in the other $d-1$ dimensions. We will refer to the model as product of Gaussian pancakes (PoGP) and show that it provides a probabilistic technique for minor component analysis (MCA). This MCA solution contrasts with Probabilistic PCA (PPCA) (Tipping and Bishop (1999); see also Roweis (1998)), which is a probabilistic method for PCA based on a factor analysis model with isotropic noise. The key difference is that PPCA is a model for the covariance matrix of the data, while PoGP is a model for the *inverse* covariance matrix of the data.

In section 2 we discuss products of Gaussians. In section 3 we consider products of Gaussian pancakes, derive analytic solutions for maximum likelihood estimators for the parameters and provide experimental evidence that the analytic solution is correct. Section 4 discusses the covariance structure of sums and products of Gaussian pancakes and 1-factor PPCA models.

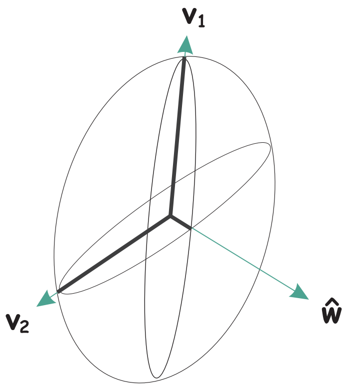

Figure 1: Probability contour of a Gaussian “pancake” in $\mathbb{R}^3$.

2 Products of Gaussians

If each expert in (1) is a Gaussian $p_i(\mathsf{x}|\Theta_i) \sim N(\mu_i, \mathsf{C}_i)$, the resulting distribution of the product may be expressed as

$$p(\mathsf{x}|\Theta) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(\mathsf{x}-\mu_i)^T \mathsf{C}_i^{-1}(\mathsf{x}-\mu_i)\right\}.$$

By completing the quadratic term in the exponent it may be easily shown that $p(\mathsf{x}|\Theta) \sim N(\mu_\Sigma, \mathsf{C}_\Sigma)$, where the inverse covariance C_Σ^{-1} of the product is given as a summation of inverse covariance matrices C_i^{-1} of each expert:

$$\mathsf{C}_\Sigma^{-1} = \sum_{i=1}^{m} \mathsf{C}_i^{-1} \subset \mathbb{R}^{d\times d}, \qquad \mu_\Sigma = \mathsf{C}_\Sigma\left(\sum_i \mathsf{C}_i^{-1}\mu_i\right) \subset \mathbb{R}^d. \tag{2}$$

To simplify the following derivations we will assume that $p_i(\mathsf{x}|\Theta_i) \sim N(\mathsf{0}, \mathsf{C}_i)$ and thus that $p(\mathsf{x}|\Theta) \sim N(\mathsf{0}, \mathsf{C}_\Sigma)$. $\mu_\Sigma \neq \mathsf{0}$ can be obtained by translation of the coordinate system. The log-likelihood for *i.i.d.* data under the PoG is then expressed as

$$\mathcal{L}(\mathsf{C}_\Sigma) \quad = \quad -\frac{N}{2}d\ln 2\pi + \frac{N}{2}\ln|\mathsf{C}_\Sigma^{-1}| - \frac{N}{2}\mathrm{tr}[\mathsf{C}_\Sigma^{-1}\mathsf{S}]. \tag{3}$$

Here N is the number of sample points and S is the sample covariance matrix with the assumed zero-mean, i.e.

$$\mathsf{S} = \frac{1}{N}\sum_{i=1}^{N}\mathsf{x}_i\mathsf{x}_i^T, \tag{4}$$

where x_i denotes the i^{th} data point.

3 Products of Gaussian Pancakes

In this section we will describe the covariance structure of a GP expert and a product of GP experts, and discuss the maximum likelihood (ML) solution for the parameters of the PoGP model.

3.1 Covariance Structure of a GP Expert

Consider a d-dimensional Gaussian whose probability contours are contracted in direction $\hat{\mathsf{w}}$ and equally elongated in directions $\mathsf{v}_1, \ldots, \mathsf{v}_{d-1}$ (see Figure 1). Its inverse covariance may be written as

$$\mathsf{C}^{-1} = \sum_{i=1}^{d-1} \mathsf{v}_i \mathsf{v}_i^T \beta_0 + \hat{\mathsf{w}}\hat{\mathsf{w}}^T \beta_{\hat{w}} \subset \mathbb{R}^{d\times d}, \tag{5}$$

where $\mathsf{v}_1, \ldots, \mathsf{v}_{d-1}, \hat{\mathsf{w}}$ form a $d \times d$ matrix of normalized eigenvectors of the covariance C. $\beta_0 = \sigma_0^{-2}$, $\beta_{\hat{w}} = \sigma_{\hat{w}}^{-2}$ define inverse variances in the directions of elongation and contraction respectively, so that $\sigma_0^2 \geq \sigma_{\hat{\mathsf{w}}}^2$. Expression (5) can be re-written in a more compact form as

$$\begin{aligned} \mathsf{C}^{-1} &= \beta_0 \mathsf{I}_d + (\beta_{\hat{w}} - \beta_0)\hat{\mathsf{w}}\hat{\mathsf{w}}^T \\ &= \beta_0 \mathsf{I}_d + \mathsf{w}\mathsf{w}^T, \end{aligned} \tag{6}$$

where $\mathsf{w} = \hat{\mathsf{w}}\sqrt{\beta_{\hat{w}} - \beta_0}$ and $\mathsf{I}_d \subset \mathbb{R}^{d\times d}$ is the identity matrix. Notice that according to the constraint considerations $\beta_0 < \beta_{\hat{w}}$, and all elements of w are real-valued.

We see that covariance of the data C of a GP expert can be uniquely determined by the weight vector w, collinear with the direction of contraction, and the variance in the direction of elongation $\sigma_0^2 = \beta_0^{-1}$. We can further notice the similarity of (6) with expression for the covariance of the data of a 1-factor probabilistic principal component analysis model $\mathsf{C} = \sigma^2 \mathsf{I}_d + \mathsf{w}\mathsf{w}^T$ (Tipping and Bishop, 1999), where σ^2 is the variance of the factor-independent spherical Gaussian noise. The only difference is that it is the *inverse* covariance matrix for the constrained Gaussian model rather than the covariance matrix which has the structure of a rank-1 update to a multiple of I_d.

3.2 Covariance of the PoGP Model

We will now consider a product of m GP experts, each of which is contracted in a single dimension. We will refer to the model as a $(1, m)$ PoGP, where 1 represents the number of directions of contraction of each expert. We also assume that all experts have identical means.

From (2) and (5) the inverse covariance of the the resulting $(1, m)$ PoGP model is expressed as

$$\mathsf{C}_\Sigma^{-1} = \sum_{i=1}^{m} \mathsf{C}_i^{-1} \quad = \quad \beta_\Sigma \mathsf{I}_d + \mathsf{W}\mathsf{W}^T \subset \mathbb{R}^{d\times d}, \tag{7}$$

where columns of $\mathsf{W} \subset \mathbb{R}^{d\times m}$ correspond to weight vectors of m PoGP experts, and $\beta_\Sigma = \sum_{i=1}^{m} \beta_0^{(i)} > 0$.

Comparing (7) with the m-factor PPCA we can make a conjecture that in contrast with the PPCA model where ML weights correspond to principal components of the data covariance (Tipping and Bishop, 1999), weights W of the PoGP model define projection onto m *minor* eigenvectors of the sample covariance in the visible d-dimensional space, while the distortion term $\beta_\Sigma \mathsf{I}_d$ explains larger variations[1]. The proof of this conjecture is discussed in Appendix A.

For the covariance described by (7), we find that

$$p(\mathsf{x}) \propto \exp -\frac{1}{2}\{\mathsf{x}^T(\beta_\Sigma \mathsf{I}_d + \sum_{i=1}^{m} \alpha_i \hat{\mathsf{w}}_i \hat{\mathsf{w}}_i^T)\mathsf{x}\}, \tag{8}$$

where $\hat{\mathsf{w}}_i$ is a unit vector in the direction of w_i and $\alpha_i = |\mathsf{w}_i|^2$. This distribution can be given a maximum entropy interpretation. From Cover and Thomas (1991, equation 11.4), the maximum entropy distribution obeying the constraints $E[f_i(\mathsf{x})] = r_i,\ i = 1, \ldots, c$ is of the form $p(\mathsf{x}) \propto \exp - \sum_{i=1}^{c} \lambda_i f_i(\mathsf{x})$. Hence we see that (8) can be interpreted as a maximum entropy distribution with constraints on $E[(\hat{\mathsf{w}}_i^T \mathsf{x})^2]$, $i = 1, \ldots, m$ and on $E[\mathsf{x}^T\mathsf{x}]$.

[1] Because equation 7 has the form of a factor analysis decomposition, but for the *inverse* covariance matrix, we sometimes refer to PoGP as the *rotcaf* model.

3.3 Maximum-Likelihood Solution for PoGP

In Appendix A it is shown that the likelihood (3) is maximized when

$$\mathsf{W}_{ML} = \mathsf{U}_m(\Lambda^{-1} - \beta_\Sigma^{ML}\mathsf{I}_m)^{1/2}\mathsf{R}^T, \qquad \beta_\Sigma^{ML} = \frac{d-m}{\sum_{i=m+1}^{d}\lambda_i}, \tag{9}$$

where U_m is the $d \times m$ matrix of the m minor components of the sample covariance S, Λ is the $m \times m$ matrix of the corresponding eigenvalues and R is an arbitrary $m \times m$ orthogonal rotation matrix. As in PPCA (Tipping and Bishop, 1999), the distortion term accounts for the variance in the directions lying outside the space spanned by W_{ML}.

Thus, the maximum likelihood solution for the weights of the $(1, m)$ PoGP model corresponds to m scaled and rotated minor eigenvectors of the sample covariance S and leads to a probabilistic model of minor component analysis. As in the PPCA model, the number of experts m is assumed to be lower than the dimension of the data space d.

3.4 Experimental Confirmation

In order to confirm the analytic results we have performed experiments comparing the derived ML parameters of the PoGP model with convergence points of scaled conjugate gradient (SCG) optimization (Bishop, 1995; Møller, 1993) performed on the log-likelihood (3).

We considered different data sets in data spaces of dimensions d, varying from 2 to 25 with $m = 1, \dots, d-1$ constrained Gaussian experts. For each choice of d and m we looked at three types of sample covariance matrices, resulting in different types of solutions for the weights $\mathsf{W} \subset \mathbb{R}^{d\times m}$. In the first case, S was set to have $s \leq d - m$ identical largest eigenvalues, and we expected all expert weights to be retained in W. In the second case, S was set up in such a way that $d - m < s < d$, so that variance in some directions could be explained by the noise term, and some columns of W could be equal to zeros. Finally, we have considered the degenerate case when the sample covariance was a multiple of the identity matrix I_d and expected $\mathsf{W} = 0$ [see (9), (6)].

For all of the considered cases, we performed 30 runs of the scaled conjugate gradient optimization of the likelihood, started at different points of the parameter space $\{\mathsf{W}, \beta_\Sigma\}$. To ensure that β_Σ was non-negative, we parameterized it as $\beta_\Sigma = e^x$, $x \in \mathbb{R}$. For each data set, the SCG algorithm converged to approximately the same value of the log-likelihood (with the largest observed standard deviation of the likelihood $\sim 10^{-12}$). The largest observed absolute error between the analytic $\{\mathsf{W}^{(A)}, \beta_\Sigma^{(A)}\}$ and the SCG $\{\mathsf{W}^{(SCG)}, \beta_\Sigma^{(SCG)}\}$ parameters satisfied

$$\begin{aligned} \max_{i,j}\left[|\mathsf{W}^{(A)} - \mathsf{W}^{(SCG)}\mathsf{R}|\right]_{ij} &\lessapprox 10^{-8} \\ |\beta_\Sigma^{(A)} - \beta_\Sigma^{(SCG)}| &\lessapprox 10^{-8}, \end{aligned}$$

where $\mathsf{R} \subset \mathbb{R}^{m\times m}$ is a rotation matrix. Moreover, each numeric solution resulted in the expected type of $\mathsf{W}^{(SCG)}$ for each type of the sample covariance.

The experiments therefore confirm that the method of the scaled conjugate gradients converges to a point in the parameter space that results in the same values of the log-likelihood as the analytically obtained solution (up to the expected arbitrary rotation factor for the weights).

4 Related Covariance Structures

Above we have discussed the covariance structure of a PoGP model. For real-valued densities, an alternative method for combining experts is to sum the random vectors produced by each expert, i.e. $\mathsf{x} = \sum_{i=1}^{m} \mathsf{x}_i$; let us call this a Sum of Experts (SoE) model. (Note that the sum of experts model refers to the sum of the random vectors, while the product of experts model refers to the product of their densities. The distribution of the sum of random variables is obtained by the *convolution* of their densities.) For zero-mean Gaussians, the covariance of the Sum of Experts is simply the sum of the individual covariances, in contrast to the PoE, where the overall inverse covariance is the sum of the individual inverse covariances. Hence we see that the PPCA model is in fact a SoE model where each expert is a 1-factor PPCA model.

This leads us to ask the question, what are the covariance structures for (a) a product of 1-factor PPCA models and (b) a sum of Gaussian pancakes? These questions are discussed below.

4.1 Covariance of the PoPPCA Model

We consider a product of m 1-factor PPCA models, denoted as $(1, m)$ PoPPCA. Geometrically, the probability contours of a one-factor PPCA model in $\mathbb{R}^d$ are d-dimensional hyper-ellipsoids elongated in one dimension and contracted in the other $d-1$ directions. The covariance matrix of 1-PPCA is expressed as

$$\mathsf{C} = \mathsf{w}\mathsf{w}^T + \sigma^2 \mathsf{I} \subset \mathcal{R}^{d\times d}, \tag{10}$$

where weight vector $\mathsf{w} \subset \mathbb{R}^d$ defines the direction of elongation and σ^2 is the variance in the direction of contraction. Its inverse covariance is

$$\mathsf{C}^{-1} = \beta \mathsf{I}_d - \frac{\beta^2 \mathsf{w}\mathsf{w}^{\mathrm{T}}}{1 + \mathsf{w}^{\mathrm{T}}\mathsf{w}\beta} = \beta \mathsf{I}_d - \beta\gamma \mathsf{w}\mathsf{w}^{\mathrm{T}}, \tag{11}$$

where $\beta = \sigma^{-2}$ and $\gamma = \beta/(1 + ||\mathsf{w}||^2\beta)$. β and γ are the inverse variances in the directions of contraction and elongation respectively.

Plugging (11) into (2) we obtain

$$\mathsf{C}_\Sigma^{-1} \;\; = \;\; \sum_{i=1}^{m} \mathsf{C}_i^{-1} = \beta_\Sigma \mathsf{I}_d - \mathsf{W}\mathsf{B}\mathsf{W}^T, \;\; \mathsf{B} = \mathrm{diag}(\beta_1\gamma_1, \; \beta_2\gamma_2, \; \dots, \beta_m\gamma_m) \tag{12}$$

where $\mathsf{W} = [\mathsf{w}^{(1)}, \dots, \mathsf{w}^{(m)}] \subset \mathbb{R}^{d\times m}$ is the weight matrix with columns corresponding to the weights of individual experts, β_Σ is the sum of the inverse noise variances for all experts, and B may be thought of as a squared scaling factor on the columns of W. We can rewrite expression (12) as

$$\mathsf{C}_\Sigma^{-1} \;\; = \;\; \beta_\Sigma \mathsf{I}_d - \tilde{\mathsf{W}}\tilde{\mathsf{W}}^T \subset \mathbb{R}^{d\times d}, \tag{13}$$

where $\tilde{\mathsf{W}} = \mathsf{W}\mathsf{B}^{1/2}$ are implicitly scaled weights.

4.2 Maximum-Likelihood Solution for PoPPCA

Our studies show that the ML solution for the $(1, m)$ PoPPCA model can be equivalent to the ML solution for m-factor PPCA, but only when rather restrictive conditions apply. Consider the simplified case when the noise variance β^{-1} is the same for each 1-factor PPCA model. Then (12) reduces to

$$\mathsf{C}_\Sigma^{-1} = m\beta \mathsf{I}_d - \beta \mathsf{W}\Gamma\mathsf{W}^T, \qquad \Gamma = \mathrm{diag}(\gamma_1, \; \gamma_2, \; \dots, \gamma_m). \tag{14}$$

A m-factor PPCA model has covariance $\sigma^2 \mathsf{I}_d + \mathsf{W}\mathsf{W}^T$ and thus, by the Woodbury formula (see e.g. Press et al. (1992)), it has inverse covariance $\beta \mathsf{I}_d - \beta \mathsf{W}(\sigma^2 \mathsf{I}_m + \mathsf{W}^T\mathsf{W})^{-1}\mathsf{W}^T$. The maximum likelihood solution for a m-PPCA model is similar to (9), i.e. $\hat{\mathsf{W}} = \mathsf{U}(\Lambda - \sigma^2 \mathsf{I}_m)^{1/2}\mathsf{R}^T$, but now Λ is a diagonal matrix of the m *principal* eigenvalues, and U is a matrix of the corresponding eigenvectors. If we choose $\mathsf{R}^T = \mathsf{I}_m$ then the columns of $\hat{\mathsf{W}}$ are orthogonal and the inverse covariance of the maximum likelihood m-PPCA model has the form $\beta \mathsf{I}_d - \beta \hat{\mathsf{W}}\Gamma\hat{\mathsf{W}}^T$. Comparing this to (14) (and setting $\mathsf{W} = \hat{\mathsf{W}}$) we see that the difference is that the first term of the RHS of (14) is $\beta m \mathsf{I}_d$, while for m-PPCA it is $\beta \mathsf{I}_d$.

In section 3.4 and Appendix C.3 of Agakov (2000) it is shown that (for $m \geq 2$) we obtain the m-factor PPCA solution when

$$\bar{\lambda} \leq \lambda_i < \frac{m}{m-1}\bar{\lambda}, \qquad i = 1, \dots, m, \tag{15}$$

where $\bar{\lambda}$ is the mean of the $d - m$ discarded eigenvalues, and λ_i is a retained eigenvalue; it is the smaller eigenvalues that are discarded. We see that the covariance must be nearly spherical for this condition to hold. For covariance matrices satisfying (15), this solution was confirmed by numerical experiments similar to those in section 3.4, as detailed in (Agakov, 2000, section 3.5).

To see why this is true intuitively, observe that C_i^{-1} for each 1-factor PPCA expert will be large (with value β) in all directions except one. If the directions of contraction for each C_i^{-1} are orthogonal, we see that the sum of the inverse covariances will be at least $(m-1)\beta$ in a contracted direction and $m\beta$ in a direction in which no contraction occurs.

The above shows that for certain types of sample covariance matrix the $(1, m)$ PoPPCA solution is not equivalent to the m-factor PPCA solution. Notice, however, that an m-PPCA can be modelled rather trivially by a m-factor k-expert $(k > 1)$ PoPPCA $[(m, k)$ PoPPCA$]$ by taking the product of a single m-PPCA expert with $k-1$ "large" spherical Gaussians, with all experts having identical means.

4.3 Sums of Gaussian Pancakes

A Gaussian pancake has $\mathsf{C}^{-1} = \beta_0 \mathsf{I}_d + \mathsf{w}\mathsf{w}^T$. By analogous arguments to those above we find that a sum of Gaussian pancakes has covariance $\mathsf{C}_{SGP} = \sigma^2 \mathsf{I}_d - \tilde{\mathsf{W}}\tilde{\mathsf{W}}^T$, where $\tilde{\mathsf{W}}$ is a rescaled W, with a somewhat different definition than in (13).

Analogously to section 4.2 above, we would expect that a ML solution for the sum of Gaussian pancakes would give a MCA solution when the sample covariance is near spherical, but that the solution would not have a simple relationship to the eigendecomposition of S when this is not the case.

5 Discussion

We have considered the product of m Gaussian pancakes. The analytic derivations for the optimal model parameters have confirmed the initial hypothesis that the PoGP gives rise to minor component analysis. We have also confirmed by experiment that the analytic solutions correspond to the ones obtained by applying optimization methods to the log-likelihood.

An intuitive interpretation of the PoGP model is as follows: Each Gaussian pancake imposes an approximate linear constraint in x space. Such a linear constraint is that x should lie close to a particular hyperplane. The conjunction of these constraints is given by the product of the Gaussian pancakes. If $m \ll d$ it will make sense to define the resulting Gaussian distribution in terms of the constraints. However, if there are many constraints $(m > d/2)$ then it can be more efficient to describe the directions of large variability using a PPCA model, rather than the directions of small variability using a PoGP model. This issue is discussed by Xu et al. (1991) in what they call the "Dual Subspace Pattern Recognition Method" where both PCA and MCA models are used (although their work does not use explicit probabilistic models such as PPCA and PoGP).

We have shown that $(1, m)$ PoGP can be viewed as a probabilistic MCA model. MCA can be used, for example, for signal extraction in digital signal processing (Oja, 1992), dimensionality reduction, and data visualization. Extraction of the minor component is also used in the Pisarenko Harmonic Decomposition method for detecting sinusoids in white noise (see, e.g. Proakis and Manolakis (1992), p. 911). Formulating minor component analysis as a probabilistic model simplifies comparison of the technique with other dimensionality reduction procedures, permits extending MCA to a mixture of MCA models (which will be modeled as a mixture of products of Gaussian pancakes), permits using PoGP in classification tasks (if each PoGP model defines a class-conditional density), and leads to a number of other advantages over non-probabilistic MCA models (see the discussion of advantages of PPCA over PCA in Tipping and Bishop (1999)).

In section 4 we have discussed the relationship of sums and products of Gaussian models, and shown that the sum of Gaussian pancakes and $(1, m)$ PoPPCA models have rather low representational power compared to PPCA or PoGP.

In this paper we have considered sums and products of Gaussians with Gaussian pancake or PPCA structure. It is possible to apply the sum and product operations to models with other covariance structures; for example Williams and Felderhof (2001) consider sums and products of tree-structured Gaussians, and study their relationship to AR and MA processes.

A ML Solutions for PoGP

In this appendix we derive the ML equations (9). In section A.1 we derive the conditions that must hold for a stationary point of the likelihood. In section A.2 these stationarity conditions are expressed in terms of the eigenvectors of S. In section A.3 it is shown that to maximize the likelihood it is the m *minor* eigenvectors that must be retained. The derivations in this appendix are the analogues for the PoGP model of those in Appendix A of Tipping and Bishop (1999) for the PPCA model.

A.1 Derivation of ML Equations

From (3) we can specify ML conditions for parameter Θ of a Gaussian:

$$\frac{\partial \mathcal{L}}{\partial \Theta} = \frac{N}{2}\left[\frac{\partial \ln |\mathsf{C}_\Sigma^{-1}|}{\partial \Theta} - \frac{\partial \mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S})}{\partial \Theta}\right] = 0. \tag{16}$$

To compute the derivatives we use the notation of Magnus and Neudecker (1999). If $y = \mathrm{tr}(\mathsf{A}^T\mathsf{X})$, then $dy = \mathrm{tr}(\mathsf{A}^T \mathrm{d}_\mathsf{X}\mathsf{X})$ implies that $\partial y/\partial \mathsf{X} = \mathsf{A}$ (see Magnus and Neudecker (1999), p. 176, Table 2). We can compute derivatives in the r.h.s. of (16) by first computing the corresponding differentials according to $\mathrm{d}_\Theta \ln |\mathsf{X}| = \mathrm{tr}(\mathsf{X}^{-1}\mathrm{d}_\Theta\mathsf{X})$ and $\mathrm{d}_\Theta \mathrm{tr}(\mathsf{X}) = \mathrm{tr}(\mathrm{d}_\Theta\mathsf{X})$.

In our case $\Theta = \{\mathsf{W}, x\}$, where W is the weight matrix and x is the log of the distortion term β_Σ. This leads to

$$\mathrm{d}_\mathsf{W} \ln |\mathsf{C}_\Sigma^{-1}| = \mathrm{tr}(\mathsf{C}_\Sigma \mathrm{d}_\mathsf{W}\mathsf{C}_\Sigma^{-1}) = 2\mathrm{tr}(\mathsf{W}^T\mathsf{C}_\Sigma(\mathrm{d}_\mathsf{W}\mathsf{W})), \tag{17}$$

$$\mathrm{d}_\mathsf{W}\mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S}) = \mathrm{tr}((\mathrm{d}_\mathsf{W}\mathsf{C}_\Sigma^{-1})\mathsf{S}) = 2\mathrm{tr}(\mathsf{W}^T\mathsf{S}(\mathrm{d}_\mathsf{W}\mathsf{W})), \tag{18}$$

$$\mathrm{d}_x \ln |\mathsf{C}_\Sigma^{-1}| = \mathrm{tr}(\beta_\Sigma \mathsf{C}_\Sigma \ \mathrm{d}_x x), \qquad \mathrm{d}_x \mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S}) = \mathrm{tr}(\beta_\Sigma m \mathsf{S} \ \mathrm{d}_x x). \tag{19}$$

Substituting into (16) and differentiating with respect to W and x we obtain expressions for the derivatives of the log-likelihood $\mathcal{L}$:

$$\frac{\partial \mathcal{L}}{\partial \mathsf{W}} = N(\mathsf{C}_\Sigma\mathsf{W} - \mathsf{S}\mathsf{W}), \qquad \frac{\partial \mathcal{L}}{\partial x} = \beta_\Sigma N \mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}). \tag{20}$$

Dropping the constant factors, we obtain ML equations for W and x:

$$\mathsf{C}_\Sigma\mathsf{W} - \mathsf{S}\mathsf{W} = 0, \qquad \mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}) = 0. \tag{21}$$

Note that in order to find the maximum-likelihood solution for the weight matrix W and term β_Σ both equations in (21) should hold simultaneously.

A.2 Stationary Points of the Log Likelihood

There are three classes of solutions to the equation $\mathsf{C}_\Sigma\mathsf{W} - \mathsf{S}\mathsf{W} = 0$, namely

$$\begin{array}{rcl} \mathsf{W} &=& 0; \\ \mathsf{S} &=& \mathsf{C}_\Sigma; \\ \mathsf{S}\mathsf{W} &=& \mathsf{C}_\Sigma\mathsf{W}, \ \ \mathsf{W} \neq 0, \ \ \mathsf{S} \neq \mathsf{C}_\Sigma. \end{array} \tag{22}$$

The first of these, $\mathsf{W} = 0$ is uninteresting, and corresponds to a minimum of the log-likelihood. In the second case, the model covariance is equal to the sample covariance and C_Σ is *exact*. In the third case, while $\mathsf{S}\mathsf{W} = \mathsf{C}_\Sigma\mathsf{W}$, $\mathsf{S} \neq \mathsf{C}_\Sigma$ and the model covariance is said to be *approximate*. By analogy with Tipping and Bishop (1999), we will consider the singular value decomposition of the weight matrix, and establish dependencies between left singular vectors of $\mathsf{W} = \mathsf{U}\mathsf{L}\mathsf{R}^T$ and eigenvectors of the sample covariance S. $\mathsf{U} = [\mathsf{u}_1, \mathsf{u}_2, \dots, \mathsf{u}_m] \subset \mathbb{R}^{d\times m}$ is a matrix of left singular vectors of W with columns constituting an orthonormal basis, $\mathsf{L} = \mathrm{diag}(l_1, l_2, \dots, l_m) \subset \mathbb{R}^{m\times m}$ is a diagonal matrix of the singular values of W and $\mathsf{R} \subset \mathbb{R}^{m\times m}$ defines an arbitrary rigid rotation of W.

A.2.1 Exact Model Covariance

Considering non-singularity of S and C_Σ, we find $\mathsf{C}_\Sigma = \mathsf{S} \Rightarrow {\mathsf{C}_\Sigma}^{-1} = \mathsf{S}^{-1}$. As ${\mathsf{C}_\Sigma}^{-1} = \beta_\Sigma \mathsf{I}_d + \mathsf{W}\mathsf{W}^T$, we obtain

$$\mathsf{W}\mathsf{W}^T = \mathsf{U}\mathsf{L}^2\mathsf{U}^T = \mathsf{S}^{-1} - \beta_\Sigma \mathsf{I}_d. \tag{23}$$

This has the known solution $\mathsf{W} = \mathsf{U}_m(\Lambda^{-1} - \beta_\Sigma \mathsf{I}_m)^{1/2}\mathsf{R}^T$, where U_m is the matrix of the m eigenvectors of S with the smallest eigenvalues and Λ is the corrresponding diagonal matrix of the eigenvalues. The sample covariance must be such that the largest $d-m$ eigenvalues are all equal to β_Σ; the other m eigenvalues are matched explicitly.

A.2.2 Approximate Model Covariance

Applying the matrix inversion lemma [see e.g. Press et al. (1992)] to the expression for the inverse covariance of the PoGP (7) gives

$$\mathsf{C}_\Sigma = \beta_\Sigma^{-1}\left(\mathsf{I}_d - \mathsf{W}(\beta_\Sigma \mathsf{I}_m + \mathsf{W}^T\mathsf{W})^{-1}\mathsf{W}^T\right) \subset \mathbb{R}^{d\times d}. \tag{24}$$

Solution of (21) for the approximate model covariance results in

$$\mathsf{C}_\Sigma \mathsf{W} = \mathsf{S}\mathsf{W} \Rightarrow \mathsf{C}_\Sigma \mathsf{U}\mathsf{L} = \mathsf{S}\mathsf{U}\mathsf{L}. \tag{25}$$

Substituting (24) for C_Σ we obtain

$$\begin{aligned}
\mathsf{C}_\Sigma \mathsf{U}\mathsf{L} &= (\beta_\Sigma^{-1}\mathsf{I}_d - \beta_\Sigma^{-1}\mathsf{W}(\beta_\Sigma + \mathsf{W}^T\mathsf{W})^{-1}\mathsf{W}^T)\mathsf{U}\mathsf{L} \\
&= (\beta_\Sigma^{-1}\mathsf{I}_d - \beta_\Sigma^{-1}\mathsf{U}\mathsf{L}\mathsf{R}^T(\beta_\Sigma \mathsf{I}_m + \mathsf{R}\mathsf{L}^2\mathsf{R}^T)^{-1}\mathsf{R}\mathsf{L}\mathsf{U}^T)\mathsf{U}\mathsf{L} \\
&= \mathsf{U}(\beta_\Sigma^{-1}\mathsf{I}_m - \beta_\Sigma^{-1}\mathsf{L}\mathsf{R}^T(\beta_\Sigma \mathsf{I}_m + \mathsf{R}\mathsf{L}^2\mathsf{R}^T)^{-1}\mathsf{R}\mathsf{L})\mathsf{L} \\
&= \mathsf{U}(\beta_\Sigma^{-1}\mathsf{I}_m - \beta_\Sigma^{-1}(\beta_\Sigma \mathsf{L}^{-2} + \mathsf{I}_m)^{-1})\mathsf{L}.
\end{aligned} \tag{26}$$

Thus

$$\mathsf{S}\mathsf{U}\mathsf{L} = \mathsf{U}(\beta_\Sigma^{-1}\mathsf{I}_d - \beta_\Sigma^{-1}(\beta_\Sigma \mathsf{L}^{-2} + \mathsf{I}_m)^{-1})\mathsf{L}. \tag{27}$$

Notice that the term $\beta_\Sigma^{-1}\mathsf{I}_d - \beta_\Sigma^{-1}(\beta_\Sigma \mathsf{L}^{-2} + \mathsf{I}_m)^{-1}$ in the r.h.s. of equation (27) is a diagonal matrix (i.e, just a scaling factor of U). Equation (27) defines the matrix form of the eigenvector equation, with both sides post-multiplied by the diagonal matrix L.

If $l_i \neq 0$ then (27) implies that

$$\mathsf{C}_\Sigma \mathsf{u}_i = \mathsf{S}\mathsf{u}_i = \lambda_i \mathsf{u}_i, \tag{28}$$

$$\lambda_i = \beta_\Sigma^{-1}(1 - (\beta_\Sigma l_i^{-2} + 1)^{-1}), \tag{29}$$

where u_i is an eigenvector of S, and λ_i is its corresponding eigenvalue. The scaling factor l_i of the i^{th} retained expert can be expressed as

$$l_i = (\lambda_i^{-1} - \beta_\Sigma)^{1/2} \Rightarrow \lambda_i \leq \frac{1}{\beta_\Sigma}. \tag{30}$$

This result resembles the solution for the scaling factors in the PPCA case [*cf* Tipping and Bishop (1999)]. However, in contrast to the PPCA solution where $l_i = (\lambda_i - \sigma^2)^{1/2}$, we notice that λ_i and σ^2 are replaced by their inverses.

Obviously, if $l_i = 0$ then u_i is arbitrary. If $l_i = 0$ we say that the direction corresponding to u_i is *discarded*, i.e. the variance in that direction is explained merely by noise. Otherwise we say that u_i is retained. All potential solutions of W may then be expressed as

$$\mathsf{W} = \mathsf{U}_m(\mathsf{D} - \beta_\Sigma \mathsf{I}_m)^{1/2}\mathsf{R}^T, \tag{31}$$

where $\mathsf{R} \subset \mathbb{R}^{m\times m}$ is an arbitrary rotation matrix, $\mathsf{U}_m = [\mathsf{u}_1\mathsf{u}_2\ldots\mathsf{u}_m] \subset \mathbb{R}^{d\times m}$ is a matrix whose columns correspond to m eigenvectors of S, and $\mathsf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_m) \subset \mathbb{R}^{m\times m}$ such that

$$d_i = \begin{cases} \lambda_i^{-1} & \text{if } \mathsf{u}_i \text{ is retained;} \\ \beta_\Sigma & \text{if } \mathsf{u}_i \text{ is discarded.} \end{cases} \tag{32}$$

A.3 Properties of the Optimal Solution

In section A.3.1 we show that for the PoGP model, the discarded eigenvalues must be adjacent within the sorted spectrum of S, and that for the maximum likelihood solution it is the smallest eigenvalues that should be retained. In section A.3.2 we show that if there are r retained eigenvectors, the log likelihood is maximized when $r = m$.

A.3.1 The Nature of the Retained Eigenvalues

Let $r \leq m$ be the number of eigenvectors of S, retained in W_{ML}. Since the SVD representation of a matrix is invariant under simultaneous permutations in the order of the left and right singular vectors (e.g. Golub and Van Loan (1997)), we can assume without any loss of generality that the first eigenvectors $\mathsf{u}_1, \ldots, \mathsf{u}_r$ are retained, and the rest of the eigenvectors $\mathsf{u}_{r+1}, \ldots, \mathsf{u}_m$ are discarded.

In order to investigate the nature of eigenvectors retained in W_{ML}, we will express the log-likelihood (3) of a Gaussian through eigenvectors and eigenvalues of its covariance matrix. From the expression for the PoGP weights (31) and the form of the model's inverse covariance (7), C_Σ^{-1} can be expressed as follows:

$$\mathsf{C}_\Sigma^{-1} \;=\; \sum_{i=1}^{r} \mathsf{u}_i \mathsf{u}_i^T \lambda_i^{-1} + \sum_{i=r+1}^{d} \mathsf{u}_i \mathsf{u}_i^T \beta_\Sigma. \tag{33}$$

Since determinants and traces can be expressed as products and summations of eigenvalues respectively, we see that

$$\ln|\mathsf{C}_\Sigma|^{-1} \;=\; \ln\left(\beta_\Sigma^{(d-r)} \prod_{i=1}^{r} \lambda_i^{-1}\right) = -\sum_{i=1}^{r} \ln(\lambda_i) + (d-r)\ln\beta_\Sigma \tag{34}$$

$$\mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S}) \;=\; r + \beta_\Sigma \sum_{i=r+1}^{d} \lambda_i. \tag{35}$$

By substituting these expressions into the form of $\mathcal{L}$ for a Gaussian (3), we get

$$\mathcal{L}(\mathsf{W}_{ML}) \;=\; -\frac{N}{2}\left[d\ln(2\pi) + \sum_{i=1}^{r}\ln(\lambda_i) - (d-r)\ln\beta_\Sigma + r + \beta_\Sigma \sum_{i=r+1}^{d} \lambda_i\right]. \tag{36}$$

Differentiating (36) with respect to β_Σ gives

$$\frac{\partial \mathcal{L}(\mathsf{W}_{ML})}{\partial \beta_\Sigma} = 0 \;\;\Rightarrow\;\; -\frac{N}{2}\left[-\frac{d-r}{\beta_\Sigma} + \sum_{i=r+1}^{d}\lambda_i\right] = 0, \tag{37}$$

which results in

$$\beta_\Sigma^{ML} = \frac{d-r}{\sum_{i=r+1}^{d}\lambda_i}. \tag{38}$$

We see that assuming non-zero noise this solution makes sense only when $d > m \geq r$, i.e. the dimension of the input space should exceed the number of experts.

Then we obtain

$$\mathcal{L}(\mathsf{W}_{ML}, \beta_\Sigma^{ML}) \;=\; -\frac{N}{2}\left[\sum_{i=1}^{r}\ln(\lambda_i) + (d-r)\ln\frac{\sum_{i=r+1}^{d}\lambda_i}{(d-r)} + r + (d-r)\right] + c \tag{39}$$

$$\;=\; -\frac{N}{2}\left[(d-r)\ln\frac{\sum_{j=r+1}^{d}\lambda_j}{(d-r)} - \sum_{j=r+1}^{d}\ln(\lambda_j)\right] + c' \tag{40}$$

where c, c' are constants, and we have used the fact that $\sum_{i=1}^{d} \ln(\lambda_i) = \ln|\mathsf{S}|$ is a constant.

Let A denote the term in the square brackets of (40). Clearly $\mathcal{L}$ is maximized with respect to the λ terms when A is minimized. We will now investigate the value of $\mathcal{L}(\mathsf{W}_{ML}, \beta_{\Sigma}^{ML})$ as different eigenvalues are discarded or retained, so as to identify the maximum likelihood solution. The expression for A is identical (up to unimportant scaling by $(d-r)$) with expression (20) in Tipping and Bishop (1999). Thus their conclusion that the retained eigenvalues should be adjacent within the spectrum of sorted eigenvalues of the sample covariance S is also also valid in our case[2]. This result, together with the constraint on the retained eigenvalues (30) and the noise parameter (38), yields

$$\forall i = 1 \ldots r, \quad \lambda_i \leq \frac{\sum_{j=r+1}^{d} \lambda_j}{d-r}, \tag{41}$$

i.e. **only the smallest eigenvalues should be retained**.

A.3.2 The Number of the Retained Eigenvalues

By analogy with the PPCA model, we may expect that retaining m eigenvectors of the sample covariance S in all m columns of $\mathsf{W} \subset \mathbb{R}^{d \times m}$ will result in a higher value of $\mathcal{L}$. To prove this, we need to show that the term in the brackets of (40) decreases as the the number of the retained eigenvectors increases.

Let the term in the brackets of (40) be denoted by B_n for $n = d - r$ discarded eigenvectors. Then

$$\begin{aligned} B_n &= \sum_{i=1}^{r} \ln(\lambda_i) + (d-r) \ln \frac{\sum_{i=r+1}^{d} \lambda_i}{(d-r)} \\ &= n \ln \bar{\hat{\lambda}}(n) - \sum_{i=1}^{n} \ln \hat{\lambda}_i + \ln|\mathsf{S}|, \end{aligned} \tag{42}$$

where $\hat{\lambda}_i \overset{def}{=} \lambda_{n+i}$ (i.e. $\hat{\lambda}_i$ is the ith discarded eigenvalue), and $\bar{\hat{\lambda}}(n)$ is the mean of n discarded eigenvalues. Again we have used the fact that $\sum_{i=1}^{d} \ln(\lambda_i) = \ln|\mathsf{S}|$ is a constant.

To show that retaining all $m = \max(r)$ eigenvectors maximises $\mathcal{L}$, we need to show that

$$B_n - B_{n+1} \leq 0. \tag{43}$$

We will simplify this inequality, and check whether it holds for all $n \in \mathbb{N}, n < d$ (i.e. with at least one vector retained). Substituting B_n and B_{n+1} into (43), we obtain

$$B_n - B_{n+1} = n \ln \bar{\hat{\lambda}}(n) - (n+1) \ln \bar{\hat{\lambda}}(n+1) + \ln \hat{\lambda}_{n+1}. \tag{44}$$

Therefore in order to prove (43) it is sufficient to show that the r.h.s. of (44) is ≤ 0.

As shown in the previous section, discarding n eigenvalues in the PoGP model corresponds to discarding n principal components of S. Since $\hat{\lambda}_{n+1}$ corresponds to the $(n+1)$'th discarded eigenvalue, then due to continuity of the discarded eigenvalues within the ordered multi-set of all the eigenvalues of S, we can see that

$$0 < \hat{\lambda}_{n+1} \leq \bar{\hat{\lambda}}(n+1) \leq \bar{\hat{\lambda}}(n). \tag{45}$$

Letting $\bar{\hat{\lambda}}(n) = z\hat{\lambda}_{n+1}$, $\bar{\hat{\lambda}}(n+1) = x\hat{\lambda}_{n+1}$, we have from (45) that $1 \leq x \leq z$. Thus

$$B_n - B_{n+1} \leq 0 \quad \Leftrightarrow \quad (n+1)\ln x - n \ln z \geq 0. \tag{46}$$

Note that

$$\bar{\hat{\lambda}}(n+1) \quad = \quad \frac{\bar{\hat{\lambda}}(n)n + \hat{\lambda}_{n+1}}{n+1} \Rightarrow x = \frac{nz+1}{n+1}. \tag{47}$$

[2]An alternative derivation of this result can be found in Appendix B.4 of Agakov (2000).

Substituting (47) into the r.h.s. of (46), we find that that

$$\Delta(z) \stackrel{def}{=} (n+1)\ln(nz+1) - n\ln z - (n+1)\ln(n+1) \geq 0 \Leftrightarrow (n+1)\ln x - n\ln z \geq 0, \tag{48}$$

where $\Delta(z)$ is a function of z defined for all $n = 1, \ldots, d-1$. It can be easily shown by means of analysis that $\forall z \geq 1,\quad \Delta(z) \geq \Delta(1) = 0$. Thus, (43) holds, and retaining all m eigenvalues of S in the weight matrix of the PoGP model leads to the same or higher log-likelihood $\mathcal{L}$ as retaining a smaller number of them.

Acknowledgements

Much of the work on this paper was carried out as part of the MSc project of FA at the Division of Informatics, University of Edinburgh. CW thanks Sam Roweis, Geoff Hinton and Zoubin Ghahramani for helpful conversations on the *rotcaf* model during visits to the Gatsby Computational Neuroscience Unit. FA gratefully acknowledges the support of the Royal Dutch Shell Group of Companies for his MSc studies in Edinburgh through a Centenary Scholarship.

References

Agakov, F. (2000). Investigations of Gaussian Products-of-Experts Models. Master's thesis, Division of Informatics, The University of Edinburgh. Available at `http://www.dai.ed.ac.uk/homes/felixa/all.ps.gz`.

Bishop, C. (1995). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory.* Wiley, New York.

Golub, G. H. and Van Loan, C. F. (1997). *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, USA, third edition.

Hinton, G. E. (1999). Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN 99)*, pages 1–6.

Magnus, J. R. and Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics.* Wiley, New York, second edition.

Møller, M. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533.

Oja, E. (1992). Principal Components, Minor Components, and Linear Neural Networks. *Neural Networks*, 5:927 – 935.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C.* Cambridge University Press, Second edition.

Proakis, J. G. and Manolakis, D. G. (1992). *Digital Signal Processing: Principles, Algorithms and Applications.* Macmillan.

Roweis, S. (1998). EM Algorithms for PCA and SPCA. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing Systems 10*, pages 626–632. MIT Press, Cambridge, MA.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components analysis. *J. Roy. Statistical Society B*, 61(3):611–622.

Williams, C. K. I. and Felderhof, S. N. (2001). Products and Sums of Tree-Structured Gaussian Processes. In *Proceedings of the ICSC Symposium on Soft Computing 2001 (SOCO 2001).*

Xu, L. and Krzyzak, A. and Oja, E. (1991). Neural Nets for Dual Subspace Pattern Recogntion Method. *International Journal of Neural Systems*, 2(3):169–184.