Investigations of Gaussian Products-of-Experts Models

Felix Agakov

MSc in Artificial Intelligence
Division of Informatics
University of Edinburgh
2000

Abstract

In 1999 Hinton introduced the Products of Experts (**PoE**) architecture for density estimation, where individual expert probabilities are multiplied and renormalized. In this work we consider specific types of PoE models where each expert is a Gaussian. Even though the resulting distributions of the products in these cases are also Gaussians, it is shown that they may lead to interesting models, such as probabilistic analogues of non-probabilistic dimensionality reduction techniques. This makes products of Gaussians potentially useful in a wide range of data analysis applications.

In first two principal parts of the investigation we consider products of constrained Gaussians (PoCG), elongated in all directions except one, and products of probabilistic principal component analyzers (PoPPCA). We prove that the PoCG gives rise to a minor component analysis solution. We also show that the product of m one-dimensional PoPPCAs where all experts have identical noise variances has a lower representational power than the m-factor PPCA.

In the third principal part of the research we investigate learning by minimizing the contrastive divergence (CD), suggested by Hinton (2000) as a significantly more efficient alternative to maximizing the log-likelihood in PoE. We consider a 1-D 1-factor linear Gaussian latent variable model and show that for the analyzed model the CD learning introduces the systematic bias by underestimating the absolute value of the gradient of the log-likelihood. However, we also analytically confirm that for the model we consider optimization of the log-likelihood and its i-step approximation lead to similar maximum-likelihood solutions.

Acknowledgements

I would first of all like to thank Dr. Chris Williams, who has introduced me to the field of Bayesian statistics and latent variable modeling, ignited my enthusiasm in probabilistic modeling, reasoning, learning, and unlearning, and helped me to undertake the procedure of supervised learning throughout my work at the project.

My deepest gratitude is to my parents, Anna, Tanya, and Sergey for their continuous support, devotion, and very interesting discussions on the nature and philosophy of intelligence.

I would also like to thank the Royal Dutch Shell Group of Companies for sponsoring my year in Edinburgh via the Centenary Scholarship, without which the vague and mostly unknown future would have definitely been significantly less artificially intelligent.

My final words of gratitude are to C.M. Bishop, G.E. Hinton, and S. Haykin for writing books that make life interesting.

Contents

1 Background **1**

1.1 Introduction 1
1.2 Structure of the Thesis 4
1.3 Latent Variable Modeling 6
1.4 Learning in Latent Variable Models 9
1.4.1 Likelihood 10
1.4.2 Analytic Methods 11
1.4.3 Iterative Methods 12
1.4.4 Methods of Non-Linear Optimization 13
1.5 Latent Variable Models 15
1.5.1 FA and PPCA 15
1.5.2 Constrained Gaussians 20
1.5.3 Mixture Models 24
1.5.4 Products of Experts 26
1.5.5 PoE *vs* Mixtures 29
1.5.6 Products of Gaussians 31
1.6 Useful Notations 33
1.7 Summary 34

2 Products of Constrained Gaussians (PoCG) **37**

2.1 Motivation 38
2.2 Analytic Parameter Estimation 38
2.2.1 Covariance in the Visible Space 39
2.2.2 ML Parameter Estimators 41
2.3 Experiments with the PoCG Model 44
2.3.1 Correspondence Between Likelihoods of Analytic and Non-Linear Optimization Solutions 45
2.3.2 Properties of the ML Solutions for the PoCG 53
2.4 Summary 55

3 Products of Probabilistic Principal Component Analysers 57
3.1 Motivation 58
3.2 Covariance in the Visible Space 59
3.2.1 A Single Expert Case 59
3.2.2 A Multiple Expert Case 60
3.3 The Problem of Modeling PPCA Data 60
3.4 Analytic Parameter Estimation 62
3.4.1 ML Parameter Estimators for Implicitly Scaled Weights 63
3.4.2 ML Parameter Estimators for Unscaled Weights 66
3.5 Experiments with the PoPPCA Model 68
3.5.1 Correspondence between Likelihoods of Analytic and SCG Solutions for Covariances Resulting in Real-Valued Weight Matrices 69
3.5.2 Modeling 1-Factor PoPPCA Data 76
3.5.3 Properties of the ML Solutions for the PoPPCA 78
3.6 Summary 80

4 CD Learning Rule in Gaussian Models 81
4.1 Motivation 82
4.2 Boltzmann Machines and PoE 82
4.2.1 The BM Overview 83
4.2.2 Boltzmann Learning 84
4.2.3 Restricted Boltzmann Machines 85
4.2.4 Gaussian Models and Boltzmann Machines 85
4.3 Learning by Minimizing Contrastive Divergence 86
4.3.1 The Problem of Estimating the Likelihood 86
4.3.2 Contrastive Divergence Learning Rule 87
4.3.3 Contrastive Divergence and Boltzmann Learning 89
4.4 Contrastive Learning in a Gaussian Model 90
4.4.1 The Model Description 90
4.4.2 Gibbs Sampling in the Gaussian Case 91
4.4.3 Estimation of the Moments for Parameter Updates . . 93
4.5 Summary 99

5 Conclusions 101
5.1 Goals 101
5.2 This Work in the Context of Previous Studies 102
5.3 Results 102
5.4 Assumptions and Limitations 104
5.5 Future Work 106

Appendices **113**

A The Fundamentals **115**
A.1 Basic Definitions 115
A.2 The Gaussian Distribution 116
A.3 Independence of the Hidden States Given the Data in PoE Models 119
A.4 Maximum Likelihood Expression for the PoE Model 121
A.5 Gibbs Sampling and MCMC 122

B Analytic Solution of the Constrained Gaussian Model **125**
B.1 Covariance of the PoCG model 125
B.1.1 Covariance in the Visible Space 125
B.1.2 Hidden Variable Representation 127
B.2 Derivation of the ML Equations for the Parameter Estimates 129
B.2.1 ML Equations Under the Assumption of Independence of W and β_0 129
B.2.2 Justification of the Simplifying Assumptions * 131
B.3 ML Solutions for the Parameter Estimates 132
B.3.1 Exact Model Covariance 133
B.3.2 Approximate Model Covariance 134
B.4 Investigation of Parameter Properties 136
B.4.1 The Nature of the Retained Eigenvalues 136
B.4.2 The Number of the Retained Eigenvalues 140

C Analytic Solution for the PoPPCA model **143**
C.1 Verification of PPCA 143
C.2 Equal Experts for Modeling 1-factor PPCA Data 146
C.3 ML Parameter Estimators for PoPPCA with Implicitly Scaled Weights 148
C.3.1 ML Equations for Implicitly Scaled Weights 148
C.3.2 ML Solutions for Implicitly Scaled Weights 149
C.3.3 Conditions on Real Weights of the PoPPCA Model . . 150
C.4 ML Parameter Estimators for PoPPCA with Unscaled Weights 152
C.4.1 Expressions of Derivatives of the Log-Likelihood 152
C.4.2 ML Equations for Unscaled Weights 155
C.4.3 ML Solutions for Parameter Estimates 159
C.4.4 Special Cases of Solutions for Parameter Estimates . . 163

D Analysis of CD Learning: Auxiliary Derivations **167**

List of Figures

1.1 A 1-factor latent variable model. 7

1.2 3-D data explained by a 1-factor PPCA. 17

1.3 A mixture of two 1-factor PPCA models in the 3-D data space. The ellipses indicate probability contours of each expert, the triangles and circles show data points generated by each expert. 18

1.4 A product of two 1-factor PPCA models in the 3-D data space. Circles show a random sample form the PoE distribution. . . . 18

1.5 Projection from the latent onto the 3-D visible space in 2-factor FA. $\lambda_1, \lambda_2, \lambda_3$ are eigenvectors of the sample covariance. W is the expert weight matrix. 19

1.6 Probability contour of a constrained Gaussian in 2D. 20

1.7 3-D data explained by a constrained Gaussian model. 22

1.8 A mixture of two constrained Gaussians in the 3D data space. 23

1.9 A product of two constrained Gaussians in the 3D data space. 23

1.10 [a] A mixture of simple parametric probabilities; [b] a mixture of latent variable models. $z \in \{z_1, \dots, z_m\}$ is a discrete index; $\acute{\mathsf{z}}$ is a vector of hidden variables of each expert in the mixture. 24

1.11 EM algorithm for a Gaussian mixture. The solid lines represent probability contours of three 2-D experts in the mixture of Gaussians model after convergence of the EM algorithm. Radii are equal to variances in the directions specified by eigenvectors of the covariances of the conditional densities $p(\mathsf{x}|z_i)$. The dashed lines show the contours of the three data-generating experts. 26
1.12 Density for a product [left] and a mixture [right] of 3 2-D Gaussians. 30
1.13 Probability surfaces of conditional densities $p(\mathsf{x}|z_i)$ in the 2D data space. Each conditional is a 1-D Gaussian. 31

2.1 Learning curves of the SCG optimization of the log-likelihood for parameters of the $(1, m)$ PoCG model in 7-dimensional data space. The number of experts m=1,2,3,6. 48
2.2 Convergence of non-singular solutions started at different points of the parameter space to the same value of $\mathcal{L}$. 50
2.3 Influence of the number of the retained eigenvalues on the value of the log-likelihood in the $(1, 19, r)$ PoCG model in the 20-D data space. 54

3.1 Learning curves of the SCG optimization of the log-likelihood for parameters of the (1,4) PoPPCA model in 7-dimensional data space. 71
3.2 Influence of the number of the retained eigenvalues on the value of the log-likelihood $\mathcal{L}$ under the $(1, 19)$ PoPPCA in the 20-D data space for the unconstrained weights. 79

4.1 Gibbs sampling. 89
4.2 Typical mean and variance of the update parameter for $k_{min} \geq 1$. 97

4.3 Typical mean and variance of the update parameter for $0 < k_{min} < 1$. 98
4.4 Typical mean and variance of the update parameter for $k_{min} < 0$. 98

A.1 A two-dimensional Gaussian distribution 117
A.2 Contour of a two-dimensional Gaussian distribution. 118

Chapter 1

Background

> The world is round and the place which may seem like the end may also be only the beginning.
>
> – Ivy Baker

1.1 Introduction

A variety of definitions of intelligence suggest that its possible measure is the ability of an expert to acquire and process information in order to solve a specific task. In highly complicated environments, an optimal solution to a problem may often require generalization over previously solved problems, observed data patterns, etc. Thus, the *ability to generalize over the observed data samples and the ability to predict states of a system in uncertain and only partially known environments* seem to form a necessary extension to the commonly accepted definition of intelligence.

In the context of this definition, one of the most important practical goals of non-symbolic Artificial Intelligence can be viewed as **creating a model**, suitable for solving various types of classification problems on the basis of observations. In its pure form, a task of classification is to map observed data onto one or more of discrete pre-defined classes according to some features of the data. Mapping between observations and real-valued spaces is generally referred to as *regression* (Bishop, 1995, p.5). Classification and regression can also be defined as *prediction* tasks.

Three most general approaches to estimating the model from observations include *supervised*, *unsupervised*, and reinforcement learning techniques, and various combinations thereof.

In supervised learning, training patterns incorporate both data points and correct outputs, and *posterior probabilities* (a probability that a given point belongs to a certain class) can be modelled directly. This may result in a relatively simple solution to classification problems. However, labelling observations with true values of the output function for each training pattern may be difficult, impossible, or expensive to achieve.

In contrast, unsupervised learning assumes that no known target outputs are associated with observed data patterns. The goal of learning is to distinguish statistical features of the given data and produce a description of the data in a form that can be suitable for solving specific tasks. For instance, classification tasks can be solved by modeling *class-conditional probabilities* (a probability that a random variable takes a certain value given that it belongs to a particular class), applying Bayesian inference rules for computing values of posterior probabilities, and constructing classification decision boundaries. If the data is observed in uncertain environments, we can assume that each data pattern occurs with a certain probability. Assuming that this probability is static in time, we can try to understand an underlying structure of the data by modeling the data-generating probability density.

If it is known a priori that the data is drawn from a relatively simple distribution, we can assume that the density can be estimated by simple parametric models and can be expressed analytically prior to observing the data. In this case, the observed data is used for adjusting parameters of the distribution.

However, if the data is drawn from a highly complicated distribution, it is very unlikely that a simple parametric model (*expert*) can fit the data well. Two most commonly used approaches for estimating densities of data-generating distributions are based on *non-parametric* and *semi-parametric* (Bishop, 1995, p.60) **mixture** techniques.

Non-parametric density estimation techniques can model an arbitrary smooth distribution. However, in general the modeled density converges to the true density when the sample size approaches infinity. The number of variables in the model grows directly with the sample size (Bishop, 1995), (Duda and Hart, 1973), what may cause a very slow evaluation of posterior densities for each observed vector.

In semi-parametric mixture models, the data-generating probability is estimated as a weighted summation of simpler expert probabilities. There are well-established techniques for estimating parameters of mixture models, what makes them an attractive tool for density estimation. However, the weighted summation of probabilities may generally result in excessively smooth densities in high-dimensional spaces. As indicated in Hinton (2000), the resulting distribution of the mixture model cannot be sharper than each particular expert, and every expert must be "broadly tuned" to explain the data in a multi-dimensional space. Thus, a mixture model may be rather inefficient for fitting local multi-dimensional data patterns.

In 1999 Hinton introduced a principally new Product of Experts (**PoE**) model for combining expert probabilities, where a data-generating distribution is estimated as a normalized multiplication of probabilities of individual experts (Hinton, 1999). In this case, each individual expert may assign high probabilities to an irrelevant region of the data space, as long as the probability of the region is small under other experts. The PoE model is believed to produce a significantly better fit to local patterns in multi-dimensional space.

This thesis will focus on investigation of the PoE architecture for the case when each expert is a Gaussian. In this case the resulting PoE model is also a Gaussian, but as indicated in (Williams and Felderhof, 2000), one of the possible practical applications of the products of Gaussians models could be a simplification of calculations with very large matrices that could otherwise be required. Moreover, if we could find a probabilistic equivalent of a non-probabilistic data analysis technique (like minor component analysis), we could compare it with other density-estimation methods by comparing likelihoods or Bayesian MAP estimators. We could also consider mixtures of such methods (what would correspond to mixtures of products of experts). Other important advantages of probabilistic models are discussed in Tipping and Bishop (1997) and Bishop and Tipping (1999). In addition, investigation of contrastive divergence learning in a Gaussian model may be performed analytically and is interesting from the theoretical point of view.

1.2 Structure of the Thesis

Chapter 1 introduces the background necessary for understanding the rest of the thesis. In the following sections we will briefly discuss general principles of latent variable modeling for analysis of data distributions. We will briefly describe various approaches to model parameters estimation, based on analytic, iterative, and non-linear techniques. Chapter 1 ends with examples of several types of latent variable models, namely Probabilistic Principal Component Analysis models (Tipping and Bishop, 1997), constrained Gaussians with "disk"-shaped probability contours, general mixture models, and Products of Experts (Hinton, 1999). We will illustrate differences between mixture models and PoE in a little more detail, and discuss the specifics of the Product of Gaussians models. The description is given on a very high level and assumes little familiarity with the investigated concepts. Appendix A is an introductory supplement of the chapter, introducing some definitions used throughout the thesis, briefly describing multi-variate Gaussian distributions, and deriving the maximum-likelihood expression for the PoE.

Chapter 2 presents a more detailed analysis of the products of constrained Gaussians (PoCG), where each Gaussian is elongated in all directions but one. We will show that parameters of the PoCG can be found analytically and lead to minor component analysis. We will consider representation of the model in the latent space, derive analytic solution for the model, and discuss its properties. Finally, we will illustrate the results by experiments and compare the analytic solution with the ones obtained by applying non-linear optimization techniques. The derivations confirming the analytical results of the chapter are shown in Appendix B.

In Chapter 3 we investigate products of PPCA models. First we consider a case when an m-expert PoPPCA with a single retained expert cannot model the same probability density as a 1-factor PPCA model. This shows limitations of the PoPPCA architecture and its lower representational power compared with that of the m-factor PPCA at least for some covariance matrices. Then we discuss two different approaches to ML parameter estimation in PoPPCA models, and show that the general analytic solution for the model parameters appears to be difficult due to the scaling factor on experts' weight, unless we allow experts of the PoPPCA model to have imaginary weights. We also determine ranges of eigenvalues of the sample covariance matrix for which a PoPPCA model leads to the same results as a PPCA, and confirm the results experimentally. The corresponding derivations and verification of some special cases are listed in Appendix C.

Chapter 4 investigates contrastive divergence learning in Gaussian models. First we briefly discuss learning in restricted Boltzmann machines and outline contrastive divergence learning in PoE (Hinton, 2000). Then we briefly show correspondence of a product of Gaussian experts to a restricted Boltzmann machine and examine analytically contrastive divergence learning for a simple 1-D 1-factor Gaussian model. We show that for the considered model maximizing the log-likelihood and its approximation by i-step Gibbs sampling in both cases leads to the same ML solutions for the model parameters. Appendix D shows some auxiliary derivations.

1.3 Latent Variable Modeling

Very often entities that we are most interested in cannot be inferred from observations deterministically. Consider, for example, an Asian character written in an ancient manuscript. If all what we know about it is a number of strokes and a type of the rice paper it is written on, the perceived information may be insufficient for determining whether it corresponds to Japanese, Korean, or Vietnamese syllable (or, perhaps, a complete paragraph). In another case, if our knowledge of a country's president is confined in his KGB past and remarks in yesterday's newspapers, it may be rather difficult to see underlying factors that will influence a country's policy for the next several years.

In many cases we cannot build a complete description of the observations because there is some data which is hidden from us. It cannot be perceived or derived from observations directly, even though it may often be the cause of the observations. It may be missing or unknown. In some cases, we (or experts) can make a guess about real underlying causes of observations. Alternatively, we can apply probabilistic inference techniques for estimation of posterior distributions (i.e. probability of the *cause* given the *effect*).

We can define a *latent variable* as an auxiliary tool for explaining patterns in observations. As Radford Neal states,

> Doctors have invented 'diseases' to explain commonalities of symp-

toms seen in patients (Neal, 2000).

By learning a mapping from diseases to symptoms, we can model a probability of a disease given a symptom.

Thus, a very important task of the latent variable modeling can be defined as creating representation of the observed data vectors $\{\mathbf{x}\}$ through the hidden data vectors $\{\mathbf{z}\}$. The hidden variables $\{\mathbf{z}\}$ capture certain properties of the visible variables $\{\mathbf{x}\}$. This properties may be important for explaining the observations and creating a *generative* model of the data (i.e. a model that generates the data).

By analogy with Factor Analysis (Bartholomew, 1987), the dimension of the latent space will be referred to as the number of *factors*.

The following discussion relates to Bishop and Tipping (1999) and Tipping and Bishop (1997), who distinguish three stages of latent variable modeling.

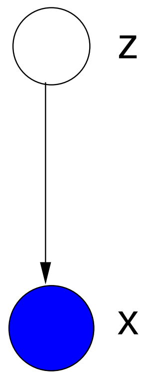

Figure 1.1: A 1-factor latent variable model.

1. **Estimation of joint probabilities:**

   Initially, the joint distribution of the *complete* model can be expressed through conditional probabilities (probability of the data given a latent variable) as

$$p(\mathsf{x},\mathsf{z}) = p(\mathsf{z})p(\mathsf{x}|\mathsf{z}) = p(\mathsf{z})\prod_{i=1}^{d} p(x_i|\mathsf{z}), \tag{1.1}$$

where d is the dimension of the data space.

The key assumption is that the **observations are independent given the latent variables**. It is further noticed that (1.1) may be represented in the framework of Bayesian nets, which are thoroughly described in Heckerman (1998), Russell and Norvig (1995), and many other sources.

2. **Estimation of posterior probabilities:**

   Representation of the observed data through the hidden variables can be expressed as

$$\mathsf{x} = \mathsf{f}(\mathsf{z},\Theta) + \epsilon \subset \mathcal{R}^d, \tag{1.2}$$

   where function f defines a mapping from the hidden to the visible data space, Θ is a vector of parameters, and ϵ is a noise, independent of the latent variables z.

   In relatively simple cases when the mapping is linear and the hidden variables have relatively simple parametric forms, expression (1.2) can be used for analytic estimation of the posterior densities $P(\mathsf{z}|\mathsf{x})$. This is the case in linear *Factor Analysis* (Bartholomew, 1987), PPCA (Tipping and Bishop, 1997), *Hidden Markov Models*, and *Kalman filters*. If the mapping from the latent space is non-linear, like in the GTM model of Bishop, Svensen and Williams (1998), posterior distributions cannot be always expressed analytically. In this case, they may be estimated by applying Monte-Carlo techniques as described, for example, in Neal (1993) and Mackay (1998). An overview of latent variable models is presented in Williams (1999b).

3. **Estimation of marginal data distributions:**

   Finally, the data distribution may be estimated as

$$p(\mathsf{x}) = \int p(\mathsf{x}|\mathsf{z})p(\mathsf{z})\mathrm{d}\mathsf{z}. \tag{1.3}$$

   $p(\mathsf{z})$ is referred to as *prior* probability on the latent variables.

In this thesis we will focus on two particular linear mappings from the latent to the visible space.

We can easily see that if latent variables $\{\mathsf{z}\}$ correspond to an unknown category of the data, modeling posterior probabilities $P(\mathsf{z}|\mathsf{x})$ may help us solve classification tasks.

1.4 Learning in Latent Variable Models

Learning in latent variable models may be defined as estimation of parameters from observations. In section 1.1 we have briefly discussed supervised and unsupervised approaches to learning, which are also applicable to learning latent variable representations of the data.

The major disadvantage of supervised learning is a general problem in collecting the labeled data. Moreover, modeling $p(\mathsf{z}|\mathsf{x})$ may results in difficulties in dealing with unclassified and missing data in classification problems.

On the other hand, unsupervised learning techniques do not require any labeled data, and the data-generating distribution $p(\mathsf{x})$ may be obtained by marginalizing over hidden (missing) entities (1.3). However, we will often need to make an assumption about a prior distribution of the latent variables $p(\mathsf{z})$. Moreover, both creation of a model description according to a sample and sampling from a highly-complicated distribution may be rather complicated.

When we speak about learning, we usually mean optimization of some criterion. In fact, the major difference between supervised and unsupervised learning techniques is the definition of this criterion, ant this determines specific features of both approaches. In supervised learning, the criterion to be minimized is a measure of the difference between target patterns and model's outputs (e.g. the sum-of-squares error, the cross-entropy, etc.)

One possible optimization criterion in unsupervised learning is a measure of likeliness that a system with certain parameters Θ[1] generates an observed data sample (*maximum likelihood*). Another possible criterion, incorporating the priors on model parameters, is a *maximum a posteriori (MAP)* measure. Differences between these measures are determined by Bayesian and frequentist approaches to defining probability and discussed in length in numerous sources, e.g. (Bishop, 1995, pp. 40-46); (Tipping, 1999, pp. 47-53).

PoE and mixture models are examples of unsupervised learning techniques, and in the following sections we will focus on parameter estimation maximizing likelihood.

1.4.1 Likelihood

Maximum likelihood (ML) parameter estimators Θ_{ML} maximize the probability of generating the data set X:

$$\Theta_{ML} = argmax_{\Theta} p(\mathsf{X}|\Theta). \tag{1.4}$$

With the assumption of *independent identically distributed* observed variables, likelihood of the parameters Θ is defined as

$$\mathbf{L}(\Theta) = p(\mathsf{X}|\Theta) = \prod_{n=1}^{N} p(\mathsf{x}^{(n)}|\Theta), \tag{1.5}$$

where N is the sample size and $\mathsf{x}^{(n)}$ is a sample point. The likelihood $\mathbf{L}$ is assumed to be a function of parameters Θ for a fixed sample of observations.

For distributions whose densities contain exponential terms, using a $\log \mathbf{L}$ error criterion may be more convenient (this criterion is valid due to the

[1]Here and further we shall focus on parametric and semi-parametric models.

monotonic growth of the log function). For example, this is the case for a very commonly used Gaussian distribution[2]. For a Gaussian, the vector of parameters Θ consists of the mean μ and the covariance C. The log-likelihood can be re-written as

$$\begin{aligned}\mathcal{L}(\mu, \mathsf{C}) = \log \mathbf{L}(\mu, \mathsf{C}) &= \sum_{n=1}^{N} \ln p(\mathsf{x}^{(n)}|\mu, \mathsf{C}) \\ &= -\frac{N}{2} d \ln 2\pi - \frac{N}{2} \ln |\mathsf{C}| - \frac{N}{2} \mathrm{tr}[\mathsf{C}^{-1}\mathsf{S}], \quad (1.6)\end{aligned}$$

where

$$\mathsf{S} = \frac{1}{N} \sum_{n=1}^{N} (\mathsf{x}^{(n)} - \mu)(\mathsf{x}^{(n)} - \mu)^T \quad (1.7)$$

is the *sample covariance* matrix with the sample mean equal to the data mean μ; C is the covariance of the model in the data space, and d is the dimension of the data space.[3]

In critical points the ML estimator for Θ of a multi-variate Gaussian distribution results in zero of the derivative, i.e.

$$\frac{\partial \mathcal{L}}{\partial \Theta} = -\frac{N}{2} \frac{\partial \ln |\mathsf{C}|}{\partial \Theta} - \frac{N}{2} \frac{\partial \mathrm{tr}[\mathsf{C}^{-1}\mathsf{S}]}{\partial \Theta} = 0. \quad (1.8)$$

1.4.2 Analytic Methods

When it is possible, the idea of analytic methods is to find such values of parameters $\Theta^\star$ that the derivative of the optimization measure $\partial\mathcal{L}/\partial\Theta|_{\Theta^\star} = 0$, and verify that determined values of the parameters in fact correspond to a maximum of the log-likelihood $\mathcal{L}$. For a Gaussian distribution it will involve an analytic solution of equation (1.8).

[2] Gaussian distributions are briefly described in Appendix A.

[3] In deriving (1.6) we have used the fact, that for a matrix X and vectors a and b their trace has the property $\mathrm{tr}(\mathsf{a}^{\mathrm{T}}\mathsf{X}\mathsf{b}) = \mathrm{tr}(\mathsf{X}\mathsf{b}\mathsf{a}^T) = \mathsf{a}^{\mathrm{T}}\mathsf{X}\mathsf{b}$. This representation will simplify calculations in the following chapters. Many useful properties of traces and determinants that we will often use are described online in Brookes (2000).

Evaluation of parameters from analytic expressions is often the fastest way of parameter estimation. However, conceptually this technique is most difficult and problem-dependent. In many cases, latent variables make analytic estimation of parameters impossible.

1.4.3 Iterative Methods

In some cases we can find an analytic solution for parameters, maximizing log-likelihood $\mathcal{L}$. However if we are dealing with a latent variable model, evaluation of these parameters from the analytic expression may be problematic or impossible. Optimal parameters Θ of the model may depend on unknown values of the latent variables, which may themselves depend on Θ.

A broad family of *Expectation-Maximization (EM)* algorithms (Dempster et al., 1977) offers an efficient approach to maximizing log-likelihood for incomplete data sets, where an analytic solution could have been possible had the data been complete. Initially, the algorithm has been designed for Hidden Markov Models, but later developed for many other types of latent variable models and their mixtures.

A very comprehensive discussion of the EM algorithm is offered in (Tipping, 1999) and (Williams, 1999a). Here we will very briefly outline its most important steps.

Two main steps of EM algorithms can be defined as follows:

1. **Expectation step**:

   The idea is to compute the *expected complete* data log-likelihood, [i.e. expectation of a log of the joint distribution $P(\mathsf{x}, \mathsf{z}|\Theta)$ under a model with parameters Θ)] with respect to the posterior distribution with the *old* values of the parameters [i.e. $P(\mathsf{z}|\mathsf{x}, \Theta^{old})$]. By doing this, we evaluate likelihood of the complete model on the basis of the current knowledge of posterior distributions with the fixed values of Θ^{old}.

$$Q(\Theta|\Theta^{old}) = \sum_{\mathsf{z}} p(\mathsf{z}|\mathsf{x}, \Theta^{old}) \log p(\mathsf{z}, \mathsf{x}|\Theta), \tag{1.9}$$

where $Q(\Theta|\Theta^{old})$ denotes expected complete data log-likelihood.

2. **Maximization step**:

   The idea here is to find model parameters Θ maximizing the expected value of the data log-likelihood:

$$\Theta = argmax_{\Theta} \left[Q(\Theta|\Theta^{old}) \right]. \tag{1.10}$$

In this thesis we will focus mainly on the analytic and non-linear approaches to parameter estimation. It will be shown that, for instance, for the PoCG model solutions can be found analytically. Comparison with non-linear techniques verifies correctness of the analytical derivation and shows interesting patterns in convergence of non-linear algorithms. For models with a known analytical solution, derivation of the slower EM algorithm is probably not very useful. However, it can be considered for learning in mixtures of products of experts.

1.4.4 Methods of Non-Linear Optimization

The third general approach to parameter estimation discussed here determines a class of methods of non-linear optimization, based on gradient descent and higher-order optimization techniques. These are probably among the most general approaches to unconstrained numerical optimization, widely used, for example, in artificial neural networks. Like EM-algorithms, they can converge to a local optima of parameter values, and are generally relatively slow.

However, sometimes numerical optimization is the only plausible approach to parameter estimation. Consider once again the equation for parameters

$$\frac{\partial \mathcal{L}}{\partial \Theta} = 0.$$

Analytical expressions of derivatives can be found relatively easily, but the resulting equations may not have analytic solutions for parameters either for the reason of the hidden data, or because they are transcendental. Non-linear optimization techniques do not require explicit analytic expression for the parameters, what makes them significantly more general than the first two of the discussed methods.

The idea of non-linear optimization approaches is to numerically update parameters in order to minimize the error measure. If $\Delta\Theta$ is a parameter update, the Taylor expansion of the error function $E(\Theta)$ is expressed as

$$E(\Theta + \Delta\Theta) = E(\Theta) + \mathsf{g}^T \Delta\Theta + \frac{1}{2}\Delta\Theta^T \mathsf{H} \Delta\Theta + o(\Delta\Theta^3), \tag{1.11}$$

where

$$\mathsf{g} = \frac{\partial E(\Theta)}{\partial \Theta} \text{ – is the gradient} \tag{1.12}$$

$$\mathsf{H} = \frac{\partial^2 E(\Theta)}{\partial \Theta^2} \text{ – is the Hessian matrix.} \tag{1.13}$$

Applying matrix differentiation rules described, for example, in Minka (1999), we can find that the analytical solution for the parameter update $\Delta\Theta$ results in the "Newton step" update

$$\Delta\Theta = -\mathsf{H}^{-1}\mathsf{g}. \tag{1.14}$$

Calculation of H^{-1} may be very computationally intensive, but *conjugate gradient* and *Quasi-Newton* methods minimize the error function without computing the inverse of the Hessian. These higher-order optimization techniques generally result in a much better convergence than the first-order gradient descent. We will apply the scaled conjugate gradient method for estimating parameters of the PoCG and PoPPCA architectures in the following chapters.

Non-linear optimization techniques are described in multiple sources on optimization and artificial neural nets. Haykin (1999) as well as Reed and Marks (1999) offer very comprehensive discussions.

1.5 Latent Variable Models

Having considered various approaches to parameter estimation, here we will discuss constrained Gaussian and probabilistic principal component analysis models, whose products are investigated in chapters 2 and 3 of the suggested work. We will also briefly describe mixture models, and illustrate their differences from the PoE models. Finally, we will introduce the products of Gaussians and describe some of their most important features.

1.5.1 FA and PPCA

Probabilistic principal component analysis (Tipping and Bishop, 1997) is a special case of a linear Factor Analysis (FA) model with a specific structure of the noise covariance matrix. In their paper, Tipping and Bishop showed that the principal component analysis can be expressed in terms of a probability model, what leads to a number of important advantages over its non-probabilistic analogue.

In linear FA, mapping from the hidden to the visible space is linear, and (1.2) can be re-written as

$$\mathsf{x} = \mathsf{W}\mathsf{z} + \mu + \epsilon, \tag{1.15}$$

where $\mathsf{z} \subset \mathcal{R}^m \sim N(0, \mathsf{I}_m)$ is the vector of latent variables, $\mathsf{W} \subset \mathcal{R}^{d\times m}$ is the *factor loading* matrix, $\mu \subset \mathcal{R}^d$ is the data mean, and $\epsilon \sim N(0, \Psi)$ is the z-independent noise term with the diagonal covariance Ψ. We use m and d to denote the dimensions of the hidden and the visible spaces respectively.

Notice that due to independence of z and ϵ, parameters of the marginal and conditional probability densities[4] in factor analysis can be expressed as

$$
\begin{aligned}
\mu_{\mathsf{x}} &= E[\mathsf{x}] = \mu && (1.16)\\
\mathsf{C}_{\mathsf{x}} &= E[(\mathsf{x} - \mu_{\mathsf{x}})(\mathsf{x} - \mu_{\mathsf{x}})^T] \\
&= \mathsf{W}E[\mathsf{z}\mathsf{z}^T]\mathsf{W}^T + E[\epsilon\epsilon^T] = \mathsf{W}\mathsf{W}^T + \Psi, && (1.17)
\end{aligned}
$$

i.e. $p(\mathsf{x}) \sim N(\mu, \mathsf{W}\mathsf{W}^T + \Psi)$. Similarly, $p(\mathsf{x}|\mathsf{z}) \sim N(\mathsf{W}\mathsf{z} + \mu, \Psi)$.

In PPCA, the noise is assumed to be a spherical Gaussian with the covariance $\sigma^2\mathsf{I}_d$. Equation (1.17) for the sample covariance matrix may then be re-written as

$$\mathsf{C} = \mathsf{W}\mathsf{W}^T + \sigma^2\mathsf{I}_d \subset \mathcal{R}^{d\times d}, \qquad (1.18)$$

and $p(\mathsf{x}) \sim N(\mu, \mathsf{W}\mathsf{W}^T + \sigma^2\mathsf{I}_d)$.

Tipping and Bishop (1997) showed that the the weights defining the mapping in this case can be found analytically and correspond to scaled and rotated principal eigenvectors of the sample covariance matrix:

$$\mathsf{W}_{ML} = \mathsf{U}_S(\Lambda_S - \sigma^2\mathsf{I}_d)^{1/2}\mathsf{R} \subset \mathcal{R}^{d\times m}, \qquad (1.19)$$

where $\mathsf{U}_S \subset \mathcal{R}^{d\times m}$ is the orthogonal matrix of m principal eigenvectors of the sample covariance S, Λ_S is the matrix of the corresponding eigenvalues, and R is an arbitrary rotation matrix.

The variance of the noise term in this case is given as the mean of $d - m$

[4]See Appendix A.1 for the definitions and Appendix A.2 for description of the Gaussian distribution.

smallest eigenvalues of the sample covariance matrix:

$$\sigma^2 = \frac{1}{d-m} \sum_{i=m+1}^{d} \lambda_i, \tag{1.20}$$

The singular values of the model covariance C have the form of

$$[\lambda_1, \lambda_2, \ldots, \lambda_m, \sigma^2 \ldots \sigma^2] \subset \mathcal{R}^{d \times d}.$$

If $\lambda_1 > \lambda_2 = \lambda_3 = \ldots \lambda_d$, then the weight matrix has a single non-zero weight vector, which determines variance in the direction of the first principal component. Variance in the other directions is explained by the noise term. In this case, we can say that the data is modeled by a 1-factor PPCA.

An example of this situation in the 3-D data space is illustrated on figure 1.2. Triangles indicate sample points from the distribution defined by a 1-factor PPCA model. The ellipses show diametral cuts of probability contours of a Gaussian with the modeled covariance C.

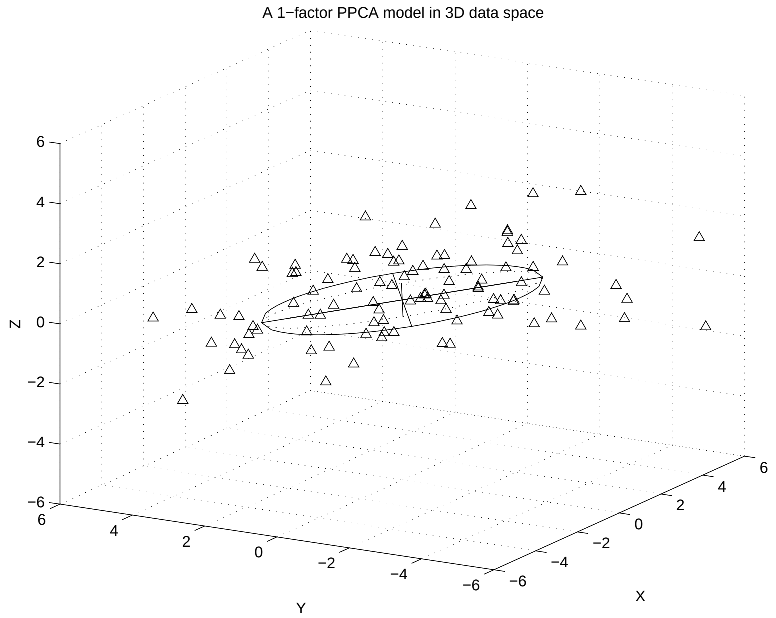

Figure 1.2: 3-D data explained by a 1-factor PPCA.

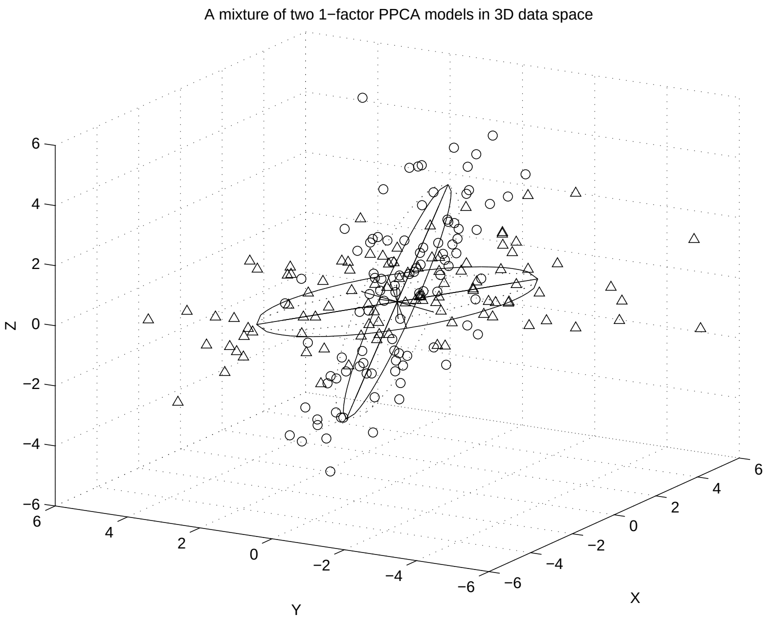

Figure 1.3: A mixture of two 1-factor PPCA models in the 3-D data space. The ellipses indicate probability contours of each expert, the triangles and circles show data points generated by each expert.

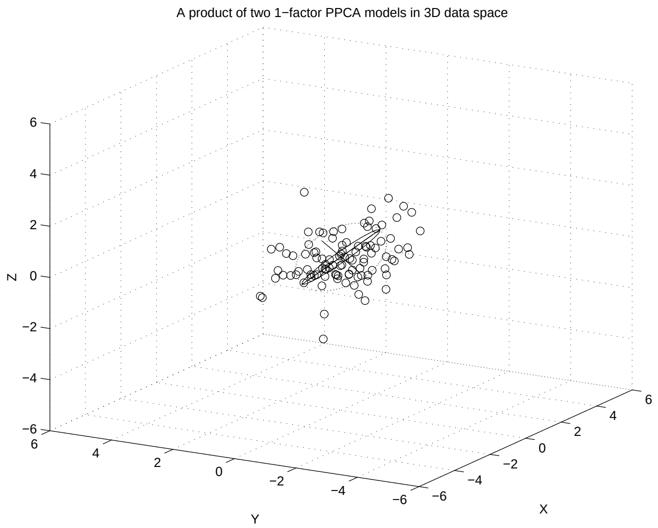

Figure 1.4: A product of two 1-factor PPCA models in the 3-D data space. Circles show a random sample form the PoE distribution.

Figures 1.3 and 1.4 illustrate a mixture and a product of two 1-factor PPCAs with identical means. Circles and triangles on figure 1.3 show the data points generated by each expert, and their combination corresponds to a sample from the resulting mixture distribution[5]. Circles on figure 1.4 represent a random sample from the PoE distribution for the same experts. We see that the sample from the PoE is significantly more localized than the sample from the mixture distribution.

Figure 1.5 gives a geometric interpretation of the FA/PPCA model (it was inspired by the discussion of the PPCA in Moerland (1998)). A spherical Gaussian is projected from the lower-dimensional latent into a higher-dimensional visible space, where it is rotated and scaled in such a way that the axis of the resulting hyper-ellipsoid correspond to principal components of the sample covariance matrix. Then it is distorted by uncorrelated Gaussian noise (spherical Gaussian for PPCA).

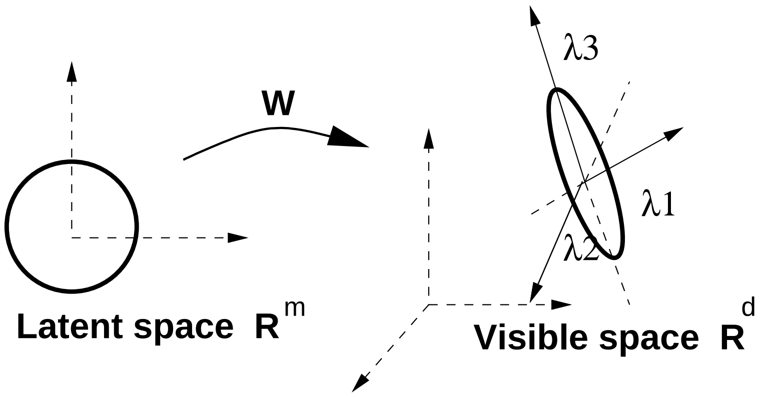

Figure 1.5: Projection from the latent onto the 3-D visible space in 2-factor FA. $\lambda_1, \lambda_2, \lambda_3$ are eigenvectors of the sample covariance. W is the expert weight matrix.

We will discuss products of PoPPCAs in Chapter 3 and Appendix C.

[5] Here we are assuming equal mixing coefficients. In case of different mixing coefficients, the resulting distribution of the model will be more influenced by one or the other expert.

1.5.2 Constrained Gaussians

Here we will consider a special case of a constrained Gaussian distribution, whose probability contours are equally elongated in the directions of eigenvectors of its covariance except one. By referring to a *constrained Gaussian*, we will always mean this particular constraint.

Consider an eigenvalue decomposition of the covariance matrix of a constrained Gaussian distribution in $\mathcal{R}^d$:

$$\mathsf{C} = \sum_{i=1}^{d-1} \mathsf{u}_i \mathsf{u}_i^T \sigma_0^2 + \hat{\mathsf{w}}\hat{\mathsf{w}}^T \sigma_{\hat{w}}^2, \tag{1.21}$$

where C is the covariance, $\mathsf{u}_1, \mathsf{u}_2, \ldots, \mathsf{u}_{d-1}, \hat{\mathsf{w}}$ are its eigenvectors, $\sigma_0^2, \ldots, \sigma_0^2, \sigma_{\hat{w}}^2$ are the corresponding eigenvalues, such that

$$\sigma_0^2 > \sigma_{\hat{w}}^2. \tag{1.22}$$

We will call a model with these properties a *constrained Gaussian*.

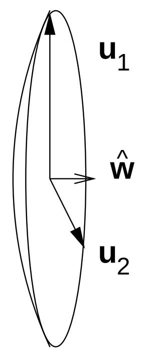

Figure 1.6: Probability contour of a constrained Gaussian in 2D.

Since C is a symmetric matrix, its eigenvectors form an orthogonal basis. Geometrically C defines probability contours resembling d-dimensional "hyper-disks". Eigenvectors $\mathsf{u}_1, \ldots, \mathsf{u}_{d-1}$ specify directions of elongation, while $\hat{\mathsf{w}}$ specifies the direction of the lower variance.

As we shall see in the following sections, the inverse covariance in this case can be expressed as

$$\mathsf{C}^{-1} = \beta_0 \mathsf{I}_d + \mathsf{w}\mathsf{w}^T, \tag{1.23}$$

$$\mathsf{w} = \hat{\mathsf{w}}\sqrt{\beta_{\hat{w}} - \beta_0}, \tag{1.24}$$

where $\beta_0, \beta_{\hat{w}}$ are inverse variances

$$\beta_0 = \sigma_0^{-2}, \ \beta_{\hat{w}} = \sigma_{\hat{w}}^{-2}. \tag{1.25}$$

Thus, the covariance of a constrained Gaussian is uniquely defined by a model with two parameters: the variance in the direction of elongation σ_0^2, and a *weight* vector w, collinear with the direction of shrinkage $\hat{\mathsf{w}}$. Note that the constraint considerations (1.22) imply $\beta_0 < \beta_{\hat{w}}$. Thus, the square root in (1.24) is positive and results in real-valued weights.

We see that the analytical form of the expression (1.23) for the inverse covariance C^{-1} in the constrained Gaussian case resembles the PPCA expression (1.18) for C, but is expressed through the inverse of the noise variance β_0. By analogy with factor analysis, we will call β_0 the [inverse] noise term. We will sometimes refer to the inverse factor analysis as the *rotcaf* model.

Considering inversion of the covariance matrices, we can make a speculative guess that wherever PPCA weights correspond to rotated and scaled *principal* eigenvectors of the sample covariance S, weights of a PoCG model correspond to rotated and scaled *minor* eigenvectors of S (possibly, with a different scaling factor). In chapter 2 we show that the assumption is indeed valid.

Figure 1.7 shows a sample from a 3-D data set, which is explained by a constrained Gaussian with a single non-zero weight vector explaining variance in the direction of shrinkage. Variance in the directions of elongation is expressed by the noise term. The sampled points are lying in approximately the same plain.

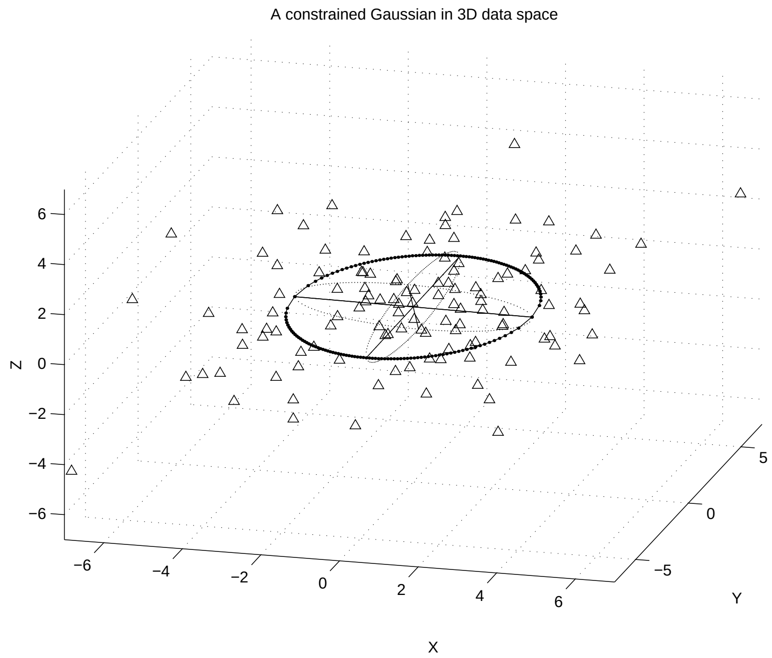

Figure 1.7: 3-D data explained by a constrained Gaussian model.

Figure 1.8 shows a mixture of two constrained Gaussians with identical means in the 3-D space. Ellipses represent diametral cross-sections of the disk-shaped probability contours with radii equal to variances in these directions. Dots and triangles indicate a random sample from each expert. The product of these experts and a sample from it are shown on figure 1.9.

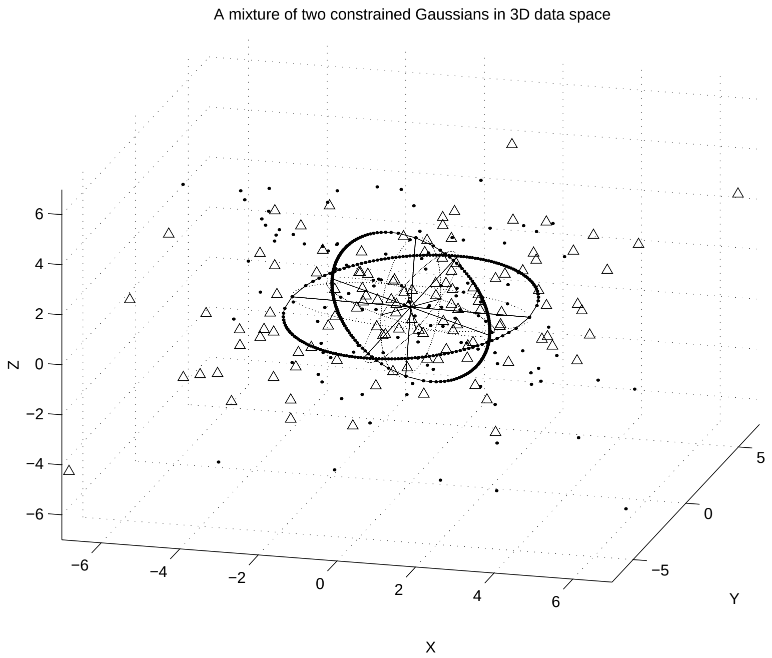

Figure 1.8: A mixture of two constrained Gaussians in the 3D data space.

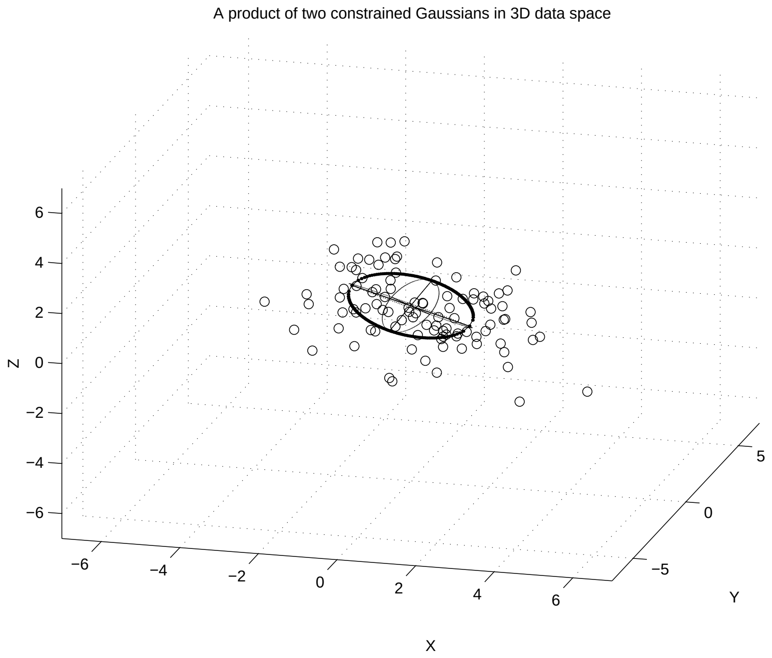

Figure 1.9: A product of two constrained Gaussians in the 3D data space.

1.5.3 Mixture Models

In cases when the data distribution is complicated and cannot be modeled by a single parametric expert (i.e. a single PPCA or a constrained Gaussian), the data density can be expressed as a superposition of simpler probability densities. Throughout the thesis, we will refer to these simpler densities as *experts*.

Note that in general we might not know posterior distributions. Given a data point, we cannot always say what expert has generated it (or, in classification tasks, what class it belongs to). Thus, we may speak about a discrete hidden variable $z \in \{z_1 \ z_2 \ \ldots \ z_m\}$, defining an index of a data-generating expert[6].

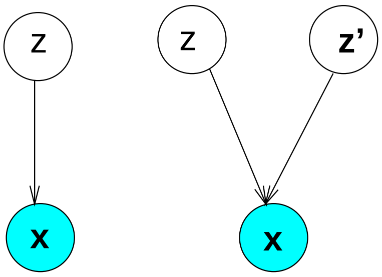

Figure 1.10: [a] A mixture of simple parametric probabilities; [b] a mixture of latent variable models. $z \in \{z_1, \ldots, z_m\}$ is a discrete index; $\acute{\mathsf{z}}$ is a vector of hidden variables of each expert in the mixture.

In general, each expert z_i may itself be a latent variable model, in which case it is necessary to distinguish between z and a vector of its hidden variables $\acute{\mathsf{z}}_i$ (see figure 1.10). These hidden variables may determine parameter values of the conditional densities, and the marginal density can be expressed

[6]We will write z_i to refer to the i^{th} expert of the mixture.

as

$$p(\mathsf{x}) = \sum_{i=1}^{m} P(z_i) \int p(\mathsf{x}|\acute{\mathsf{z}}, z_i) p(\acute{\mathsf{z}}|z_i) \mathrm{d}\acute{\mathsf{z}}. \tag{1.26}$$

(see p.3 in Ghahramani and Hinton (1996) for discussion of a mixture of factor analyzers). However, we shall limit our introduction of the mixture models to the discussion of Gaussian mixtures, where only indices z are hidden.

In *mixture models* (Titterington and Makov, 1985) with simple experts, the probability $p(\mathsf{x})$ of a data point x is expressed as a summation of conditional probabilities $p(\mathsf{x}|z_i, \Theta_i)$, weighted by the discrete *prior* expert probability $P(z_i)$:

$$p(\mathsf{x}) = \sum_{i=1}^{m} p(\mathsf{x}|z_i, \Theta_i) P(z_i). \tag{1.27}$$

Priors on components of z are also to as *mixing coefficients*.

In a generative model, the prior $P(z_i)$ defines the probability of selecting the i^{th} expert. Each data point is assumed to be generated by one of the experts, so the priors and the posteriors sum up to one:

$$\sum_{i=1}^{m} P(z_i) = 1, \ \sum_{i=1}^{m} P(z_i|\mathsf{x}) = 1. \tag{1.28}$$

Moreover

$$\begin{aligned} p(\mathsf{x}|z_i)P(z_i) &= P(z_i|\mathsf{x})p(\mathsf{x}) \Rightarrow \\ P(z_i) \int p(\mathsf{x}|z_i)\mathrm{d}\mathsf{x} &= P(z_i) \Rightarrow \\ \int p(\mathsf{x}|z_i)\mathrm{d}\mathsf{x} &= 1. \end{aligned} \tag{1.29}$$

This implies a direct mapping between $p(\mathsf{x}|z_i)$ and class-conditional densities, what makes mixture models a convenient tool for solving classification tasks. More discussion of this can be found in Bishop (1995) and Tipping (1999).

Figure 1.11 illustrates convergence of the EM algorithm for a mixture of three Gaussians in the 2-D data space. We see that the algorithm has converged to a local optimum.

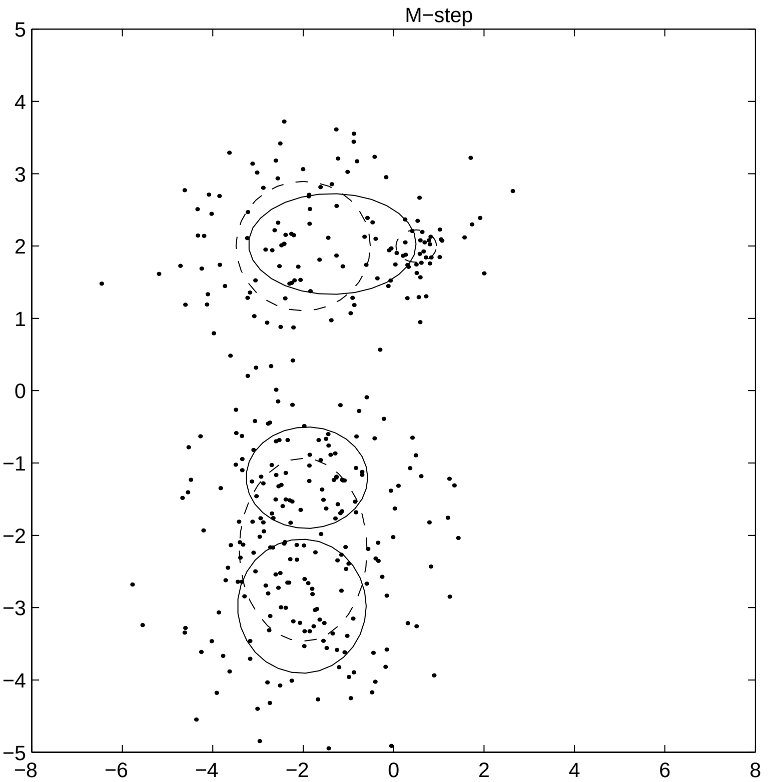

Figure 1.11: EM algorithm for a Gaussian mixture. The solid lines represent probability contours of three 2-D experts in the mixture of Gaussians model after convergence of the EM algorithm. Radii are equal to variances in the directions specified by eigenvectors of the covariances of the conditional densities $p(\mathsf{x}|z_i)$. The dashed lines show the contours of the three data-generating experts.

1.5.4 Products of Experts

Mixture models provide a powerful technique for estimation of complex probabilities. However, in some cases the weighted summation of conditional probabilities may result in excessively smooth marginal densities. This is often the case for high-dimensional spaces. Each parametric expert generalizes over observed examples, and may assign high probabilities to irrelevant parts of the data space (see figure 1.11). When conditional probabilities $p(\mathsf{x}|z_i)$ are weighted by positive mixing coefficients $P(z_i)$ and summed, the resulting probability of a point will often be vaguer than the probability of a point under each individual expert. For fitting localized patterns of multi-dimensional data, this may be a rather undesirable phenomena.

Recently Hinton (1999) introduced the products of experts (PoE) model – a principally new way of combining expert probabilities. This method is potentially very efficient for fitting local constraints of multi-dimensional data. The probability density of a point under a PoE model $p(\mathsf{x}|\Theta)$ is modeled as a normalized multiplication of probabilities of a point under each individual expert. This leads to the density that is much sharper than that under each individual expert. If each expert generates the data that satisfies some subset of the set of constraints, their product will satisfy all of the constraints (Hinton, 1999). For example, in a speech recognition problem some experts can generate signals from a certain frequency range, while others can impose finer constraints on changes of frequency in time, lexical, and semantic plausibility of the combinations of sounds.

In the PoE models, each individual expert may assign high probabilities to an irrelevant region of the data space, as long as the probability of the region is small under other experts. Thus, the PoE model is believed to produce a significantly better fit to local patterns in multi-dimensional space than mixture models.

The probability density of generating a point x under the PoE model is defined at Hinton (1999) as follows:

$$p(\mathsf{x}|\Theta_1,\ \ldots\ ,\ \Theta_m) = \frac{\prod_{i=1}^m p_i(\mathsf{x}|\Theta_i)}{\int \prod_{i=1}^m p_i(\acute{\mathsf{x}}|\Theta_i)\mathrm{d}\acute{\mathsf{x}}}. \tag{1.30}$$

Here $p_i(\mathsf{x}|\Theta_\mathsf{i})$ is the probability of the data point x under the expert with parameters Θ_i, m is the number of experts, and $\acute{\mathsf{x}}$ represents all the possible data points generated by the PoE model (Hinton, 1999).

Hinton (1999) refers to points $\acute{\mathsf{x}}$ as the *fantasy* data. Moreover, he notices that the component $p_i(\mathsf{x}|\Theta_i)$ does not have to represent probability and can define any positive function. However, we will limit our discussion only to the case when $p_i(\mathsf{x}|\Theta_i)$ denotes conditional probability of x given expert parameters Θ_i.

Notice that the denominator of (1.30) is a normalizing factor, ensuring that

$$\int p(\mathsf{x})\mathrm{d}\mathsf{x} = 1. \tag{1.31}$$

Due to the way of modeling the marginal density, **the hidden states of experts are independent given the data**, i.e.

$$p(z_1, \ldots, z_m|\mathsf{x}) = \prod_{i=1}^{m} p(z_i|\mathsf{x}), \tag{1.32}$$

where $z_1, \ldots, z_m$ denote the latent states. See Appendix A.3 for the derivation of this result.

As shown in Appendix A.4, the log-likelihood of the point x under the PoE model can be defined as

$$\frac{\partial \log p(\mathsf{x}|\tilde{\Theta})}{\partial \Theta_j} = \frac{\partial \log p_j(\mathsf{x}|\Theta_j)}{\partial \Theta_j} - E_{p(\acute{\mathsf{x}}|\tilde{\Theta})}\left[\frac{\partial \log p_j(\acute{\mathsf{x}}|\Theta_j)}{\partial \Theta_j}\right], \tag{1.33}$$

where $\tilde{\Theta} = [\Theta_1\ \Theta_2\ \ldots \Theta_m]$ is a vector of parameters of the PoE model, and the rightmost term of the r.h.s. is the expected value of the derivative of the log probability of the PoE-generated data point $\acute{\mathsf{x}}$ under the j^{th} expert.

This mathematical expectation may be evaluated by estimating the corresponding integral. Hinton (2000) describes difficulties involved its estimation by means of the rejection sampling and Gibbs sampling in Markov Chain Monte Carlo [see, e.g., Mackay (1998), Neal (1993) for description of sampling techniques]. He further suggests using contrastive divergence as an alternative error measure.

1.5.5 PoE *vs* Mixtures

In the previous sections we have seen that a marginal distribution under a PoE model may be much sharper than that under a mixture of the same experts. Here we will once again illustrate the idea by an example.

Any data-generating (or *generative*) model should satisfy two major requirements. First, it should assign high probabilities to the points from the observed sample. Second, it should not assign very high probabilities to irrelevant parts of the data space (see (Hinton, 2000, p.2) for a discussion of this).

Due to the nature of density approximation in mixture models, they generate irrelevant data whenever individual experts assign high probabilities to extraneous regions of the data space. This is not a problem under the PoE model, as long as different experts assign high probabilities to different irrelevant sub-spaces (see Hinton (2000) and the discussion in the previous section).

Figure 1.12 shows the marginal probability surface of a product and a mixture of 3 two-dimensional Gaussians[7]. The probability surface for the product of Gaussians is a single bump-shaped Gaussian, while the probability surface for a mixture of Gaussians is a vaguer superposition of bumps.

Another potential advantage of the PoE architecture is its ability to model the data with lower-dimensional experts. Consider, for example, a d-dimensional data space. The number of parameters of a mixture of m Gaussians is $d(d+3)m/2$. It is $q(q+3)m/2$ for the m-expert PoE, where q is the dimension of each expert. When each Gaussian expert imposes constraints on just one dimension, the PoE of Gaussians will require only $2m$ parameters.

[7] In chapter 1 we will consider models where the number of experts m is lower than dimensionality of the data d. This is a reasonable assumption for many practical applications. In this example, $m > d$ mostly for illustrative purposes

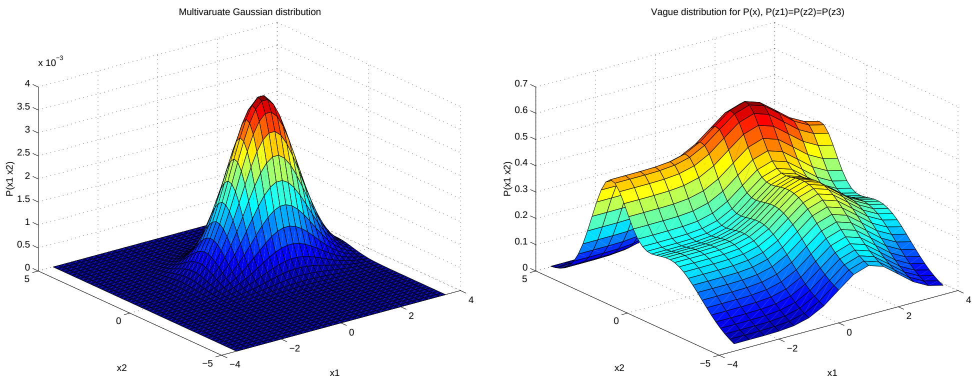

Figure 1.12: Density for a product [left] and a mixture [right] of 3 2-D Gaussians.

If a mixture model and a PoE have the same number of parameters, the marginal probability of the mixture models will generally be significantly vaguer than that under PoE. For example, if each expert in a mixture generates one dimension from a Gaussian, and the others uniformly, the probability surface of the mixture will have the shape of very vague super-imposed ridges (see figure 1.13). A PoE model with 1-D Gaussian experts, having the same number of parameters, will still result in a localized d-D Gaussian (see the discussion of the product of Gaussians).

The potential problem of PoE models may be an excessive localization of the modeled data densities. Clearly, a single product of Gaussians cannot be used for an accurate approximation of an arbitrary density. However, Hinton (1999) argues that the PoE model with more complicated experts may be more powerful.

Another possible extension of the PoE model could be seen by considering PoE mixtures. A normalized multiplication of probabilities for a PoE model roughly corresponds to some probabilistic equivalent of the logical *AND* operation. A weighted summation of probabilities for mixture models is somewhat analogous to probabilistic *OR*. A mixture of PoE models

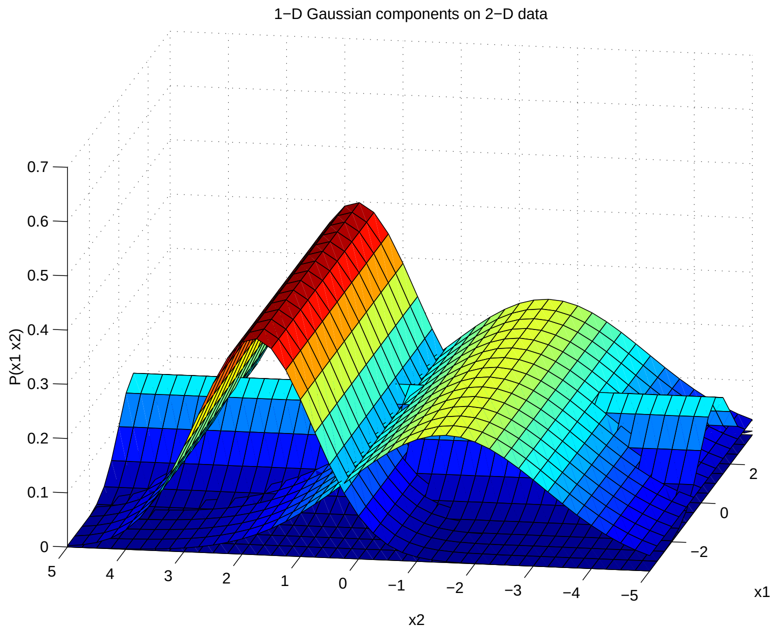

Figure 1.13: Probability surfaces of conditional densities $p(\mathsf{x}|z_i)$ in the 2D data space. Each conditional is a 1-D Gaussian.

might be thought of as a probabilistic equivalent of disjunctive normal forms (DNFs). From mathematical logic we know that there is a DNF representation of an arbitrary binary function. We may also express a speculative hope that a mixture of products of experts can accurately approximate an arbitrary probability density with a smaller overall number of parameters.

1.5.6 Products of Gaussians

In the previous sections we have mentioned several times that a product of Gaussians is a Gaussian. Indeed, let probability of a data point under the i^{th} expert be a multi-variate Gaussian distribution:

$$p(\mathsf{x}|z_i) \propto \exp\left\{-\frac{1}{2}(\mathsf{x}-\mu_i)^T \mathsf{C}_i^{-1}(\mathsf{x}-\mu_i)\right\}. \qquad (1.34)$$

Under the product of m Gaussian experts, the probability density can be expressed as

$$
\begin{aligned}
p(\mathsf{x}) &\propto \prod_{i=1}^{m} \exp\left\{-\frac{1}{2}(\mathsf{x}-\mu_i)^T \mathsf{C}_i^{-1}(\mathsf{x}-\mu_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{m}(\mathsf{x}-\mu_i)^T \mathsf{C}_i^{-1}(\mathsf{x}-\mu_i)\right\} \qquad (1.35)
\end{aligned}
$$

By completing the square in the exponent, we see that

$$p(\mathsf{x}) \propto N(\mu_\Sigma, \mathsf{C}_\Sigma),$$

where the inverse of the resulting covariance C_Σ is expressed as

$$\mathsf{C}_\Sigma^{-1} = \sum_{i=1}^{m} \mathsf{C}_i^{-1} \subset \mathcal{R}^{d\times d}. \qquad (1.36)$$

For simplification of derivations in the following chapters, we will assume that $\mu_\Sigma = 0$. This assumption does not violate generality, since we can always obtain an arbitrary mean by introducing a coordinate system with the origin at μ_Σ (see figure A.2 in Appendix A.2).

Thus, the probability of the data under the PoE model can be expressed as

$$p(\mathsf{x}) \propto N(\mathsf{0}, \mathsf{C}_\Sigma),$$

with the likelihood (for *i.i.d.* data) expressed as

$$\mathcal{L}(\mathsf{C}_\Sigma) = -\frac{N}{2} d \ln 2\pi + \frac{N}{2}\ln|\mathsf{C}_\Sigma^{-1}| - \frac{N}{2}\mathrm{tr}[\mathsf{C}_\Sigma^{-1}\mathsf{S}]. \qquad (1.37)$$

Here S is the sample covariance matrix with the assumed zero-mean, i.e.

$$\mathsf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathsf{x}_i \mathsf{x}_i^T, \qquad (1.38)$$

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

We can note that for Gaussian experts the inverse covariance C^{-1} has a simpler form of the derivative w.r.t. parameters of an individual expert, than C. For this reason, we will find it more convenient to express the log-likelihood through the inverse covariance C_{Σ}^{-1}, since it simplifies the differentiation step of the ML parameter estimation procedure.

In this work we will focus on products of Gaussians. We will investigate analytical properties of the ML parameter estimates for the product of the constrained Gaussians (PoCG) and the product of PPCA models (PoPPCA) in the visible space. In both cases we assume that each expert is a one-factor latent variable model. For simplifying the notation, we will disregard the normalizing factor in the PoE expression for marginal densities (1.30).

1.6 Useful Notations

Here we will introduce some useful notations, which we will use throughout the following chapters.

By m-PPCA we will refer to a m-factor PPCA model. Equivalently, we can say that the latent space of the PPCA model has m dimensions.

In this work we will often speak about an (e, m, r) PoE model. Here e is the dimension of each expert, m is the number of experts, and r is the number of non-zero [retained] weights in the expert weight matrix. In this work we will limit our attention only to $e = 1$-dimensional expert models. Note the difference between dimensionality of an expert and dimensionality of the data space.

For example, a 1-dimensional 3-expert PoPPCA model with 2 retained weights in 5-D data space will be denoted as a $(1, 3, 2)$ PoPPCA, and its weight matrix $\mathsf{W} \subset \mathcal{R}^{m \times r}$ may be expressed as

$$\mathsf{W} = \begin{bmatrix} w_{11} & w_{12} & 0 \\ w_{21} & w_{22} & 0 \\ w_{31} & w_{32} & 0 \\ w_{41} & w_{42} & 0 \\ w_{51} & w_{52} & 0 \end{bmatrix}. \tag{1.39}$$

If the weight matrix of the model has no $\mathsf{0}$-valued [discarded] weights, we will sometimes omit the third element of the triple and refer to it as the (e, m) model.

1.7 Summary

In Chapter 1 we have introduced the most important concepts necessary for understanding the results presented in the following chapters. We have looked at latent variables as a tool for explaining patterns in observations. The important assumption in most latent variable models is that observations are independent given the latent variables. One type of latent variable models, which we have considered, is probabilistic principal component analysis, where the probability contour in the visible space is a hyper-ellipsoid elongated in one direction and equally contracted in the others.

Mixture models estimate a complicated data density as a weighted summation of simpler *expert* probability densities. This is a well-established technique widely used in practice. However, mixture models lead to vague distributions in higher dimensions and might generate irrelevant data.

In Product of Experts (PoE) models, the data density is modeled as renormalized multiplication of expert densities. Generally, the resulting density of the PoE model is much sharper than that under each particular expert. Thus, PoE models are expected to produce a significantly better fit to local patterns in multi-dimensional data than mixture models. It is important to notice independence of hidden states given the data in PoE models, what leads to computational efficiency of Gibbs sampling procedure (Hinton, 2000) for drawing samples from PoE distributions.

If each expert is a Gaussian, their product will still be a Gaussian with the inverse covariance matrix expressed as the summation of inverse expert covariances. In the following chapters we will concentrate on investigation of properties of products of Gaussians.

Chapter 2

Products of Constrained Gaussians (PoCG)

> An expert is someone who knows some of the worst mistakes, which can be made, in a very narrow field.
>
> – Niels Bohr

In chapter 2 we will consider a product of constrained Gaussians in the d-dimensional data space, whose probability contours are equally elongated in $d-1$ directions, and are smaller in the other. We will refer to the model as the product of constrained Gaussians (PoCG), or the *rotcaf* model. Properties of data sets generated by individual experts of a PoCG model, were briefly discussed in the previous chapter (see the section on constrained Gaussians).

In the following sections we will derive the maximum likelihood solution for parameters of the PoCG model. Then, we will find an explicit expression for the model parameters, discuss their properties, and show that the model provides a probabilistic technique for minor component analysis (MCA). In the experimental section we will compare solutions, obtained by applying non-linear optimization techniques, with those obtained analytically. Moreover, we will investigate how changing the number of non-zero experts in an

m-expert model influences the value of the log-likelihood.

2.1 Motivation

The main task we will consider in this chapter is analytic investigation of products of constrained Gaussians. It will be shown that the PoCG representation has a close relation to the minor component analysis, which may be a useful dimensionality reduction technique when we are interested in dimensions of the least variation. For example, one of the possible applications of PoCG could be estimation of frequencies of sinusoidal signals, buried in white noise (Oja, 1992), (Thompson, 1979).

As discussed in Tipping and Bishop (1997), probabilistic models have a number of important advantages compared with non-probabilistic ones. The major advantage is the definition of the measure of likelihood, which allows comparison of PoCG with other dimensionality reduction techniques. We could also consider mixtures of such methods (that would correspond to mixtures of products of constrained Gaussians), and evaluate MAP estimates by combining priors on parameters with values of the log-likelihood.

However, the major motivation of this and the following chapters is to show that even though products of Gaussians are still Gaussian, they may lead to interesting models, which can potentially be used in a number of useful applications.

2.2 Analytic Parameter Estimation

In this section we will show the maximum likelihood solution for the PoCG model, derive an explicit expression for the PoCG parameters, and investigate properties of the ML solution. Each section focuses on discussions of the most important results, whenever their detailed derivations are shown in Appendix

B.

2.2.1 Covariance in the Visible Space

For Gaussian distributions, log-likelihood $\mathcal{L}$ is the function of the data covariance C. In this section we find the analytic form of the PoCG covariance matrix C_Σ. We will discuss a 1-expert case (briefly introduced in chapter 1), and a multiple expert case.

A single expert case

Let us first consider a single-expert PoCG, where the resulting covariance of the data corresponds to the covariance of a single constrained Gaussian. In this case,

$$\mathsf{C}_\Sigma^{-1} = \mathsf{C}_1^{-1} \Rightarrow \mathsf{C}_\Sigma = \mathsf{C}_1. \tag{2.1}$$

In the discussion of the constrained Gaussians in chapter 1 we have seen that the analytical form for the inverse covariance $\mathsf{C}_i^{-1} \subset \mathcal{R}^{d\times d}$ of a single constrained Gaussian corresponds to the analytical form of the covariance matrix C under the PPCA model in the visible space $\mathcal{R}^d$ (see Appendix B.1 for the derivation):

$$\mathsf{C}_i^{-1} \quad = \quad \beta_0 \mathsf{I}_d + \mathsf{w}\mathsf{w}^T, \tag{2.2}$$

where w is the weight vector, collinear with the direction of shrinkage, and $\beta_0 = \sigma_0^{-2}$ is the inverse variance in the direction of elongation.

It is easy to see that the data covariance C_i under the model can be uniquely determined by the weight vector w and the variance in the direction of elongation σ_0^2. We can further note the similarity of (2.2) with equation (1.18) for 1-factor PPCA, except that it is the *inverse* covariance matrix for the PoCG model rather than the covariance matrix which has the structure of a rank-1 update to a multiple of I_d.

A multiple expert case

For the multiple expert case we will make an important assumption that **all m experts have equal variances in the directions of elongation**, i.e.

$$\beta_0^{(0)} = \beta_0^{(1)} = \ldots = \beta_0^{(m)}. \tag{2.3}$$

A possible extension of the PoCG model could consider a product of disk-shaped experts with distinct variances (see a brief discussion in section 5.4 on assumptions and limitations of the considered models).

From (1.36), the resulting inverse covariance in the multiple-expert case may be found as

$$\mathsf{C}_\Sigma^{-1} = \sum_{i=1}^{m} \mathsf{C}_i^{-1} \quad = \quad \beta_\Sigma \mathsf{I}_d + \mathsf{W}\mathsf{W}^T \subset \mathcal{R}^{d\times d}, \tag{2.4}$$

where columns of $\mathsf{W} \subset \mathcal{R}^{d\times m}$ correspond to weight vectors of m individual experts, and

$$\beta_\Sigma = m\beta_0. \tag{2.5}$$

Applying the Woodbury transform (B.8), we obtain the expression for the covariance of the PoCG model:

$$\mathsf{C}_\Sigma = \beta_\Sigma^{-1}\left(\mathsf{I}_d - \frac{\mathsf{W}\mathsf{W}^T}{\beta_\Sigma \mathsf{I}_m + \mathsf{W}^T\mathsf{W}}\right) \subset \mathcal{R}^{d\times d}. \tag{2.6}$$

From the inverse form of the factor analysis [compare (2.4) with (1.17)] we can make a conjecture that, in contrast with the PPCA model, weights W of the PoCG model define projection of the spherical Gaussian distribution from the m-dimensional latent space onto the *minor* eigenvectors of the sample covariance in the visible d-dimensional space, while the noise term explains the larger variations.

In the following section we will derive the maximum likelihood solution for parameters β_Σ and W of the resulting PoCG distribution $N(0, \mathsf{C}_\Sigma)$.

2.2.2 ML Parameter Estimators

The first step of the analytic solution assumes calculation of derivatives of the log-likelihood $\mathcal{L}$ with respect to parameters of a probability distribution. This results in an ML equation for model parameters. The second step, which might not be always addressed to analytically, assumes solving the ML equation and deriving an explicit analytic expression for the model parameters. As we have noticed in chapter 1, derivation of the analytic solution for ML parameters may be impossible if the resulting equations are transcendental. In latent variable models, analytic solutions might not exist even if the explicit expressions can be found (for instance, when model parameters depend on unknown variables).

In this section we will show that parameter estimates for the PoCG model can, however, be found analytically. In the subsequent sections we will investigate properties of the solutions and discuss their geometric interpretation.

Derivation of the ML equation

Since the inverse noise parameter β_0 is strictly positive, we have a constraint on possible solutions for β_0. To ensure that all of the possible solutions for β_0 are positive, we express β_0 as a function of its logarithm x, i.e.

$$\beta_0 = e^x, x \in \mathcal{R}. \tag{2.7}$$

Then instead of solving $\partial\mathcal{L}/\partial\beta_0 = 0$ for β_0 directly, we will find the ML estimate for x by solving $\partial\mathcal{L}/\partial x = 0$. Clearly, the monotonic growth of the *log* function implies $\beta_0^{ML} = e^{x_{ML}}$. We will refer to x as the *log-noise* parameter.

In Appendix B.2 we show that differentiation of the log-likelihood with respect to the weight matrix W and the log-noise parameter x results in[1]

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathsf{W}} &= N(\mathsf{C}_\Sigma \mathsf{W} - \mathsf{S}\mathsf{W}) && (2.8)\\
\frac{\partial \mathcal{L}}{\partial x} &= Nm\beta_0 \mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}) && (2.9)
\end{aligned}
$$

Dropping the constant factors in (2.8) and (2.9), we see that the maximum-likelihood equations for the parameters are

$$
\begin{aligned}
\mathsf{C}_\Sigma \mathsf{W} - \mathsf{S}\mathsf{W} &= 0 && (2.10)\\
\mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}) &= 0. && (2.11)
\end{aligned}
$$

Assuming independence of W and x (see the discussion in B.2.2), we can first evaluate W_{ML}, and then express x_{ML} by solving

$$\partial \mathcal{L}(\mathsf{W}_{ML})/\partial x = 0.$$

ML solution for the experts weight matrix

It can be easily seen that equation (2.10) results in three classes of solutions:

$$
\left[\begin{array}{rcl}
\mathsf{W} &=& 0\\
\mathsf{S} &=& \mathsf{C}_\Sigma\\
\mathsf{S}\mathsf{W} &=& \mathsf{C}_\Sigma \mathsf{W}, \mathsf{W} \neq 0, \mathsf{S} \neq \mathsf{C}_\Sigma
\end{array}\right. \tag{2.12}
$$

Apart from the last case, these classes are similar to the ones described in Tipping and Bishop (1997).

In Appendix B.3 we show that all the non-zero solutions of the ML equations for W_{ML} correspond to scaled and rotated eigenvectors of the sample covariance S:

$$\mathsf{W}_{ML} = \mathsf{U}_S(\tilde{\Lambda}_S - \beta_\Sigma \mathsf{I}_m)^{1/2}\mathsf{R}^T, \tag{2.13}$$

[1]These expressions will be important when we apply gradient ascent on the log-likelihood $\mathcal{L}$ in experiments of the subsequent sections.

where $\beta_\Sigma = m\beta_0$, $\mathsf{R} \subset \mathcal{R}^{m\times m}$ is an arbitrary rotation matrix, $\mathsf{U}_S = \mathsf{U} = [\mathsf{u}_1\mathsf{u}_2\ldots\mathsf{u}_m] \subset \mathcal{R}^{d\times m}$ is a matrix whose columns correspond to m eigenvectors of S, and $\tilde{\Lambda}_S = diag(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_m) \subset \mathcal{R}^{m\times m}$, such that

$$\tilde{\lambda}_i = \begin{cases} \lambda_i^{-1} & \text{if } \mathsf{u}_i \text{ is retained;} \\ \beta_\Sigma & \text{if } \mathsf{u}_i \text{ is discarded.} \end{cases} \tag{2.14}$$

As we can see from (2.13), W_{ML} corresponds to scaled and rotated eigenvectors of S. From the expression for the negative covariance (2.4) it is easy to see that C_Σ^{-1} and, consequently, $\mathcal{L}$ are invariant under orthonormal transformations of W.

We call the projection weight $\mathsf{w} \subset \mathcal{R}^d$ *discarded* if the data from the latent space $\mathcal{R}^m$ is not projected onto the dimension w of the visible space $\mathcal{R}^d$. In the other case we say that w is *retained*. In section 2.3.2 we provide an experimental confirmation of the analytic results expressing influence of the number of the retained eigenvectors on the value of $\mathcal{L}$.

Investigation of parameter properties

In Appendix B.4 we show that

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \frac{\partial \mathcal{L}(\mathsf{W}_{\mathcal{ML}})}{\partial \beta_0} = 0 \;\Rightarrow\; \beta_0^{ML} = \frac{d-r}{m\sum_{i=r+1}^{d}\lambda_i}, \tag{2.15}$$

where r is the number of retained eigenvalues of $\mathsf{W} \subset \mathcal{R}^{d\times m}$. Like in PPCA (Tipping and Bishop, 1997), the noise term accounts for the variance in the directions lost in the projection of the sample points onto the m-D hidden space, where m is the number of 1-factor experts.

It may be shown that the likelihood $\mathcal{L}$ is maximized when all m weights, defining the $d\times m$ weight matrix, are retained, i.e. $r = m$ in expression (2.15). Then the matrix $\tilde{\Lambda}_S$ in expression (2.13) for the weight estimates is, in fact, an $m \times m$ diagonal matrix of the inverse eigenvalues of S:

$$\tilde{\Lambda}_S = \Lambda_S^{-1}. \tag{2.16}$$

The ML expressions (2.15) and (2.13) may, therefore, be re-written as follows:

$$\boxed{\beta_0^{ML} = \frac{d-m}{m\sum_{i=m+1}^{d}\lambda_i}, \ \mathsf{W}_{ML} = \mathsf{U}_S(\Lambda_S^{-1} - m\beta_0^{ML}\mathsf{I}_m)^{1/2}\mathsf{R}^T}. \tag{2.17}$$

We also prove [see Appendix B.4.1] that in contrast with Tipping and Bishop's PPCA model, the maximum likelihood solutions for the weights of the $(1, m)$ PoCG model correspond to m scaled and rotated eigenvectors of the sample covariance S with the *lowest* eigenvalues. Thus we may speak about correspondence between the PoCG model and the minor component analysis. As in the PPCA model, the number of experts m is assumed to be lower than the dimension d of the data space.

There appears to be no intuitive linear latent variable interpretation of the PoCG model. One of the possible representations, leading to the correlated noise term, is discussed in Appendix B.1.2.

2.3 Experiments with the PoCG Model

In this section we will describe experiments with the PoCG model, confirming the analytic solutions of Appendix B.3. In the first set of the experiments, we will compare the analytic solutions with the the ones obtained by applying non-linear optimization techniques[2]. We expect the parameters obtained analytically to result in the same or higher values of the log-likelihood as SCG parameters. The latter case is due to the possibility of convergence to local optima of non-linear optimization methods.

[2]We will focus on the method of scaled conjugate gradients, briefly introduced in chapter 1.

Then we will check the conditions on the number and nature of the retained eigenvalues, derived in Appendix B.4 and discussed in the previous section.

Due to limitation of space, we will not discuss experiments that are identical to one another except for a few minor changes in the values of the data dimension and the number of experts, and only briefly summarize their results.

2.3.1 Correspondence Between Likelihoods of Analytic and Non-Linear Optimization Solutions

The problem

Non-linear optimization techniques perform a search for a local minima of the error measure in the parameter space. They provide the most universal (but, quite often, the slowest) approach to estimation of the maximum-likelihood parameters. We will consider several d-dimensional m-expert PoCG models and apply method of the scaled conjugate gradients for minimizing the value of the negative log-likelihood (see Chapter 1 for the objective of non-linear optimization methods).

It is important to notice that the log-likelihood of a Gaussian may be infinite for some values of the parameters (see Day (1969) or a brief discussion on p.63 in Bishop (1995)). This may happen, for example, when the initial search point results in almost zero variance in at least one direction. Bishop (1995) mentions various techniques for handling the problem of singularity of the resulting covariance. However, in comparison of the SCG and analytical solutions we will focus only on non-singular cases.

Let s be the number of equal largest eigenvalues of S. We can think of s as the number of dimensions of the data space with the variance explainable by the distortion term I_d/β_0. We can then see that the weight matrix explains variance in at most $d - s$ dimensions, i.e.

$$rank(\mathsf{W}\mathsf{W}^T) \leq d - s. \tag{2.18}$$

We will consider three types of sample covariance matrices. In the first case, $\mathsf{S} \subset \mathcal{R}^{d \times d}$ is an arbitrary symmetric positive-definite matrix, whose $d - m + 1$ largest eigenvalues are not all identical, i.e.

$$s < d - m + 1. \tag{2.19}$$

In this case, as we can see from expression (2.17), each weight of the expert weight matrix W explains some variance of the data in its direction and should be retained. Thus, in accordance with the analytic results, we expect all columns of the weight matrix W to consist of non-zero vectors, and $rank(\mathsf{W}\mathsf{W}^T) = m$.

In the second case

$$d - m + 1 \leq s < d, \tag{2.20}$$

i.e. variance of S in $s - d + m$ directions defined by the weight matrix can be explained by the distortion term defined by β_0. Thus, $s - d + m$ expert weights can be discarded (with the corresponding columns of W set to zero). In this case, $0 < rank(\mathsf{W}\mathsf{W}^T) < m$.

Finally, we can consider the degenerate case when the sample covariance is a multiple of the identity matrix I_d. From (2.6) we can clearly see that all the data variance in this case can be modeled solely by the distortion term, and the analytic solution results in $\mathsf{W} = 0$.

In order to apply method of scaled conjugate gradients to the PoCG parameter search, we make use of the expressions of derivatives of the log-likelihood with respect to the model parameters [see (2.8) and (2.9), derived in the previous sections].

Case 1: $rank(\mathsf{W}\mathsf{W}^T) = m$**.**

Initial Parameters

Let S be a random positive-definite covariance matrix; $d = 7$ - the data dimension; $m = 1, 2, 3, 6$ - the number of experts. The initial search point is chosen at random in the parameter space $[\mathcal{R}^{d\times m+1}]$. The results are given for the sample covariance matrix with the singular values

$$L = [3.1710,\ 1.0814,\ 0.3994,\ 0.3717,\ 0.3690,\ 0.0584,\ 0.0323];$$

Results

Graph 2.1 represents typical learning curves of the PoCG model for the considered singular values of the sample covariance matrix and 1,2,3, and 6 experts in 7-D. The smoothest curve corresponds to the 1-expert PoCG model, the roughest one - to the 6-expert PoCG. In all cases, the scaled error of the SCG had the order of 10^{-9} or less after 20 iterations, what shows that the model has converged with a high degree of accuracy.

For both the analytical and the SCG solutions, the values of the log-likelihood for 1, 2, 3, and 6 experts lead to $\mathcal{L}$ equal to -15.8563, -13.9700, -13.5066, and -11.9200 respectively.

Interpretation

Comparison of the analytic solution with the results obtained by applying the SCG shows that

$$|\mathcal{L}^{(A)} - \mathcal{L}^{(SCG)}| \sim 10^{-8}, \tag{2.21}$$

where $\mathcal{L}^{(A)}$ is the log-likelihood of the parameters obtained analytically [see (2.13), (2.15)], $\mathcal{L}^{(SCG)}$ is the log-likelihood of the parameters obtained by applying the method of the scaled conjugate gradients.

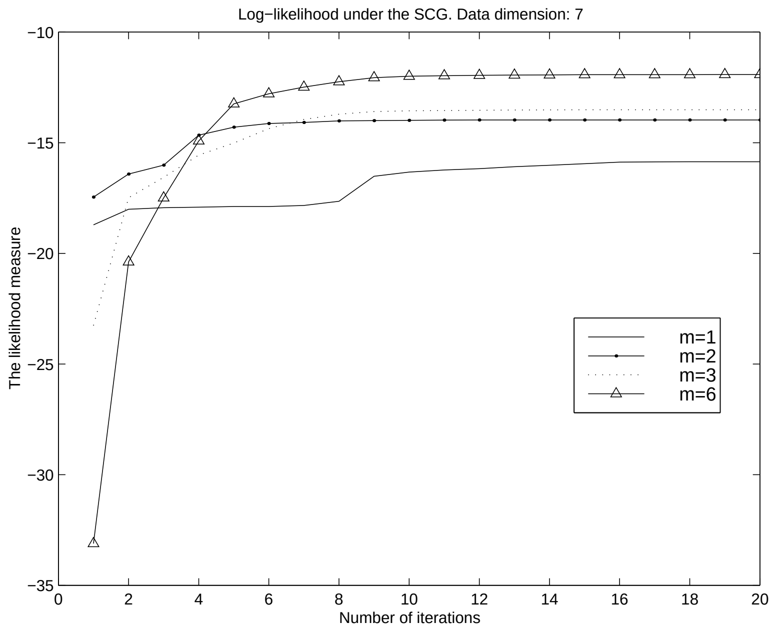

Figure 2.1: Learning curves of the SCG optimization of the log-likelihood for parameters of the $(1, m)$ PoCG model in 7-dimensional data space. The number of experts m=1,2,3,6.

Moreover, when the search was started at different initial points, non-singular solutions for the parameters still lead to the same model covariance (error $\sim 10^{-8}$) and, consequently, the same value of the log-likelihood $\mathcal{L}$. Following the notation described above, the parameters obtained in the experiment satisfy

$$|\mathsf{W}^{(A)} - \mathsf{W}^{(SCG)}\mathsf{R}| \lessapprox 10^{-8} \tag{2.22}$$

$$|\beta^{(A)} - \beta^{(SCG)}| \lessapprox 10^{-8}, \tag{2.23}$$

where $\mathsf{R} \subset \mathcal{R}^{m\times m}$ is a rotation factor, such that $\mathsf{R}^T\mathsf{R} = \mathsf{I}_m$. The sample covariance S is assumed to be the same under both the analytical and the non-singular optimization approach.

The experiment confirms the initial hypothesis that the method of the scaled conjugate gradients converges to a point in the parameter space

$d = 3$	
m	$\mathcal{L}$
1	-9.0326
2	-8.7562

$d = 7$	
m	$\mathcal{L}$
1	-15.8563
2	-13.9700
3	-13.5066
6	-11.9200

$d = 10$	
m	$\mathcal{L}$
2	-21.3525
3	-21.0174
7	-19.0683
8	-18.9270

Table 2.1: Values of the log-likelihood of the $(1, m)$-PoCG model in d-D. In all cases, $\mathcal{L}$ is practically identical for analytic and the SCG solutions.

that results in the same values of the log-likelihood as the analytically obtained solution (up to an arbitrary rotation factor for the weights). Thus we can say that the results of the experiment suggest that the analytic solution, found in the section 2.2.2, is correct and leads to the highest values of the log-likelihood. The results for this and some other similar experiments for different spaces and numbers of experts are briefly summarized in table 2.1.

The following briefly described experiments are reiterations of this case for different types of the sample covariance matrix. We will outline only differences in parameters and results.

Case 2: $0 < rank(\mathsf{W}\mathsf{W}^T) < m$

Initial Parameters

Data space: $d = 7$, number of experts $m = 6$. The singular values of the sample covariance S are $L = [1.1,\ 2.2,\ 2.7,\ 3,\ 4,\ 5,\ 5]$.

We see that $s = 2$ is the number of identical largest eigenvalues of S (i.e. the number of dimensions explained by the noise factor). We expect that $rank(\mathsf{W}^{(\mathsf{SCG})}{\mathsf{W}^{(\mathsf{SCG})}}^T) = min(d - s, m)$.

Results

In 30 runs of the SCG, started at different points of the parameter space, all the non-singular solutions converged to the same value of the log-likelihood $\mathcal{L} = -27.4459$ with the standard deviation $\sigma \approx 7.3822 \times 10^{-12}$. The scaled error of the SCG had the order of 10^{-15} after 100 iterations. As before, this shows that the model converges to the same value of $\mathcal{L}$ with a high degree of accuracy. Figure 2.2 shows the learning curves for 4 random initial search points in the parameter space.

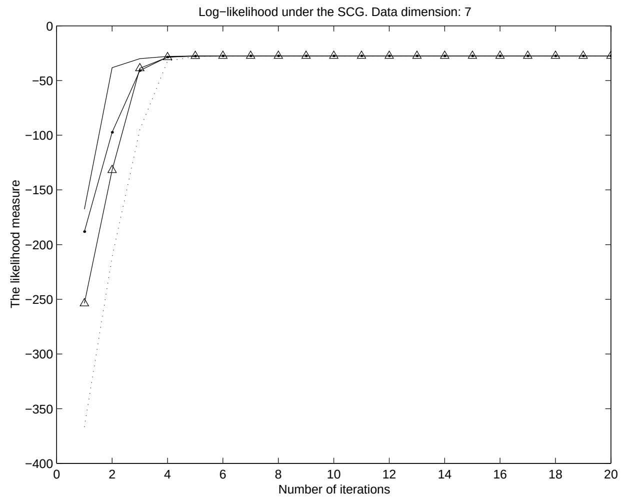

Figure 2.2: Convergence of non-singular solutions started at different points of the parameter space to the same value of $\mathcal{L}$.

Interpretation

Comparison of the likelihood of the analytic solution with that of the model with SCG parameters results in

$$|\mathcal{L}^{(A)} - \mathcal{L}^{(SCG)}| \sim 10^{-11}. \tag{2.24}$$

As before, $\mathcal{L}^{(A)}$ denotes the log-likelihood for the parameters obtained analytically; $\mathcal{L}^{(SCG)}$ is the log-likelihood for the SCG parameters. Thus, in this experiment the analytical method of the previous chapter and the SCG resulted in approximately equal values of the log-likelihood.

We can also see that if

$$\mathsf{P} = \mathsf{W}^{(SCG)} \backslash \mathsf{W}^{(A)} \subset \mathcal{R}^{m \times m}, \tag{2.25}$$

then $\mathsf{P}^T\mathsf{P}$ is a diagonal matrix of $d-s$ or less ones with other elements equal to zero.[3] In other words, the rank of $\mathsf{P} \subset \mathcal{R}^{m \times m}$ is at most $d-s$. This shows that the SCG weight matrix $\mathsf{W}^{(SCG)}$ and the weights obtained analytically $\mathsf{W}^{(A)}$ are equal up to an orthogonal factor.

Analogously to the first considered type of the sample covariance matrices,

$$|\mathsf{W}^{(A)} - \mathsf{W}^{(SCG)}\mathsf{P}| \lessapprox 10^{-6} \tag{2.26}$$

$$|\beta^{(A)} - \beta^{(SCG)}| \lessapprox 10^{-5}. \tag{2.27}$$

However, here $\mathsf{P} \subset \mathcal{R}^{m \times m}$ is a factor of rigid rotation, cancelling $m + s - d$ dimensions defined by the expert weight matrix.

To check the initial assumption that $rank(\mathsf{W}^{(\mathsf{SCG})}\mathsf{W}^{(\mathsf{SCG})^T}) = d - s$, we may also consider the singular values of $\mathsf{W}^{(\mathsf{SCG})}\mathsf{W}^{(\mathsf{SCG})^T}$:

$$\tilde{L} = [0.709,\ 0.255,\ 0.170,\ 0.133,\ 0.050,\ 5.032 \times 10^{-7},\ 5.536 \times 10^{-19}].$$

We see that the last two singular values are approximately zero, i.e. the initial assumption holds. Therefore, both the SCG and the analytical parameters explain variance in $d - s = 5$ dimensions, while the distortion term I_d/β_0 models the variance in the remaining $s = 2$ dimensions.

[3] Here $\mathsf{A}\backslash\mathsf{B}$ is defined as the left matrix division, i.e. $\mathsf{P} = \mathsf{W}^{(SCG)}\backslash\mathsf{W}^{(A)}$ is the solution to the equation $\mathsf{W}^{(SCG)}\mathsf{P} = \mathsf{W}^{(A)}$.

	$d=4, m=1,\dots,3$	$d=7, m=1,\dots,6$	$d=10, m=1,\dots,9$
$\mathcal{L}$	-14.1241	-24.7172	-35.3402
ϵ	4.473×10^{-12}	1.243×10^{-13}	1.421×10^{-13}

Table 2.2: Values of the negative log-likelihood of the $(1, m)$-PoCG models in d-D for spherical sample covariances.

As in the previous case, the experiment has shown that, for the considered conditions, the parameter values obtained by the scaled conjugate gradients correspond to the ones obtained by the analytical methods, up to a rigid projection factor for W, independently of the initial search point.

Case 3: $\mathsf{W} = \mathsf{0}$

Initial Parameters

In the last case we consider, the sample covariance I_d is a multiple of the identity matrix. The results are shown for $\mathsf{S} = 2\mathsf{I}_d$. We expect that the analytical solution will result in $\mathsf{W} = \mathsf{0}$ with the spherical noise explaining variance in all d directions. In this case we expect that the value of $\mathcal{L}$ is independent of the number of experts m.

Results

The experiment was conducted for a number of cases, and each run has confirmed the initial assumption. Table 2.2 briefly summarizes the results. The column for ϵ indicates the absolute error between $\mathcal{L}^{(A)}$ and $\mathcal{L}^{(SCG)}$.

Interpretation

While errors ϵ in Table 2.2 change for each run of the experiment, their order stays at about 10^{-11} to 10^{-13}. As before, this shows factually identical values of $\mathcal{L}$ for the analytical and non-linear cases.

Thus, the experiments have confirmed that the analytical solutions and led to the same values of log-likelihood as non-linear ones for all of the considered cases. They have also indicated that the number of equal largest eigenvalues influences the number of weights retained in W_{ML}.

2.3.2 Properties of the ML Solutions for the PoCG

The Problem

Here we will investigate how alteration of the number of the non-zero [retained] eigenvectors influences the values of $\mathcal{L}$ in the PoCG models. As before, we will use (2.17) to determine analytic values of the parameters. For the considered dimension of the data space d, we will generate $(1, m, r)$ PoCG models, where the number of experts m changes from 1 to $d-1$, and the number of the non-zero weights r changes from 0 to m. We will see how changing the number of the retained dimension r influences the log-likelihood.

We expect that, in accordance with the results described in Appendix B.4, the likelihood of $(1, m, r)$ PoCG models is greater or equal to that of $(1, m, r_1)$ PoCG models, where $r_1 \leq r \leq m$.

Initial Parameters

Data space: $d = 20$, number of experts: $m = 0, \dots, 19$, number of retained eigenvalues: $r = 0, \dots, m$. The sample covariance is set to be a random symmetric positive-definite $d \times d$ matrix.

Results

The experiment shows that the value of the log-likelihood is a non-decreasing function of the number r of non-zero eigenvalues. The number of experts m does not directly influence the value of the log-likelihood, but defines the upper bound on r.

Figure (2.3) represents a change in the value of the log-likelihood of the data under the $(1, 19, r)$ PoCG models with $r = 0, \ldots, 19$.

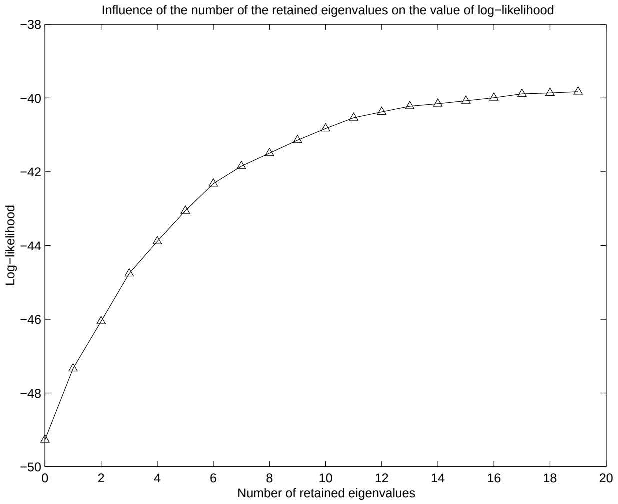

Figure 2.3: Influence of the number of the retained eigenvalues on the value of the log-likelihood in the $(1, 19, r)$ PoCG model in the 20-D data space.

Interpretation

The results of the experiment agree with the analytic form (2.6) of the sample covariance matrix of a PoCG model and confirm the analytic result of Appendix B.4 that increasing the number of non-zero weights r does not lead to a decrease in the log-likelihood of the data under the models.

2.4 Summary

In this chapter we have considered the product of m 1-dimensional constrained Gaussians [$(1, m)$ PoCG], where each expert has a probability contour equally elongated in all directions but one. All experts were assumed to have equal variances in the directions of elongation.

The analytic derivations for the model parameters have confirmed the initial hypothesis that the PoCG gives rise to minor component analysis. We have also confirmed by experiments that the derived analytic solutions correspond to the ones obtained by applying higher-order non-linear optimization techniques on the log-likelihood.

Thus, $(1, m)$ PoCG can be viewed as a probabilistic minor component analysis (MCA) model that can be used, for example, for signal extraction in digital signal processing applications, dimensionality reduction, and data visualization. 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 constrained Gaussians), permits using MCA in classification tasks (if each MCA model defines a class-conditional density), and leads to a number of other advantages over non-probabilistic MCA models[4].

We have seen that even though a product of Gaussians is still a Gaussian, it may give rise to interesting models.

[4]See the discussion of advantages of PPCA over PCA in Tipping and Bishop (1997).

Chapter 3

Products of Probabilistic Principal Component Analysers

An expert is someone who does everything else worse.

– Ruggiero Ricci

The intuition we had about the product of m one-factor probabilistic principal component analyzers was that the ML solution for the expert weights corresponds to first m principal components of the data, what is qualitatively similar to the m-PPCA case. In this chapter we investigate this hypothesis.

We begin the discussion of the chapter by considering the covariance form of an $(1, m)$ PoPPCA in the visible space, and use it to show that in contrast to the m-PPCA case, eigenvalues of the $(1, m)$ PoPPCA covariance matrix have an upper bound. This implies that for certain types of data sets a $(1, m)$ PoPPCA model has a lower representational power than an m-PPCA.

Then we discuss two approaches to deriving ML solutions for parameters of PoPPCA models. In one approach we consider implicitly scaled weights and ignore constraints on the weight magnitudes. We show that for some types of sample covariance matrices $(1, m)$ PoPPCA and m-PPCA models indeed lead to similar maximum-likelihood solutions. However, we also confirm the result of the first sections of the chapter that for certain types of sample covariance matrices the ML solution for expert weights of a $(1, m)$ PoPPCA results in imaginary weights of some of the experts. This shows a limitation of the considered PoPPCA model. We also investigate the conditions on real-valued expert weights.

In another approach to ML parameter estimation we specify the scaling factor on the expert weights explicitly. In this case we show that even though the analytic form of the ML equation for the matrix of expert weights is similar to that of the PoCG model, derivation of the analytic solutions under a PoPPCA model is significantly more complicated.

The experimental part confirms the derived results and compares the analytic solutions with the ones obtained by applying non-linear optimization methods.

3.1 Motivation

The main task of the chapter is derivation of the ML solutions for parameters of a $(1, m)$ PoPPCA model and investigation of their correspondence to m-PPCA. The results of the comparison show the difference between the models and are mostly of theoretical interest.

However, in the cases when PoPPCA gives rise to real-valued solutions, the model can be used as a possible alternative to PPCA, combining the advantages of probabilistic models (e.g. the measure of likelihood) with the framework of a product of lower-dimensional experts. Some of the possible applications of the method in these cases can be dimensionality reduction, compression, data visualization, etc.

3.2 Covariance in the Visible Space

In this section we will show the analytic expression for the covariance matrix C_Σ of the PoPPCA model. The single-expert case corresponds to the PPCA model briefly described in Chapter 1.

3.2.1 A Single Expert Case

As shown in section 1.5.1, linear mapping from the latent to the visible space is expressed as

$$\mathsf{x} = \mathsf{w}z + \epsilon, \tag{3.1}$$

where $\epsilon \sim N(0, \sigma^2 \mathsf{I})$ represents Gaussian noise, $\mathsf{x}, \mathsf{w} \subset \mathcal{R}^d$ are, respectively, the observed variable vector and the rotation vector for a single 1-D hidden variable z[1]. Then the covariance matrix may be expressed as

$$\mathsf{C} = \mathsf{w}\mathsf{w}^T + \sigma^2 \mathsf{I} \subset \mathcal{R}^{d \times d} \tag{3.2}$$

Geometrically, the probability contours of one-factor PPCA models in $\mathcal{R}^d$ have the shapes of d-dimensional hyper-ellipsoids, elongated in one and small in $d-1$ directions.

Applying the generalized Woodbury transform (B.8) to (3.2), we obtain the expression for the inverse covariance

$$\mathsf{C}^{-1} = \beta \mathsf{I} - \frac{\beta^2 \mathsf{w}\mathsf{w}^{\mathrm{T}}}{1 + \mathsf{w}^{\mathrm{T}}\mathsf{w}\beta} = \beta - \beta\gamma\mathsf{w}\mathsf{w}^{\mathrm{T}}, \tag{3.3}$$

where

$$\beta = \sigma^{-2}, \tag{3.4}$$

$$\gamma = \frac{\beta}{1 + \|\mathsf{w}\|^2\beta} = \frac{1}{\sigma^2 + \|\mathsf{w}\|^2}. \tag{3.5}$$

[1]This corresponds to 1-factor PPCA.

We see that the scaling factor γ decreases with the growth of $||\mathsf{w}||^2$ and ensures that C^{-1} in (3.3) remains positive-definite even for large weights w.

In Appendix C.1 we verify that the ML equation for the weights in the single-expert PoPPCA corresponds to the equation derived in Tipping and Bishop (1997) for PPCA.

3.2.2 A Multiple Expert Case

Similarly to the assumption we have made in the discussion of $(1, m)$ PoCG, we assume that all m experts have equal noise variances, which in this case correspond to the directions of contraction. Generalizing for products of multiple experts with constant noise variances,

$$\mathsf{C}_\Sigma^{-1} \;=\; \sum_{i=1}^{m} \mathsf{C}_i^{-1} = \beta\; m\mathsf{I}_d - \beta \mathsf{W}\Gamma\mathsf{W}^T \subset \mathcal{R}^{d\times d}, \tag{3.6}$$

where $\Gamma = diag(\gamma_1, \gamma_2, \dots, \gamma_d)$, and $\mathsf{W} = [\mathsf{w}_1\ \mathsf{w}_2\ \dots\ \mathsf{w}_m] \subset \mathcal{R}^{d\times m}$ is a matrix of experts' weights. Columns of $\mathsf{W} \subset \mathcal{R}^{d\times m}$ correspond to weight vectors of m individual experts.

Applying the Woodbury transform (B.8), we obtain the expression for the covariance of the combined experts:

$$\mathsf{C}_\Sigma \;=\; \beta^{-1}m^{-1}[\mathsf{I}_d + \mathsf{W}(m\Gamma^{-1} - \mathsf{W}^{\mathrm{T}}\mathsf{W})^{-1}\mathsf{W}^T]. \tag{3.7}$$

In the following sections we will see that for some types of sample covariance matrices a $(1, m)$ PoPPCA has a lower representational power than an m-PPCA.

3.3 The Problem of Modeling PPCA Data

Consider a dataset generated by 1-PPCA. From the form (1.18) of the covariance matrix of PPCA it is clear that this data can be generated by an

m-PPCA model ($m > 1$) by discarding (i.e. setting to zero) $m-1$ columns of the factor-loading matrix.

If we assume that an m-PPCA and a $(1, m)$ PoPPCA have the same representational power, we may argue that a $(1, m, 1)$ PoPPCA (i.e. a m-expert PoPPCA model with a single non-zero weight) should be sufficient to fit the data. If $(1, m, 1)$ PoPPCA cannot explain some of such datasets, we will find a contradiction to the assumption of computational equivalence.

□ Let us assume that the data is generated by a 1-factor PPCA with the covariance C, the noise variance σ^2, and the principal eigenvalue λ, and is being fit by a $(1, m, 1)$ PoPPCA model with the expert weight matrix $\mathsf{W} = [\mathsf{w}_1, \mathsf{0}, \dots, \mathsf{0}] \subset \mathcal{R}^{d \times m}$ and the noise variance $\sigma_\Sigma^2 = \beta^{-1}$. The number of experts m is assumed to be greater than 1.

From (3.6) we see that

$$\mathsf{C}_\Sigma^{-1} = \beta m \mathsf{I} - \beta \sum_{p=1}^{m} \gamma_p \mathsf{w}^{(\mathrm{p})} \mathsf{w}^{(\mathrm{p})\mathrm{T}}. \tag{3.8}$$

If the data can be explained by a single 1-factor expert, then

$$\mathsf{w}_2 \;=\; \ldots = \mathsf{w}_m = \mathsf{0} \tag{3.9}$$

$$\mathsf{C}_\Sigma^{-1} \mathsf{w}_1 \;=\; (\beta m - \beta \gamma_1 \|\mathsf{w}_1\|^2) \mathsf{w}_1. \tag{3.10}$$

From (3.8) and (3.10) we can notice that C_Σ has eigenvalues $(\beta m)^{-1}$ in $d-1$ directions and $(\beta m - \beta \gamma_1 \|\mathsf{w}_1\|^2)^{-1}$ in the direction of w_1. On the other hand, the covariance C of the data-generating PPCA has eigenvalues σ^2 in $d-1$ directions and the principal eigenvalue λ. If C is modeled exactly, parameters of the PoPPCA model should satisfy the following equalities:

$$\begin{cases} \lambda = (\beta m - \beta \gamma_1 \|\mathsf{w}_1\|^2)^{-1}; \\ \sigma^2 = (\beta m)^{-1}. \end{cases} \tag{3.11}$$

We can further notice that for real-valued weights and the non-zero noise variance the range of eigenvalues of C_Σ has an upper bound. Indeed,

$$\frac{\|\mathsf{w}_i\|^2}{\sigma_\Sigma^2 + \|\mathsf{w}_i\|^2} < 1 \quad \Rightarrow \quad \gamma_i \|\mathsf{w}_i\|^2 < 1 \tag{3.12}$$

$$\Rightarrow \quad \beta m - \beta \gamma_1 \|\mathsf{w}_1\|^2 > \beta(m-1), \tag{3.13}$$

where we used expression (3.5) for expressing the scaling factor γ_1. Considering this inequality and solving the system (3.11), we obtain

$$\lambda < \sigma^2 \frac{m}{m-1}. \tag{3.14}$$

However, the range of eigenvalues of the covariance matrix for the PPCA model does not need to have an upper-bound. This implies that there exist data sets generated by a 1-factor PPCA model, which cannot be explained by an m-expert PoPPCA with a single retained weight. This is the case when (3.14) does not hold, i.e. the variance in the direction of elongation of the probability contour for a data-generating PPCA is larger than $m/(m-1)$ times the variance in the other directions. This is clearly a restrictive condition, showing limitations of the PoPPCA model. ■

One possible way of modeling these data sets is by ensuring that all m experts have collinear weight vectors, so that the "team" of them models a larger variance in the direction of elongation. Appendix C.2 shows the analytic solution for the parameters of the $(1, m)$ PoPPCA model with identical weights that model arbitrary 1-factor PPCA data.

The other possible way of extending the PoPPCA model assumes distinct means and noise variances for each expert (see the brief discussion in section 5.4).

3.4 Analytic Parameter Estimation

We have seen that the role of the scaling factor Γ is to ensure that the model covariance C_Σ is positive definite and is, therefore, a valid covariance of the

Gaussian distribution. In order to simplify derivation of the ML expressions for the parameters, we will consider implicitly re-scaled weights incorporating the Γ-factor. The resulting expression will resemble the PoCG case described in section 2.2.2. The more complicated approach for the unscaled weights and explicitly specified scaling factor Γ will be considered in section 3.4.2.

3.4.1 ML Parameter Estimators for Implicitly Scaled Weights

In this section we consider unconstrained optimization on the weights incorporating the scaling factor. It will be shown that for certain types of sample covariance matrices the maximum-likelihood solution for the expert weight matrix may result in imaginary weight vectors w. We may think of imaginary weights as a trick ensuring that the PoPPCA covariance matrix remains positive-definite, what shows, however, limitations of PoPPCA models.

Derivation of the ML Equations

Let $\tilde{\mathsf{W}}$ be a $d \times m$ weight matrix such that the inverse covariance (3.6) is expressed as

$$\mathsf{C}_{\Sigma}^{-1} = \beta_{\Sigma}\mathsf{I}_d - \tilde{\mathsf{W}}\tilde{\mathsf{W}}^T, \tag{3.15}$$

where $\beta_{\Sigma} = \beta m$. We will refer to $\tilde{\mathsf{W}}$ as a matrix of implicitly scaled weights, incorporating the scaling Γ factor, and W as a matrix of unscaled weights.

From (3.15) and (3.6), the matrix of scaled weights can be expressed as

$$\tilde{\mathsf{W}} = \mathsf{W}\Gamma^{1/2}\beta^{1/2}. \tag{3.16}$$

We derive ML equations for parameters $\tilde{\mathsf{W}}$ and β of the PoPPCA model. As in the discussion of the PoCG models in section 2.2.2, we need to ensure that the inverse noise variance β is positive. We ensure this by introducing the log-noise parameter x [see (2.7)].

Under the assumption of independence of $\tilde{\mathsf{W}}$ and x, derivatives of the log-likelihood $\mathcal{L}$ w.r.t. model parameters, shown in Appendix C.3, lead to expressions

$$\frac{\partial \mathcal{L}}{\partial \mathsf{W}} = N(\mathsf{S}\tilde{\mathsf{W}} - \mathsf{C}_\Sigma \tilde{\mathsf{W}}) \tag{3.17}$$

$$\frac{\partial \mathcal{L}}{\partial x} = Nm\beta \mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}). \tag{3.18}$$

This leads to exactly the same ML equations as the ones for the PoCG model:

$$\mathsf{S}\tilde{\mathsf{W}} - \mathsf{C}_\Sigma \tilde{\mathsf{W}} = 0 \tag{3.19}$$

$$\mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}) = 0. \tag{3.20}$$

ML Solutions for the Implicitly Scaled Weights

The ML equation for the scaled weight matrix $\tilde{\mathsf{W}}$ of the PoPPCA model results in the same classes of solutions as (2.12).

In Appendix C.3.2 we show that

$$\boxed{\beta_{ML} = \frac{d-m}{m\sum_{i=m+1}^{d} \lambda_i}}, \tag{3.21}$$

where λ_i correspond to the discarded [smallest] eigenvalues of S. Moreover, if $\beta_\Sigma = \beta m$ and $\beta_\Sigma^{ML} = \beta_{ML} m$, the maximum-likelihood solution for the scaled weight matrix is expressed as

$$\boxed{\tilde{\mathsf{W}}_{ML} = \mathsf{U}_S(\beta_\Sigma^{ML}\mathsf{I}_m - \Lambda_S^{-1})^{1/2}\mathsf{R}^T}. \tag{3.22}$$

Here $\mathsf{U}_S = [\mathsf{u}_1\ \mathsf{u}_2 \ldots \mathsf{u}_m] \subset \mathcal{R}^{d\times m}$ is composed of eigenvectors of the sample covariance S, $\mathsf{R} \subset \mathcal{R}^{m\times m}$ is a rotation matrix, and $\Lambda_S = diag(\lambda_1, \ldots, \lambda_m) \subset \mathcal{R}^{m\times m}$ is a diagonal matrix of m largest eigenvalues of S, where each λ_i corresponds to the i^{th} column vector of U_S. For more discussions on the derivation of this result see derivations of the maximum-likelihood solution for the PoCG model and comments to expressions (B.46) - (B.49).

It can be easily verified that the value of the log-likelihood $\mathcal{L}$ is invariant under rotations of $\tilde{\mathsf{W}}$. Thus we can limit our attention to the simplest case when the unit right singular vectors $\mathsf{R} = \mathsf{I}_m$.

Note that since the covariance matrix is to be positive-definite, the retained eigenvalues of the sample covariance should satisfy

$$\forall i = 1, \ldots, r. \ (\beta_\Sigma \mathsf{I}_m - \Lambda_S^{-1}) > 0. \ \Rightarrow \lambda_i > \frac{1}{\beta_\Sigma}. \tag{3.23}$$

This is the key difference between PoPPCA with rescaled weights and PoCG, leading to the weight matrix corresponding to the scaled and rotated *principal* eigenvectors of the sample covariance matrix with all m weights of $\tilde{\mathsf{W}}$ retained [derivation of this result is quite similar to that of B.3.2 and B.4].

We see that the solution for $\tilde{\mathsf{W}}$ in $(1, m)$ PoPPCA corresponds to the m-factor PPCA model (see Tipping and Bishop (1997)).

Conditions for Real-Valued Unscaled Expert Weights

Expressing the unscaled weights of the i^{th} expert of the PoPPCA model, we obtain (see Appendix C.3.3)

$$\mathsf{w}^{(i)} = \frac{\tilde{\mathsf{w}}^{(i)}\beta^{-1}}{\sqrt{1 - \beta^{-1}\|\tilde{\mathsf{w}}^{(i)}\|^2}}. \tag{3.24}$$

It can also be shown that the weight of the i^{th} expert is real-valued if

$$\begin{cases} \lambda_i > 0 & \text{if } m = 1 \\ 0 < \bar{\lambda} \leq \lambda_i < m\bar{\lambda}/(m-1) & \text{if } m \geq 2 \end{cases} \tag{3.25}$$

where λ_i is the i^{th} retained eigenvalue of the sample covariance S, $\bar{\lambda}$ is the mean of the discarded eigenvalues. In other cases, $\mathsf{w}^{(i)}$ is imaginary. Clearly, (3.25) is very restrictive[2], especially for large numbers of experts m.

[2]Consider, for example, a 2-expert model in a higher-dimensional data space. For all the weights of the expert weight matrix to be real-valued, the principal eigenvalue λ of the

Notice the similarity between the upper bound on eigenvalues of the sample covariance in modeling 1-factor PPCA data (3.14).

Thus, we have shown what types of sample covariance matrices result in imaginary maximum-likelihood weight vectors W if we do not impose explicit constraints on expert weight magnitudes.

3.4.2 ML Parameter Estimators for Unscaled Weights

In this section we derive ML equations for unscaled weights, explicitly multiplied by the scaling Γ factor, which ensures that the covariance matrix of the PoPPCA model remains positive-definite even for expert weights with large magnitudes. We will see that the general analytic solution for model parameters appears difficult, and the log-likelihood is not invariant under rotation of expert weights.

Derivation of the ML equations

To ensure that the noise variance of the PoPPCA model is positive, we once again use the log-noise parameter x such that $\beta = e^x$ [see the discussion in (2.7)]. After dropping the constant factor N, derivatives of the log-likelihood $\mathcal{L}$ in each point of the parameter space can be expressed as

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial w_i^{(p)}} &= 2\beta\gamma_p \mathsf{w}^{(p)T}\left(\gamma_p w_i^{(p)} \mathsf{C}_\Sigma \mathsf{w}^{(p)} - \mathsf{c}_i\right) \\
&- 2\beta\gamma_p \mathsf{w}^{(p)T}\left(\gamma_p w_i^{(p)} \mathsf{S} \mathsf{w}^{(p)} - \mathsf{s}_i\right), \qquad (3.26) \\
\frac{\partial \mathcal{L}}{\partial x} &= \operatorname{tr}[\mathsf{I}_d - \mathsf{C}_\Sigma \mathsf{W}\Gamma^2\mathsf{W}^T - \mathsf{C}_\Sigma^{-1}\mathsf{S} + \mathsf{W}\Gamma^2\mathsf{W}^{\mathrm{T}}\mathsf{S}] = 0. \qquad (3.27)
\end{aligned}$$

Here $\mathsf{C}_\Sigma = [\mathsf{c}_1\ \mathsf{c}_2\ \ldots\ \mathsf{c}_d] \subset \mathcal{R}^{d\times d}$ is the covariance of the PoPPCA model, $\mathsf{S} = [\mathsf{s}_1\ \mathsf{s}_2\ \ldots\ \mathsf{s}_d] \subset \mathcal{R}^{d\times d}$ is the sample covariance, and $w_i^{(p)}$ is the i^{th} element

sample covariance should satisfy $\bar{\lambda} < \lambda < 2\bar{\lambda}$. Since $\lim_{m\to\infty} m/(m-1) = 1$, the interval for λ shrinks with an increase in the number of experts. For example, if $m = 10$ then the interval for the principal eigenvalue can be rewritten as $\bar{\lambda} < \lambda < (10/9)\cdot\bar{\lambda}$.

of the weight for the p^{th} expert. Derivation of this result is shown in Appendix C.4.1.

In the matrix form, expression (3.26) can be re-written as

$$\frac{\partial \mathcal{L}}{\partial \mathsf{W}} = \mathsf{W}diag(\Gamma \mathsf{W}^T(\mathsf{C}_\Sigma - \mathsf{S})\mathsf{W}\Gamma) - (\mathsf{C}_\Sigma - \mathsf{S})\mathsf{W}\Gamma. \tag{3.28}$$

In the following sections we will use the expressions for derivatives of the log-likelihood for performing gradient ascent on $\mathcal{L}$ in the parameter space.

As we show in Appendix C.4.2, the ML equation for the expert weight matrix may be written as

$$\begin{cases} \mathrm{tr}[\mathsf{I}_d - \mathsf{C}_\Sigma \mathsf{W}\Gamma^2\mathsf{W}^T - \mathsf{C}_\Sigma^{-1}\mathsf{S} + \mathsf{W}\Gamma^2\mathsf{W}^\mathrm{T}\mathsf{S}] = 0 \\ (\mathsf{C}_\Sigma - \mathsf{S})\mathsf{W}\Gamma = 0. \end{cases} \tag{3.29}$$

Since Γ is a non-singular matrix, the second equation of (3.29) results in

$$\mathsf{C}_\Sigma \mathsf{W} \quad = \quad \mathsf{S}\mathsf{W}. \tag{3.30}$$

This leads to the same classes of solutions as in the PoCG model [see section 2.2.2]. However, the analytic forms of ML solutions for parameters of the PoPPCA model are significantly more complicated.

ML Solutions for the Unscaled Weights

In Appendix C.3.2 we show that if r is the number of expert weight vectors retained in the weight matrix $\mathsf{W} \subset \mathcal{R}^{d\times m}$, then (3.30) can be re-written as

$$\mathsf{S}\mathsf{U}_1 \quad = \quad \mathsf{U}_1(\beta m \mathsf{I}_r - \mathsf{P})^{-1} \tag{3.31}$$

$$\mathsf{P} \quad = \quad \mathsf{L}_1\mathsf{R}_1^\mathrm{T}\Gamma\mathsf{R}_1\mathsf{L}_1 \subset \mathcal{R}^{r\times r}, \tag{3.32}$$

where columns of $\mathsf{U}_1 \subset \mathcal{R}^{d\times r}$ form an orthonormal basis, $\mathsf{L}_1 \subset \mathcal{R}^{r\times r}$ is a non-singular scaling matrix for r retained weights, and $\mathsf{R} \subset \mathcal{R}^{m\times r}$ is a projection matrix from m to $r < m$ dimensions.

From (3.31) we can see that if P is not a diagonal matrix, then $\mathsf{U}_1 \subset \mathcal{R}^{d \times r}$ is not necessarily a matrix of eigenvectors of S. In this case, the analytic solution for the weight matrix W of the PoPPCA model appears to be difficult. If P is a diagonal matrix, then solutions for the expert weights W may be shown to correspond to rotated and scaled principal eigenvectors of S, but the analytic expression for the scaling matrix L may be complicated. We comment on these cases in Appendix C.4.3 and Appendix C.4.4.

3.5 Experiments with the PoPPCA Model

In this section we will describe experiments with the PoPPCA model, confirming the analytic solutions of Appendix C.3. In the first set of the experiments (section 3.5.1) we will consider a sample covariance matrix whose eigenvalues satisfy the conditions (3.25) and should result in a real-valued maximum-likelihood solution for W, and compare analytic solutions with the the ones obtained by applying the method of scaled conjugate gradients to optimize the log-likelihood (1.6). As in experiments with the PoCG model [see section 2.3] we expect the parameters obtained analytically to result in the same or higher values of the log-likelihood as SCG parameters.

In the second set of experiments, described in section 3.5.2, we will investigate the "team" of equal experts modeling certain types of 1-factor PPCA data, verifying the derivation of Appendix C.2. We will also show that certain types of covariance matrices result in imaginary maximum-likelihood solutions for the expert weight matrix.

Finally in section 3.5.3, we will check the conditions on the number of the retained eigenvalues for implicitly scaled weight matrices.

3.5.1 Correspondence between Likelihoods of Analytic and SCG Solutions for Covariances Resulting in Real-Valued Weight Matrices

The problem

As in section 2.3, we will consider several d-dimensional m-expert PoPPCA models and apply method of the scaled conjugate gradients performing gradient ascent on the log-likelihood $\mathcal{L}$. As before, we will search for parameters resulting only in non-singular values of the log-likelihood.

If s is the number of equal smallest eigenvalues of S, we can think of s as a number of dimensions of the data space with the variance explainable by the noise term σ^2. As in the PoCG case, we can say that the weight matrix explains variance in at most $d - s$ dimensions, i.e.

$$rank(\mathsf{WW}^T) = d - s. \tag{3.33}$$

We can further consider three types of sample covariance matrices, such that $rank(\mathsf{WW}^T) = m,\ 0 < rank(\mathsf{WW}^T) < m$, and $\mathsf{W} = 0$ [see the discussion in section 2.3.1]. In the first case, $\mathsf{S} \subset \mathcal{R}^{d\times d}$ is an arbitrary symmetric positive-definite matrix, whose $d - m + 1$ smallest eigenvalues are not all identical, i.e.

$$s < d - m + 1.$$

In the second case

$$d - m + 1 \leq s < d,$$

i.e. variance of S in $s - d + m$ directions defined by the weight matrix can be explained by the noise term defined by σ^2, and $s - d + m$ expert weights can be discarded. Finally, the third case considers the degenerate situation when the sample covariance is a multiple of the identity matrix I_d.

Case 1: $rank(\mathsf{W}\mathsf{W}^T) = m$

Here we consider the case when the matrix of unscaled weights W is real-valued, with all m weights retained.

Initial Parameters

Let $d = 7$ be the dimension of the data space, $m = 4$ be the number of experts, and the smallest $d-m$ eigenvalues of the sample covariance be equal to 1, 1.1, and 1.2. According to condition (3.25) for eigenvalues of S resulting in real-valued weights, we can find that all weight vectors of the expert weight matrix W are real if the largest eigenvalue λ_7 of S satisfies $\lambda_7 \lesssim 1.4667$. Let us consider a sample covariance matrix with singular values [1 1.1 1.2 1.3 1.35 1.37 1.44], which satisfies these conditions.

The initial search point of the parameter space $[\mathcal{R}^{d \times m+1}]$ is chosen at random. In order to apply method of scaled conjugate gradients to the PoPPCA parameter search, we make use of expressions (3.27) and (3.28) for derivatives of the log-likelihood $\mathcal{L}$ with respect to the log-noise parameter x and the weight matrix W.

Results

Graph 3.1 represents typical learning curves of the PoPPCA model with search started at different points of the parameter space for $(4, 1)$ PoPPCA in 7-D data space. In all cases, the scaled error of the SCG had the order of 10^{-9} or less after 20 iterations, what shows that the model has converged with a high degree of accuracy. The analytically obtained weight matrix is indeed real-valued. In 30 runs of the SCG from different initial points in the parameter space, both the analytic

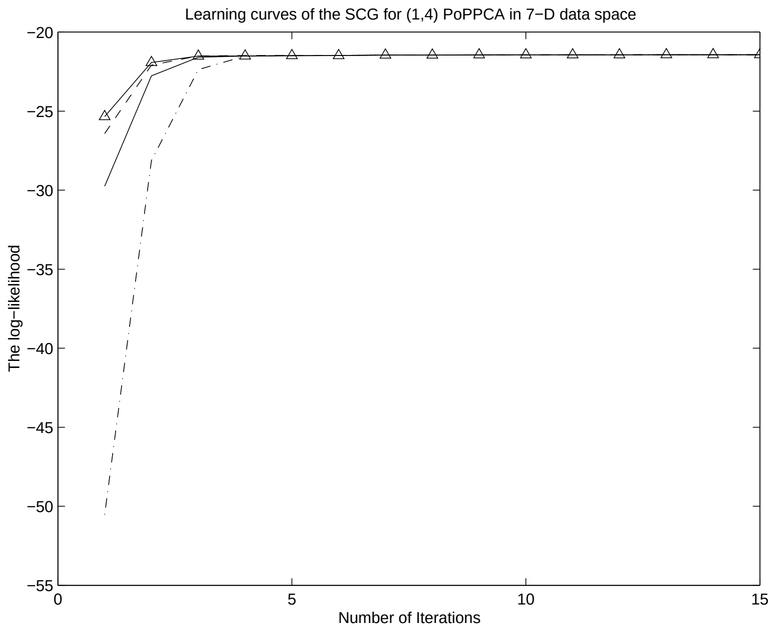

Figure 3.1: Learning curves of the SCG optimization of the log-likelihood for parameters of the (1,4) PoPPCA model in 7-dimensional data space.

and the SCG solutions converged to the value of the log-likelihood $\mathcal{L} \approx 21.3930$ with the standard deviation of the SCG solutions equal to $\sigma \approx 2.1249 \cdot 10^{-6}$.

Interpretation

Comparison of the analytic solution (3.24) for each column of W with the results obtained by applying the method of scaled conjugate gradients shows that

$$|\mathcal{L}^{(A)} - \mathcal{L}^{(SCG)}| \sim 10^{-6}, \tag{3.34}$$

where $\mathcal{L}^{(A)}$ is the log-likelihood of the parameters obtained analytically, and $\mathcal{L}^{(SCG)}$ is the log-likelihood of the parameters, obtained by applying the method of scaled conjugate gradients.

Moreover, the SCG search started at different initial points still leads to approximately the same model covariance (error $\sim 10^{-3}$) and, consequently, approximately the same value of the log-likelihood $\mathcal{L}$ (error $\sim 10^{-6}$). This indicates that the value of $\mathcal{L}^{(SCG)}$ is not significantly influenced by the initial point of the search[3].

By applying SVD to both the weight matrix, obtained from (3.24), and the sample covariance, we can confirm that the analytically obtained expert weights are indeed real-valued and corresponding to 4 rotated and scaled principal components of S. This result has been checked in other similar experiments for various data dimensions, numbers of experts, and singular values of sample covariance matrices, which were analytically chosen to lead to a real-valued W with all m retained columns.

The experiment confirms the analytic result that for certain types of sample covariance matrices the SCG converges to parameters that give rise to the same value of the log-likelihood as the analytic solution for W, obtained from implicitly scaled weights $\tilde{\mathsf{W}}$. It also confirms the analytic results that expert weights W correspond to rotated and scaled principal eigenvectors of S. Thus we can say that the experiment suggests that the analytic solution, found in the previous section, is correct and leads to the highest values of the log-likelihood.

The following briefly described experiments are reiterations of this case for different types of the sample covariance matrix. We will outline only differences in parameters and results.

Case 2: $0 < rank(\mathsf{W}\mathsf{W}^T) < m$

Here we consider the case when the matrix of unscaled weights W is real-valued, and some of its m weights are discarded.

[3] As in the PoCG case, we are considering only the values of parameters resulting in non-singular covariance matrices.

Initial Parameters

Data space: $d = 7$, number of experts: $m = 4$. The singular values of the sample covariance S are [1 1 1 1 1 1.1 1.2]. Since $\lambda_7 < 4/3$, we expect that according to (3.25) the expert weight matrix W is real-valued. Since $s = 5$ is the number of identical largest eigenvalues of S (i.e. the number of dimensions explained by the noise factor), we expect that $rank(\mathsf{W}^{(\mathsf{SCG})}\mathsf{W}^{(\mathsf{SCG})^T}) \leq d - s = 2$.

Results

In 30 runs of the SCG, started at different points of the parameter space, the parameters of the PoPPCA model have converged to log-likelihood $\mathcal{L} = 20.142771$ with the standard deviation $\sigma \approx 1.5052 \cdot 10^{-7}$. The scaled error of the SCG had the order of 10^{-15} after 100 iterations. As before, this shows that the model converges to the same value of $\mathcal{L}$ with a high degree of accuracy.

Interpretation

Let $\mathcal{L}^{(A)}$, $\mathsf{C}_\Sigma^{(A)}$, and $\mathsf{W}^{(A)}$ be the log-likelihood, the covariance, and the expert weight matrix for PoPPCA parameters obtained analytically by applying the derived learning rules, and $\mathcal{L}^{(SCG)}$, $\mathsf{C}_\Sigma^{(SCG)}$, and$\mathsf{W}^{(SCG)}$ denote the same entities for parameters obtained by applying the SCG. Comparison of analytic and SCG solutions shows that

$$max_{ij}(|[\mathsf{C}_\Sigma^{(A)} - \mathsf{C}_\Sigma^{(SCG)}]_{ij}|) \sim 10^{-6} \quad (3.35)$$

$$|\mathcal{L}^{(A)} - \mathcal{L}^{(SCG)}| \sim 10^{-10}. \quad (3.36)$$

We can say that in this experiment both the analytical method of section (3.4.1) and the SCG result in approximately equal values of the log-likelihood.

To check the initial conjecture that $rank(\mathsf{W}^{(\mathsf{SCG})}\mathsf{W}^{(\mathsf{SCG})^T}) \leq d - s$, we may consider the singular values of $\mathsf{W}^{(\mathsf{SCG})}\mathsf{W}^{(\mathsf{SCG})^T}$, which in the considered case are equal to $[0.6667,\ 0.3636,\ 2.7\cdot 10^{-15},\ 8.7\cdot 10^{-16},\ 5.4\cdot 10^{-17},\ 3.4\cdot 10^{-17},\ 9.0\cdot 10^{-19}]$. We see that the last five singular values are approximately zero, i.e. the initial assumption about the rank of $\mathsf{W}\mathsf{W}^T$ holds. Therefore, both the SCG and the analytical parameters explain variance in $d - s = 2$ dimensions, while the noise term $\sigma^2/m\mathsf{I}_d$ models the variance in the remaining $s = 5$ dimensions. From (3.7), this agrees with the analytic considerations.

Experiments with the PoCG model have shown that $\mathsf{W}^{(A)} = \mathsf{W}^{(SCG)}\mathsf{P}$, where $\mathsf{P} \subset \mathcal{R}^{m\times m}$ is a projection matrix such that $\mathsf{P}\mathsf{P}^T$ is a diagonal matrix of $min(m, d-s)$ ones. Due to the presence of the scaling factor $\Gamma \subset \mathcal{R}^{m\times m}$ in PoPPCA models, we cannot speak about equality of the weight matrices $\mathsf{W}^{(SCG)}$ and $\mathsf{W}^{(A)}$ as in the PoCG case. However we could notice that the similar identity holds for rescaled weights, i.e.

$$\mathsf{W}^{(A)}(\Gamma^{(A)})^{1/2} \approx \mathsf{W}^{(SCG)}(\Gamma^{(SCG)})^{1/2}\mathsf{P}, \tag{3.37}$$

where

$$\mathsf{P} = \frac{\mathsf{W}^{(SCG)}(\Gamma^{(SCG)})^{1/2}}{\mathsf{W}^{(A)}(\Gamma^{(A)})^{1/2}} \tag{3.38}$$

$$\mathsf{P}^T\mathsf{P} = diag(1, \dots, 1, 0, \dots, 0) \tag{3.39}$$

$$rank(\mathsf{P}) = min(d - s, m). \tag{3.40}$$

Indeed, in the considered run of the experiment

$$|\beta^{(A)} - \beta^{(SCG)}| \lesssim 10^{-6} \tag{3.41}$$

$$|\mathsf{W}^{(A)}(\Gamma^{(A)})^{1/2} - \mathsf{W}^{(SCG)}(\Gamma^{(SCG)})^{1/2}\mathsf{P}| \lesssim 10^{-8}, \tag{3.42}$$

where $\mathsf{P}^T\mathsf{P} \subset \mathcal{R}^{4\times 4}$ is a diagonal matrix with singular values $[1\ 1\ 0\ 0]$.

As in the PoCG case, the experiment has shown that the parameter values, obtained by the scaled conjugate gradients, result in the same value of the log-likelihood $\mathcal{L}$ as the analytic solutions. Rescaled analytic and SCG weights are equal up to a rigid projection factor.

Case 3: $\mathsf{W} = 0$

Initial Parameters

Here we will consider the sample covariance equal to a multiple of the identity I_d matrix. The results are shown for $\mathsf{S} = 1.5\mathsf{I}_d$. We expect that the analytic solution will result in $\mathsf{W} = 0$ with the spherical noise explaining variance in all d directions. In this case, we expect that the value of $\mathcal{L}$ is independent of the number of experts m. The data dimension d is taken to be equal to 7.

Results

The values of the log-likelihood $\mathcal{L}^{(A)}$ and $\mathcal{L}^{(SCG)}$ do not change with the number of experts, and $\mathcal{L} \approx 22.7034$ for all $m = 1, \ldots, 6$. The difference in log-likelihoods for the analytic and the non-linear solutions is

$$|\mathcal{L}^{(A)} - \mathcal{L}^{(SCG)}| \leq 10^{-11}.$$

The results obtained by applying the analytic procedure for parameter estimation (3.24) show that $\mathsf{W}^{(A)} = \mathsf{I}_d \cdot 10^{-7} \approx 0 \subset \mathcal{R}^{d\times m}$. The expert weight matrix obtained by applying the SCG results in $\mathsf{W}^{(SCG)} \sim \mathsf{I}_d \cdot 10^{-3}$, what shows close correspondence to the analytic solution:

$$|\mathsf{W}^{(A)} - \mathsf{W}^{(SCG)}| \lesssim 10^{-3} \tag{3.43}$$

$$|\beta^{(A)} - \beta^{(SCG)}| \lesssim 10^{-6}. \tag{3.44}$$

Notice that the noise variance $\beta^{(A)} = 1/(1.5m)$.

Interpretation

As in the previous cases, the experiment shows very close values of $\mathcal{L}$ for analytical and non-linear approaches to parameter estimation.

Thus, the experiments have confirmed that the analytic solutions lead to the same values of the log-likelihood as non-linear ones when eigenvalues of the sample covariance matrix result in real-valued expert weight matrices [see condition (3.25)]. They have also indicated that the number of equal largest eigenvalues influences the number of weights retained in W_{ML} just like in PoCG models.

3.5.2 Modeling 1-Factor PoPPCA Data

The Problem

Here we assume that the data is generated by a 1-factor PPCA with the exact sample covariance matrix, i.e. $\mathsf{S} = \mathsf{C}$. C is the covariance of the PPCA model. According to the results obtained in section 3.3, this data cannot be always modeled by a $(1, m)$ PoPPCA with a single retained weight, where the number of experts m is greater than 1. In the experiment of this section we will verify the analytic results of section 3.3. We will further confirm the analytic solution for parameters of the "team" of experts [see Appendix C.2] by showing that the SCG does not diverge from the analytically determined point in the PoPPCA parameter space.

Initial Parameters

Data dimension: d=7, number of experts: m=4. From (3.14) we see that a $(1, m)$ PoPPCA results in imaginary maximum-likelihood solution for expert weights W_{ML} if the variance λ of the data-generating PPCA in the direction of elongation of the probability contour is larger than $m/(m-1)$ times the variance σ^2 in the other directions:

$$\lambda > \sigma^2 \frac{m}{m-1}. \tag{3.45}$$

We consider the covariance of the PoPPCA with $\lambda = 3,\ \sigma^2 = 1$ (i.e. the singular values of C are [3 1 1 1 1 1 1]).

Appendix C.2 shows that one of the possible ML solutions in this case is

$$\begin{aligned} \mathsf{w}_1 = \mathsf{w}_2 = \ldots = \mathsf{w}_m &= \sqrt{8}\mathsf{u}_1/\|\mathsf{u}_1\| \\ \sigma_\Sigma^2 &= 4. \end{aligned} \tag{3.46}$$

We will check whether the non-linear search started at this point is indeed a local optimum of $\mathcal{L}$.

Results

Expression(3.24) results in the ML solution for the weight matrix with a single retained weight such that $\mathsf{W}_{ML} = [\mathsf{w}_1\ 0\ \ldots 0] \subset \mathcal{R}^{d\times m}$, where $\mathsf{w}_1 \subset \mathcal{C}^d$:

$$\mathsf{w}_1 = [-0.44i,\ -0.88i,\ -0.01i,\ 0.40i,\ 0.23i,\ -1.76i,\ 1.45i]^T.$$

Started at a random point, the SCG converges to $\mathcal{L}_0 \approx 20.963752$ after 150 iterations. The weight parameters in this case are real-valued. When started at the analytically obtained point of (3.46), the SCG quits after 1 iteration with the same value of the log-likelihood $\mathcal{L}_1$, so that $|\mathcal{L}_1 - \mathcal{L}_0| \approx 4.95 \cdot 10^{-7}$.

Interpretation

The experiment shows that if the data is generated by 1-PPCA with the sample covariance satisfying (3.45), then it is modeled by $(1, m)$ PoPPCA with one complex weight. It also shows that this data can be modeled by equal experts and confirms the analytical results of Appendix C.2 for the parameters of the corresponding PoPPCA model.

Finally, we should say that in general predicting the SCG solution in PoPPCA is difficult. If the sample covariance satisfies (3.25) then solutions of $(1, m)$ PoPPCA correspond to solutions of m-factor PPCA exactly. But if (3.25) does not hold, the exact correspondence would result in imaginary unscaled weights W of the PoPPCA model.

Separate tests have shown that in this case left singular vectors of the rescaled weights $\mathsf{W}^{(SCG)}(\Gamma^{(SCG)})^{1/2}$ are still closely related to principal components of the sample covariance S [with the absolute error of approximately $10^{-3} - 10^{-1}$]. However, the weights of the SCG solution remain real-valued.

3.5.3 Properties of the ML Solutions for the PoPPCA

The Problem

Here we will investigate how alteration of the number of the retained eigenvectors r influences the values of the log-likelihood of a $(1, m, r)$ PoPPCA. We consider only the unconstrained case and allow the weights of the PoPPCA model to be imaginary. We expect that in this case the $(1, m)$ PoPPCA is similar to the m-PPCA (Tipping and Bishop, 1997), and the log-likelihood of the data under the model is a nondecreasing function of r, where $r \leq m$.

We will use (3.24) to determine expert weights analytically. For the considered dimension d of the data space, we will generate a sequence of $(1, m, r)$ PoPPCA models, where the number of experts m changes from 1 to $d - 1$, and the number of the non-zero weights r changes from 0 to m.

Initial Parameters

Data space: $d = 20$, number of experts: $m = 0, \dots, 19$, number of retained eigenvalues: $r = 0, \dots, m$. The sample covariance is set to be a random symmetric positive-definite $d \times d$ matrix. We consider $(1, m, r)$ PoPPCA models with unconstrained weights.

Results

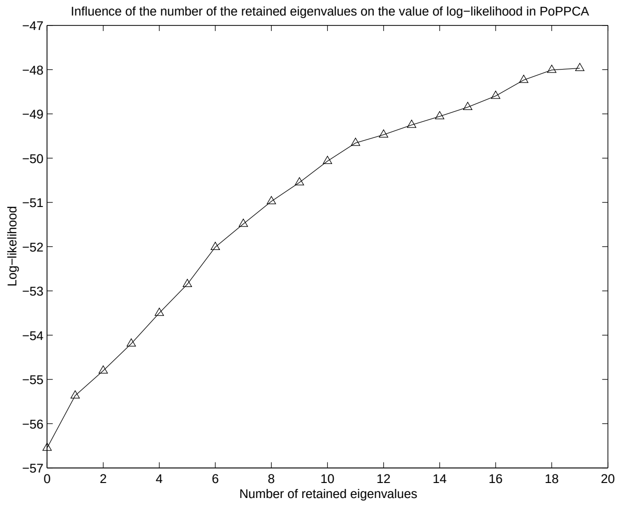

Figure 3.2: Influence of the number of the retained eigenvalues on the value of the log-likelihood $\mathcal{L}$ under the $(1, 19)$ PoPPCA in the 20-D data space for the unconstrained weights.

Figure 3.2 represents changes in the value of the log-likelihood $\mathcal{L}$ with the growth in the number of the retained expert weights. The experiment shows that in the considered case the value of the log-likelihood is a non-decreasing function of the number r of non-zero eigenvalues.

Interpretation

The experiment agrees with the analytic results for PPCA models (Tipping and Bishop, 1997) and confirms that an increase in the number of retained unconstrained implicitly-scaled [and possibly imaginary] weights in a $(1, m)$ PoPPCA model does not lead to a decrease of the log-likelihood $\mathcal{L}$ of the data under the models.

3.6 Summary

In this chapter we have investigated the product of m 1-dimensional PPCA models, whose probability contours in d-D space correspond to d-D hyper-ellipsoids, elongated in one direction and equally contracted in the remaining $d-1$ directions. All experts were assumed to have equal variances in the directions of contraction.

Initially we had expected that a product of m 1-D PPCA models [$(1,m)$ PoPPCA] would be equivalent to an m-factor PPCA, therefore the maximum-likelihood solution for the expert weight matrix in $(1,m)$ PoPPCA would correspond to m rotated and scaled principal components of the sample co-variance matrix, and that both models would have the same representational power. Since it is known that for $m \geq r$ an m-PPCA can model the data, generated by r-PPCA, by retaining only r and setting the extraneous $m-r$ weights of the factor loading matrix to zero, we had expected that a data set, generated by an r-PPCA, could be explained by a product of m experts with only r non-zero weights [$(1,m,r)$ PoPPCA].

It was shown analytically and confirmed experimentally that for certain types of data the ML solution for the expert weight matrix in a $(1,m,r)$ PoP-PCA model indeed corresponds to that in r-PPCA. However, for other data sets the unconstrained maximum-likelihood solution gives rise to imaginary expert weights, what indicates limitations of the considered PoPPCA model and its lower representational power compared with PPCA at least for some sample covariance matrices. These problematic data sets may be modeled by increasing the number of non-zero ["retained"] weights in the expert weight matrix. For example, it was shown that one way to model data generated by an arbitrary 1-PPCA is was by using a $(1,m,m)$ PoPPCA model with collinear weights, which we refer to as the "team of experts" solution.

Chapter 4

CD Learning Rule in Gaussian Models

> Make three correct guesses consecutively, and you will establish a reputation as an expert.
>
> – Laurence J. Peter

In the previous chapters we have tried to estimate analytic forms of the maximum likelihood solutions for parameters of the PoCG and PoPPCA models. As we have seen, the analytic solutions are not always straightforward even for relatively simple Gaussian experts. Here we consider sampling methods for approximation of optimal parameters of Gaussian models.

One of the unsupervised learning techniques, using stochastic approximation for density estimation, is the *Boltzmann Machine* (BM), introduced in Ackley, Hinton and Sejnowski (1985). We begin the discussion of this chapter by briefly describing BM and Boltzmann learning, and outline the results of Freund and Haussler (1992) and Hinton (2000), showing the correspondence of the special case of Boltzmann machines to PoE.

Then we describe the contrastive divergence (CD) learning rule for products of experts. The rule may be viewed as optimization of the log-likelihood, approximated by a single iteration of Gibbs sampling, and was suggested in Hinton (2000) as an alternative to optimization of the true log-likelihood [e.g. in Boltzmann learning]. However, there have been no analytic results confirming equality of optimal solutions in both cases.

We will consider a 1-D 1-factor linear Gaussian latent variable model, corresponding to a very simple Gaussian Boltzmann machine, and see how i-step CD learning compares with Boltzmann learning in this case. We also estimate moments of the distribution of the weight update for this model.

4.1 Motivation

The principal goal of the chapter is analytical investigation of the contrastive divergence learning for the Gaussian case, and comparison of the resulting optimal solutions with the ones maximizing the true log-likelihood in Boltzmann learning. The other goal is investigation of the moments of the distribution for the weight update in i-step CD learning in a simple Gaussian model.

The investigated topics are mostly of theoretical interest. However, the derived results do not contradict the claim that the contrastive divergence and the maximum-likelihood optimization lead to similar solutions, and imply that the CD learning may indeed be used instead of maximization of the log-likelihood at least in some Gaussian models.

4.2 Boltzmann Machines and PoE

In this section we briefly discuss the BM architecture and Boltzmann learning, outline the analytic results of Freund and Haussler (1992) and Hinton (2000), showing the correspondence of restricted Boltzmann machines to PoE, and briefly discuss the correspondence of linear Gaussian hidden variable models to Boltzmann machines.

4.2.1 The BM Overview

Boltzmann machines (Ackley, Hinton and Sejnowski, 1985) are described in detail in multiple sources on neural nets (see, for example, Haykin (1999) or Hertz, Krogh, and Palmer (1991)), and here we only very briefly outline their properties and architecture.

A Boltzmann machine is a latent variable model providing a mechanism for modeling the data-generating distribution. As before, the visible variables accept certain patterns from the environment, while hidden variables capture higher-order statistical correlations in the observations.

Let $\mathsf{x} \subset \mathcal{R}^d$ and $\mathsf{z} \subset \mathcal{R}^m$ be stochastic visible and hidden variables of a BM. The state of the Boltzmann machine is determined by the vector of all the variables $\mathsf{y} = \begin{bmatrix} \mathsf{x} \\ \mathsf{z} \end{bmatrix} \subset \mathcal{R}^{d+m}$. The connections between the hidden and the visible variables are specified by the non-zero weight matrix $\mathsf{W} \subset \mathcal{R}^{(d+m)\times(d+m)}$, which is assumed to be symmetric[1]. There are no other restrictions on W, i.e. hidden and visible variables may have intra-layer connections.

The joint probability distribution of hidden and visible variables of Boltzmann machines follows the Boltzmann-Gibbs distribution, i.e.

$$p(\mathsf{x}, \mathsf{z}) = p(\mathsf{y}) = \frac{1}{Z} \exp\{-E(\mathsf{y})\}, \tag{4.1}$$

where Z is the renormalizing factor, and

$$E(\mathsf{y}) = \frac{\beta}{2} \sum_{i,j} y_i y_j w_{ij} = \frac{\beta}{2} \mathsf{y}^T \mathsf{W} \mathsf{y} \tag{4.2}$$

is the *energy* of the BM state.

[1] We assume that the elements w_{ij} and w_{ji} of W specify the weight of the connection from y_i to y_j and from y_j to y_i respectively, and $w_{ij} = w_{ji}$. Each y_k can be either a hidden or a visible variable.

4.2.2 Boltzmann Learning

Since a Boltzmann machine is a probabilistic model, its parameters may be learned by maximizing the log-likelihood.

Marginalizing the joint probability (4.1) over the latent variables, we obtain

$$p(\mathsf{x}) = \frac{1}{Z}\int \exp\{-E(\mathsf{x},\mathsf{z})\}d\mathsf{z}, \tag{4.3}$$

where

$$Z = \int\int \exp\{-E(\mathsf{x},\mathsf{z})\}d\mathsf{x}\ d\mathsf{z}. \tag{4.4}$$

If $\mathcal{X}_0$ denotes the sample from the data distribution, then the log-likelihood may be expressed as

$$\begin{aligned}\mathcal{L}(\Theta) &= \log \prod_{\mathsf{x}\in\mathcal{X}_0} p(\mathsf{x}) = \sum_{\mathsf{x}\in\mathcal{X}_0} \log p(\mathsf{x}) \\ &= \sum_{\mathsf{x}\in\mathcal{X}_0}\left[\log\left(\int \exp\{-E(\mathsf{x},\mathsf{z})\}d\mathsf{z}\right)\right. \\ &- \left.\log\left(\int\int \exp\{-E(\mathsf{x},\mathsf{z})\}d\mathsf{x}\ d\mathsf{z}\right)\right].\end{aligned} \tag{4.5}$$

If w_{ij} is the weight connecting y_i and y_j, expressions (4.2) and (4.5) lead to[2]

$$\begin{aligned}\frac{\partial\mathcal{L}(\mathsf{w})}{\partial w_{ij}} &\propto \sum_{\mathsf{x}}\left[\int p(\mathsf{z}|\mathsf{x})\ y_i y_j d\mathsf{z} - \int\int p(\mathsf{x},\mathsf{z})\ y_i y_j d\mathsf{x}\ d\mathsf{z}\right] \\ &= \langle\langle y_i y_j\rangle_+ - \langle y_i y_j\rangle_-\rangle_{x_0},\end{aligned} \tag{4.6}$$

where

$$\begin{cases}\langle y_i y_j\rangle_+ \stackrel{\text{def}}{=} \int p(\mathsf{z}|\mathsf{x})\ y_i y_j d\mathsf{z} & \text{– the “positive stage”} \\ \langle y_i y_j\rangle_- \stackrel{\text{def}}{=} \int\int p(\mathsf{z},\mathsf{x})\ y_i y_j d\mathsf{z}\ d\mathsf{x} & \text{– the “negative stage”},\end{cases} \tag{4.7}$$

[2]For the derivation see, for example, Hertz, Krogh, and Palmer (1991).

and $\langle \ldots \rangle_{x_0}$ indicates averaging over the data distribution.

At this stage we can notice that averaging over the joint distribution of all the variables of the BM at the negative stage of learning may be very computationally intensive.

4.2.3 Restricted Boltzmann Machines

Note that in the derived system (4.7) variables y_i and y_j can be either hidden or visible. Smolensky (1986) introduced the *harmony* theory where he considered restricted Boltzmann machines (RBM) with one layer of latent variables, one layer of visible variables, and no intra-layer connections. For this case we can re-write (4.6) as

$$\frac{\partial \mathcal{L}(\mathsf{w})}{\partial w_{ij}} = \langle \langle x_i z_j \rangle_+ - \langle x_i z_j \rangle_- \rangle_{x_0}, \tag{4.8}$$

where x_i and z_j is a visible and a hidden variable.

Freund and Haussler (1992) further investigated the *harmonium* models and showed that the probability of generating a visible vector x in a RBM is proportional to product of probabilities that the vector was generated by each of the hidden variables. Thus, PoE can be considered to be an RBM with one expert per hidden variable (Hinton, 2000), and Boltzmann learning, maximizing the likelihood, may be applied to PoE. It may also be shown [see (Hinton, 2000, p. 8)] that differentiation of the expression (1.33) for the log-likelihood of the PoE w.r.t. the weight w_{ij} indeed leads to the Boltzmann learning rule (4.8). Hinton further notes that PoE and BM are very different methods of density estimation, but their intersection is a restricted Boltzmann machine.

4.2.4 Gaussian Models and Boltzmann Machines

Here we can note a correspondence of a latent variable Gaussian model to a restricted Boltzmann machine.

Indeed, in the Gaussian case with $\mathsf{0}$-mean the joint probability is expressed as

$$p(\mathsf{x},\mathsf{z}) \quad \propto \quad \exp\left\{-\frac{1}{2}[\mathsf{x}^T\mathsf{z}^T]\mathsf{C}^{-1}\begin{bmatrix}\mathsf{x}\\ \mathsf{z}\end{bmatrix}\right\}, \tag{4.9}$$

where C is the covariance of the complete data. We may notice the correspondence of this expression to the form (4.1) of the joint distribution in a BM, where the inverse covariance C^{-1} corresponds to the weight matrix of the BM. Therefore, we can use Boltzmann-type learning to determine parameters of the inverse covariance of the complete data of a Gaussian latent variable model.

4.3 Learning by Minimizing Contrastive Divergence

In this section we briefly discuss the motivation behind an alternative approach to optimizing the log-likelihood and outline the contrastive divergence learning rule suggested in Hinton (2000).

4.3.1 The Problem of Estimating the Likelihood

Consider the derivative of the log-likelihood (1.33) with respect to a PoE parameter at a single point x. We see that estimation of the likelihood requires evaluation of the normalizing factor averaged over the distribution of the PoE-generated "fantasy" data $\acute{\mathsf{x}}$. This distribution is unknown at the learning stage, but can be modeled and sampled from by using Monte Carlo sampling techniques [see Appendix A.5].

Averaging expression (1.33) over the data distribution and using the notation of Hinton (2000), we obtain

$$\left\langle \frac{\partial \log p(\mathsf{x}|\hat{\Theta})}{\partial \Theta_i} \right\rangle_{Q^0} = \left\langle \frac{\partial \log p_i(\mathsf{x}|\Theta_i)}{\partial \Theta_i} \right\rangle_{Q^0} - \left\langle \frac{\partial \log p_i(\acute{\mathsf{x}}|\Theta_i)}{\partial \Theta_i} \right\rangle_{Q^\infty} \quad (4.10)$$

where $\hat{\Theta} = [\Theta_1, \Theta_2, \ldots, \Theta_m]$ denotes parameters of the PoE model, $Q^0 \stackrel{\text{def}}{=} p(\mathsf{x})$, and $Q^\infty \stackrel{\text{def}}{=} p(\mathsf{x}|\hat{\Theta})$.

In general, this derivative is difficult to evaluate by sampling techniques. Some sampling methods (e.g., rejection sampling) are computationally infeasible for the large number of experts m and high-dimensional data spaces. The Markov Chain Monte Carlo method, which is significantly more computationally efficient for modeling a PoE distribution, leads to sampling from the equilibrium distribution[3] of the Markov Chain from all over the entire model, and may result in very high sample variance. Additional problems may occur due to the fact that the parameters may diverge from the regions of high variance even if they result in [sub]-optimal values of the likelihood. This leads to difficulties in estimating derivatives of the PoE renormalizing factor and may result in wrong approximation of the gradient of the log-likelihood and incorrect learning in PoE models.

Therefore, it would be desirable to perform some other type of learning, which would result in the same optimal solutions as optimization of the log-likelihood $\mathcal{L}$, but would not require sampling from Q^∞.

4.3.2 Contrastive Divergence Learning Rule

This presentation is based on the analytic results of Hinton (2000).

From (4.10) it can be shown that maximizing the log-likelihood $\mathcal{L}$, averaged over the data, is equivalent to minimizing the KL divergence[4] $p(\mathsf{x})||p(\mathsf{x}|\hat{\Theta})$

[3]In order to get the unbiased estimate of the derivative $\partial \mathcal{L}_x / \partial x$, the Markov Chain, modeling the PoE-generated data, should converge to its equilibrium density [see Hinton (2000)].

[4]KL divergence for densities P(x) and Q(x) is a measure of the distance between the

between the data and the equilibrium distributions:

$$\left\langle \frac{\partial \log p(\mathsf{x}|\hat{\Theta})}{\partial \Theta_i} \right\rangle_{Q^0} = -\frac{\partial Q^0 \| Q^\infty}{\partial \Theta_i}. \tag{4.11}$$

Hinton claims that instead of minimizing the KL divergence between the data and the target distributions $Q^0\|Q^\infty$, it is sufficient to minimize the difference $Q^0\|Q^\infty - Q^1\|Q^\infty$, where Q^1 is the distribution of the visible variables reconstructed after 1 iteration of Gibbs sampling [see Appendix A.5].

The gradient descent on the log-likelihood in the parameter space may therefore result in the parameter update

$$\begin{aligned} \Delta\Theta_i^{(1)} &\propto -\frac{\partial Q^0\|Q^\infty - Q^1\|Q^\infty}{\partial \Theta_i} \\ &\propto \left\langle \frac{\partial \log p_i(\mathsf{x}|\Theta_i)}{\partial \Theta_i} \right\rangle_{Q^0} - \left\langle \frac{\partial \log p_i(\acute{\mathsf{x}}|\Theta_i)}{\partial \Theta_i} \right\rangle_{Q^1} + e. \end{aligned} \tag{4.12}$$

Here index 1 in the subscript of the rightmost term of the r.h.s. and superscript of the l.h.s. indicates sampling from Q^1, and $e = (\partial Q^1/\partial \Theta_i) \cdot \partial(Q^1\|Q^\infty)/\partial Q^1$.

Hinton (2000) shows experimentally that the term e in (4.12) is small and can safely be ignored, and that the variance of $\Delta\Theta_i^{(1)}$ is generally a very small non-negative number that is equal to 0 if the equilibrium distribution models the data perfectly. Then on the basis of simulations he claims that optimization of the approximate gradient of the contrastive divergence overcomes the problem of the high variance of the modeled gradient and decreases computational expense of sampling from the equilibrium distribution Q^∞, while maintaining good approximation of the data density.

two densities expressed as $\mathrm{KL}(P\|Q) = \int p(\mathsf{x}) \log \frac{p(\mathsf{x})}{q(\mathsf{x})} \mathrm{d}\mathsf{x}$. We will write this as $P\|Q$ for simplicity of notation.

4.3.3 Contrastive Divergence and Boltzmann Learning

Let w be the weight between a hidden variable z with a visible variable x, and let the contrastive divergence learning be performed for a single fixed sample point. Then by analogy with Boltzmann learning in a RBM (4.8), the approximate weight update [(4.12) with e assumed to be equal to 0] may be re-written as

$$\Delta w_{01} \propto \langle xz \rangle_{p(z_0|x_0)} - \langle xz \rangle_{p(z_1|x_1)}, \tag{4.13}$$

where x_1 is the 1-step reconstruction of the data point, z_1 is sampled from $p(z_1|x_1)$ with x_1 fixed, and Δw_{01} indicates the update of the weight between x and z after 1 iteration of the Gibbs sampler from the fixed initial point x_0. For simplicity of notation we will drop $p(z_i|x_i)$ in the terms' subscripts, and write the iteration step i as the index of the variables inside the $\langle \ldots \rangle$ term.

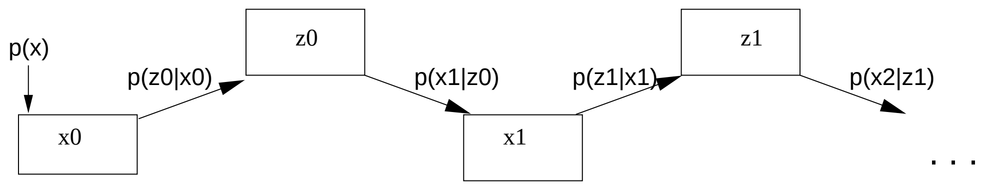

Figure 4.1: Gibbs sampling.

Ignoring the constant proportionality factor, we can extend (4.13) and define the weight update after i-step Gibbs sampling from the fixed initial point:

$$\Delta w_{0i} = \langle x_0 z_0 \rangle - \langle x_i z_i \rangle = \langle \Delta_{0i} \rangle. \tag{4.14}$$

Here Δ_{0i} is defined as

$$\Delta_{0i} = x_0 z_0 - x_i z_i, \tag{4.15}$$

and again, x_i is the i^{th} reconstruction of the data point, z_i is sampled from the posterior $p(z_i|x_i)$ [see figure 4.1]. We will refer to Δ_{0i} as the *update term*.

We can see that for $i \to \infty$ the weight update (4.14) for a single point is analogous to that in Boltzmann learning for restricted Boltzmann machines[5]:

$$\Delta w^{BM} = \langle \Delta^{BM} \rangle = \langle x_0 z_0 \rangle - \langle x_\infty z_\infty \rangle, \tag{4.16}$$

where $\Delta^{BM} = x_0 z_0 - x_\infty z_\infty$.

We expect that the parameter update averaged over the data points in the i-step contrastive divergence approaches that of the Boltzmann learning as the number of iterations i of Gibbs sampling increases:

$$\lim_{i \to \infty} \langle \langle \Delta_{0i} \rangle \rangle_{x_0} = \langle \langle x_0 z_0 \rangle - \langle x_\infty z_\infty \rangle \rangle_{x_0} = \langle \langle \Delta^{BM} \rangle \rangle_{x_0}, \tag{4.17}$$

where $\langle \dots \rangle_{x_0}$ indicates averaging over the data points. We also expect that the variance $var(\Delta_{0i})$ increases with the growth of i (i.e. as Q^i approaches the equilibrium distribution Q^∞).

4.4 Contrastive Learning in a Gaussian Model

In this section we will consider a linear Gaussian 1-D 1-factor latent variable model and the PoE learning rule to estimate the model weight update Δw. The model was initially discussed in private communication with Williams. The results for estimation of the mean and the variance of $p(\Delta w)$ have been obtained by two independent methods and lead to two identical solutions, so we discuss only one derivation.

4.4.1 The Model Description

Consider a linear Gaussian hidden variable model $\mathsf{x} = \mathsf{Wz} + \epsilon \subset \mathcal{R}^d$, where $\mathsf{z} \sim N(\mathsf{0}, \mathsf{C}_Z) \subset \mathcal{R}^m$, and $\epsilon \sim N(\mathsf{0}, \Psi) \subset \mathcal{R}^d$ is the z-independent Gaussian noise. It can be shown that the inverse covariance of the Gaussian model is

[5]See (Hinton, 2000, pp. 7-8) for discussion of correspondence between PoE and RBM.

expressed as

$$\mathsf{C}^{-1} = \begin{bmatrix} \tilde{\mathsf{C}}_{zz} & \tilde{\mathsf{C}}_{zx} \\ \tilde{\mathsf{C}}_{xz} & \tilde{\mathsf{C}}_{xx} \end{bmatrix} = \begin{bmatrix} \mathsf{W}^T\Psi^{-1}\mathsf{W} + \mathsf{C}_Z^{-1} & -\mathsf{W}^T\Psi^{-1} \\ -\Psi^{-1}\mathsf{W}^T & \Psi^{-1} \end{bmatrix} \subset \mathcal{R}^{d\times d}. \tag{4.18}$$

For the simplest case of 1-D 1-factor model with the Gaussian noise and equal variances of the hidden and visible variables, (4.18) can be rewritten as

$$\mathsf{C}^{-1} = \begin{bmatrix} \alpha_0 & -\omega_0 \\ -\omega_0 & \alpha_0 \end{bmatrix}. \tag{4.19}$$

Applying rules for inverting partition matrices (Press et al., 1992, p. 77), we obtain

$$\mathsf{C} = \frac{1}{a_0^2 - w_0^2} \begin{bmatrix} \alpha_0 & \omega_0 \\ \omega_0 & \alpha_0 \end{bmatrix} = \begin{bmatrix} a & w \\ w & a \end{bmatrix}, \tag{4.20}$$

where $a = \alpha_0/(\alpha_0^2 - \omega_0^2),\ w = \omega_0/(\alpha_0^2 - \omega_0^2),\ |w| < a$.

In section 4.2.4 we have seen that the inverse covariance C^{-1} of the complete data in the Gaussian model corresponds to the weight matrix of a Boltzmann machine. If we are free to change parameter ω_0 in C^{-1}, this change will affect the model covariance C. The task of learning may be seen as finding a parameter ω_0 that changes variances a in such a way that the model, defined by the covariance C, fits the data. If $a_t = var(x_0)$ is the variance of the data distribution $p(x_0)$, we expect that

$$\lim_{i\to\infty} a^{(i)} = a_t, \tag{4.21}$$

where i is the number of iterations of the PoE learning rule, described in section 4.3.2. We will refer to a_t as the *target* variance.

4.4.2 Gibbs Sampling in the Gaussian Case

Here we will extend the 1-step PoE learning rule (1-step Gibbs sampling) to i-step Gibbs sampling for estimation of the PoE distribution in the Gaussian case.

Let x and z be a one-dimensional visible and hidden variable respectively. As before, we will write x_i and z_i to denote reconstructed visible and hidden variables after i iterations of Gibbs sampling.

Figure 4.1 of section 4.3.3 presents a diagram showing the Gibbs sampling procedure [see Appendix A.5 for a brief discussion of Gibbs sampling and MCMC]. Note that in Gibbs sampling each random variable is sampled from its conditional distribution with all other variables fixed at their current values. For example, x_1 is sampled from $p(x_1|z_0)$, where z_0 is fixed. For the 1-D Gaussian case we can note that if $p(x_1|z_0) \sim N(\mu_{x_1|z_0}, \sigma^2_{x_1|z_0})$, then $x_1 = \mu_{x_1|z_0} + \sigma_{x_1|z_0} u$, where $u \sim N(0, 1)$.

In order to proceed with estimation of the PoE distribution, we need to estimate conditional distributions of the model. Using the results of pp. [426-228] of Mises (1964), the conditional probabilities of partition matrices can be expressed as

$$p(\mathsf{x}_1|\mathsf{x}_2) \sim N(\mu_1 + \mathsf{S}_{12}\mathsf{S}_{22}^{-1}(\mathsf{x}_2 - \mu_2), \mathsf{S}_{11} - \mathsf{S}_{12}\mathsf{S}_{22}^{-1}\mathsf{S}_{21}), \tag{4.22}$$

where

$$\mathsf{S} = \begin{bmatrix} \mathsf{S}_{11} & \mathsf{S}_{12} \\ \mathsf{S}_{21} & \mathsf{S}_{22} \end{bmatrix} \tag{4.23}$$

is the covariance of the complete model.

From (4.20) and (4.22) the conditional distributions of the considered model are expressed as

$$p(x|z) \sim N(\sigma z, \tau^2) \tag{4.24}$$

$$p(z|x) \sim N(\sigma x, \tau^2), \tag{4.25}$$

with the auxiliary parameters

$$\sigma \stackrel{\text{def}}{=} w/a \tag{4.26}$$

$$\tau^2 \stackrel{\text{def}}{=} (1 - \sigma^2)a. \tag{4.27}$$

Notice that for the covariance of the complete model to be positive-definite σ should belong to the open interval $(-1, 1)$.

Let x_0 be a sample from the data distribution $p(x) \sim N(0, a_t)$. If $v_1, v_2, \ldots$ and $u_0, u_1, \ldots$ are independent uniform Gaussians, then on the first iteration of the Gibbs sampling procedure samples from $p(z_0|x_0)$ and $p(x_1|z_0)$ can be expressed as

$$\begin{aligned} z_0 &= \sigma x_0 + \tau u_0 && (4.28) \\ x_1 &= \sigma z_0 + \tau v_1. && (4.29) \end{aligned}$$

Here we have chosen the indices of u and v to correspond to indices of z and x respectively[6]. In general

$$\begin{aligned} z_i &= \sigma x_i + \tau u_i,\ i \geq 0 && (4.30) \\ x_j &= \sigma z_{j-1} + \tau v_j,\ j \geq 1. && (4.31) \end{aligned}$$

4.4.3 Estimation of the Moments for Parameter Updates

In this section we will estimate the mean and the variance of the distribution of the update term Δ_{0i} in i-step PoE learning.

Estimation of the Mean

Expression (4.30) leads to

$$\begin{aligned} & x_i z_i = \sigma x_i^2 + \tau u_i x_i \\ \Rightarrow\ & \langle x_i z_i \rangle = \sigma \langle x_i^2 \rangle. && (4.32) \end{aligned}$$

[6]Notice that the expectations of v and u satisfy $\langle u_i v_k \rangle = 0$, $\langle u_i u_k \rangle = \langle v_i v_k \rangle = \delta_{ik}$, where δ_{ik} is the Kronecker delta.

It can be found [see Appendix D] that

$$\begin{aligned}\langle x_i^2 \rangle &= \sigma^{4i} x_0^2 + \frac{\tau^2(\sigma^{4i}-1)}{\sigma^2-1} \\ &= \sigma^{4i}(x_0^2 - a) + a. \end{aligned} \tag{4.33}$$

Then (4.32) and (4.33) lead to

$$\langle x_i z_i \rangle = \sigma^{4i+1}(x_0^2 - a) + a\sigma, \tag{4.34}$$

and the mean of the update term Δ_{0i} under the PoE learning, averaged over the data distribution, results in

$$\boxed{\langle\langle \Delta_{0i} \rangle\rangle_{x_0} = \sigma(a_t - a)(1 - \sigma^{4i})}. \tag{4.35}$$

From (4.34) we can find the weight update in Boltzmann learning $\langle \Delta^{BM} \rangle$, averaged over data points

$$\begin{aligned}\langle\langle \Delta^{BM} \rangle\rangle_{x_0} &= \langle\langle x_0 z_0 \rangle - \langle x_\infty z_\infty \rangle\rangle_{x_0} \\ &= \langle \sigma(x_0 - a) + a\sigma \rangle_{x_0} - a\sigma \\ &= \sigma(a_t - a), \end{aligned} \tag{4.36}$$

where we assumed that $\langle x_\infty z_\infty \rangle = w = a\sigma$ from the definition of the model covariance (4.20).

Comparing the weight updates under the PoE and Boltzmann learning procedures and considering the fact that $|\sigma| < 1$, we see that

$$\begin{aligned}\lim_{i\to\infty} \langle\langle \Delta_{0i} \rangle\rangle_{x_0} &= \lim_{i\to\infty} (a_t - a)\sigma(1 - \sigma^{4i}) \\ &= \sigma(a_t - a). \end{aligned} \tag{4.37}$$

This result agrees with expression (4.36) for $\langle\langle \Delta^{BM} \rangle\rangle_{x_0}$ and meets our original expectations. Thus, the gradient of the log-likelihood in i-step PoE learning approaches that in the learning rule for restrictive Boltzmann machines with the growth of the number of Gibbs sampling iterations i.

Let $\mathcal{L}^{BM}$ and $\mathcal{L}^{(i)}$ denote the log-likelihood in the restricted Boltzmann machines and its approximation by i-step Gibbs sampling, used in minimizing contrastive divergence. Then

$$\frac{\partial \mathcal{L}^{BM}}{\partial w} = 0 \quad \Rightarrow \quad \sigma(a_t - a) \Rightarrow a_{ML} = a_t. \tag{4.38}$$

In i-step PoE learning we obtain exactly the same maximum-likelihood solution:

$$\frac{\partial \mathcal{L}^{(i)}}{\partial w} = 0 \Rightarrow \sigma(a_t - a)(1 - \sigma^{4i}) \Rightarrow a_{ML} = a_t. \tag{4.39}$$

Thus, the optimal solutions for the weights are identical in PoE and Boltzmann learning and result in equality of the model and target variances.

From (4.38) and (4.39) we can further notice that since $|\sigma| < 1$

$$\left|\frac{\partial \mathcal{L}^{(i)}}{\partial w}\right| < \left|\frac{\partial \mathcal{L}^{BM}}{\partial w}\right|. \tag{4.40}$$

Thus, we can see that the estimation of the absolute value of the gradient by minimizing contrastive divergence underestimates the absolute value of the true gradient and increases with the growth of the number of iterations of the Gibbs sampling.

Estimation of the Variance

In a PoE model we can think of at least two sources of variance for the update term Δ_{0i}. First, variation may be possible due to the different data points x_0, sampled form the data distribution $p(x)$. The other source of variance is stochasticity of the reconstruction procedure in Gibbs sampling. Here we estimate the variance of Δ_{0i} due to the stochastic reconstruction, and average it over the data distribution.

Recall that $x_0 z_0$ and $x_i z_i$ are evaluated at the positive and negative stages of learning. If we assume that the positive and the negative phases are started at different points[7] then $x_i z_i$ is independent from $x_0 z_0$, and the variance of the update parameter may be expressed as

$$\begin{aligned} var(\Delta_{0i}) &= var(x_0 z_0 - x_i z_i) \\ &= var(x_0 z_0) + var(x_i z_i). \end{aligned} \tag{4.41}$$

The leftmost term of the r.h.s. is calculated as

$$\begin{aligned} var(\langle x_0 z_0 \rangle) &= \langle (x_0 z_0 - \langle x_0 z_0 \rangle)^2 \rangle \\ &= \langle (\sigma x_0^2 + \tau u_i x_0 - \sigma x_0^2) \rangle \\ &= \langle (\tau u_i x_0)^2 \rangle \\ &= (1 - \sigma^2) a x_0^2. \end{aligned} \tag{4.42}$$

We can now consider the term $var(x_i z_i)$.

The variance $var(x_i z_i)$ can be expressed as

$$var(x_i z_i) = \langle (x_i z_i)^2 \rangle - (\langle x_i z_i \rangle)^2. \tag{4.43}$$

The expression for $\langle x_i z_i \rangle$ has been derived above and is shown in (4.34). Expanding $\langle (x_i z_i)^2 \rangle$, we obtain

$$\begin{aligned} \langle (x_i z_i)^2 \rangle &= \sigma^2 \langle x_i^4 \rangle + 2\sigma\tau \langle x_i^3 u_i \rangle + \tau^2 \langle u_i^2 x_i^2 \rangle \\ &= 3\sigma^2 (\langle x_i \rangle)^2 + a(1 - \sigma^2) \langle x_i^2 \rangle, \end{aligned} \tag{4.44}$$

where we have calculated the higher-order moments by using the formula (Chacon and Iranpour, 1988)

$$\begin{aligned} \langle x_1 x_2 x_3 x_4 \rangle &= \langle x_1 x_2 \rangle \langle x_3 x_4 \rangle \\ &+ \langle x_1 x_3 \rangle \langle x_2 x_4 \rangle + \langle x_1 x_4 \rangle \langle x_2 x_3 \rangle. \end{aligned} \tag{4.45}$$

[7]Notice that if the both stages of learning are started at the same point then z_1 depends on x_0 and (4.41) does not hold.

Substituting (4.44) into (4.43) and (4.41) and calculating expectations w.r.t. the posterior and the data distributions, we obtain

$$\begin{aligned}\langle var(\Delta_{0i})\rangle_{x_0} &= 2\sigma^2(\langle x_i^2\rangle)^2 + a(1-\sigma^2)\langle x_i^2\rangle \\ &= 2\sigma^2(3a_t^2 - 2aa_t + a^2)k^2 \\ &- a(a-a_t)(1+3\sigma^2)k + a^2 + a^2\sigma^2 + (1-\sigma^2)aa_t, \quad (4.46)\end{aligned}$$

where $k = \sigma^{4i} \subset (0,1]$.

Let $f(k)$ be the r.h.s. of (4.46). The graph of $f(k)$ is a convex parabola. If k_{min} denotes the vertex of the parabola [the minimum of $f(k)$], we will distinguish three types of behaviour for the behaviour of the variance of the update term, averaged over the data distribution.

Case 1: $k_{min} \geq 1$.

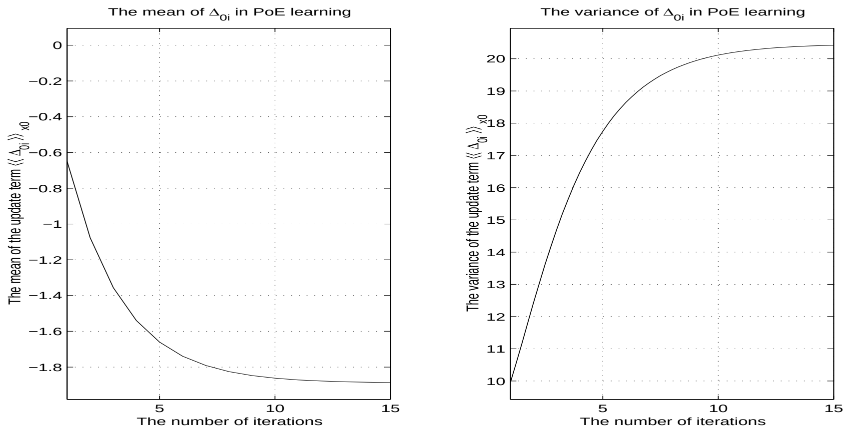

Figure 4.2: Typical mean and variance of the update parameter for $k_{min} \geq 1$.

In this case $f(k)$ monotonically decreases as k changes from 0 to 1, i.e. the variance $\langle var(\Delta_{0i})\rangle_{x_0}$ monotonically increases with the growth of the number i of Gibbs sampling iterations. Figure 4.2 illustrates the typical curves for the mean and the variance of the update parameter in this situation.

Case 2: $0 < k_{min} < 1$. For $0 < k_{min} < 1$, the variance $\langle var(\Delta_{0i})\rangle_{x0}$ decreases for a finite number of steps, and starts increasing after i_+ iterations of Gibbs sampling, where

$$i_+ = \frac{1}{4}\left\lceil \frac{\ln k_{min}}{\ln \sigma} \right\rceil . \tag{4.47}$$

The case is illustrated on figure 4.3.

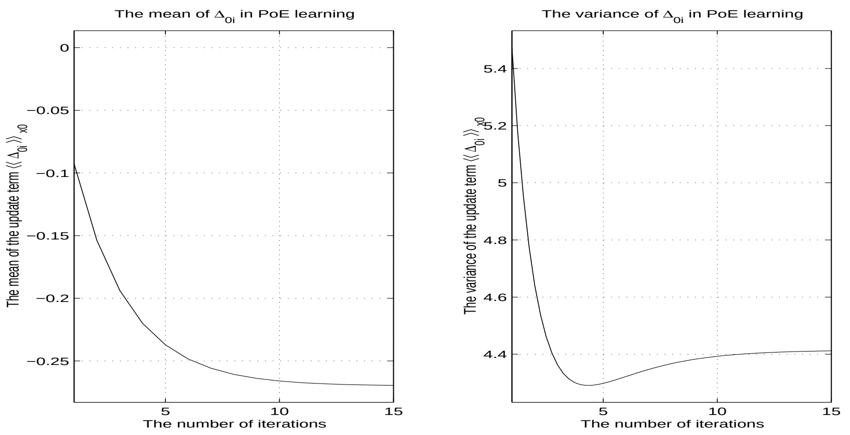

Figure 4.3: Typical mean and variance of the update parameter for $0 < k_{min} < 1$.

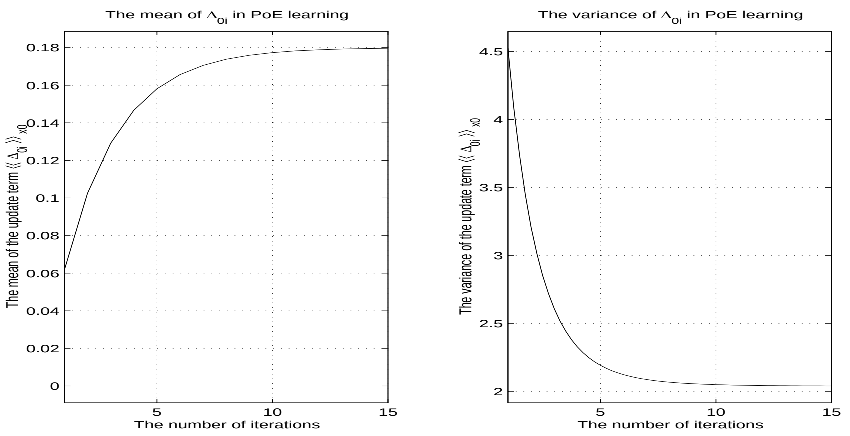

Figure 4.4: Typical mean and variance of the update parameter for $k_{min} < 0$.

Case 3: $k_{min} \leq 0$. If $k_{min} \leq 0$ the variance monotonically decreases with the growth of i (see figure 4.4). It is easy to see that this case occurs whenever $a \leq a_t$, where a is the variance of the model, and a_t is the sample variance.

At this stage, this result still requires further investigation and better understanding (see section 5.4).

4.5 Summary

In this chapter we have extended the contrastive divergence learning rule (1-step Gibbs sampling) to i steps by using the analytic results of Gibbs sampling for Gaussians. We have estimated the mean and the variance of the weight update as a function of the number of Gibbs sampling iterations i for a simple example of a (1,1) linear Gaussian latent variable model.

It has been shown that in the considered model learning by minimizing contrastive divergence introduces the systematic bias by underestimating the absolute value of the gradient of the log-likelihood. However, it has been found that the ML solutions for the weight in PoE learning by minimizing contrastive divergence and in Boltzmann learning for restricted Boltzmann machines are identical.

We had expected that the variance of the update term would increase with the number of Gibbs sampling iterations (i.e. as the negative phase of learning approaches to sampling from the equilibrium distribution). However, the obtained analytic results for the variance indicate that this is not necessarily the case for all model and data distributions. One of the possible explanations could be the simplicity of the considered model, where a sample from the joint distribution does not necessarily need to have a high variance. This result requires further understanding and experimental confirmation.

Chapter 5

Conclusions

> The end is never as satisfying as the journey. To have achieved everything but to have done so without integrity and excitement is to have achieved nothing.
>
> – Anonymous

5.1 Goals

The main goal of the conducted study was investigation of the Products of Experts model when each expert is a Gaussian. The first two principal parts of the investigation corresponded to derivation and experimental confirmation of maximum-likelihood solutions for PoE parameters in the cases of constrained Gaussian and PPCA experts. The main goal of these parts was to confirm or refute initial hypotheses that the ML solution gives rise to PMCA (probabilistic minor component analysis) and PPCA respectively. The third principal part of the project concerned investigation of i-step contrastive divergence learning for the Gaussian case, and comparison of the solutions obtained by minimizing contrastive divergence with the maximum-likelihood solutions of Boltzmann learning in restricted Boltzmann machines.

5.2 This Work in the Context of Previous Studies

Since the PoE paradigm was introduced relatively recently (September 1999), there has been relatively little work conducted in this field. The general idea of PoE and 1-step Gibbs sampling were introduced in Hinton (1999). The work also discusses correspondence of PoE and standard Boltzmann learning, that is described in more detail in Hinton and Brown (2000). Hinton (2000) discusses problems of estimating maximum log-likelihood in PoE and suggests minimization of 1-step contrastive divergence as an alternative to maximizing the likelihood.

Hinton ((Hinton, 1999) and (Hinton, 2000)) claims that approximation of the likelihood by 1-step Gibbs sampling is sufficient for obtaining satisfactory solutions for parameters, and confirms his results experimentally. However, there have been no analytic considerations confirming equality of ML solutions maximizing the exact log-likelihood (in Boltzmann learning) and its i-step approximation (PoE learning, extended to i-step Gibbs sampling). We use the results of Gibbs sampling for Gaussians to verify the claim analytically.

Current research of PoE architectures includes investigations of products of Hidden Markov Models (Hinton and the UCL Gatsby group), and products of tree-structured Gaussians (Williams and Felderhof, 2000). It appears that the architectures investigated in this work have not yet been studied elsewhere.

5.3 Results

The analytic derivations of the maximum-likelihood solutions for parameters of the product of constrained Gaussians (PoCG) have confirmed the initial hypothesis that the PoCG gives rise to minor component analysis. We have also confirmed by experiments that the derived solutions correspond to the ones obtained by applying higher-order non-linear optimization techniques on the log-likelihood.

Thus, a $(1, m)$ PoCG model can be viewed as a probabilistic minor component analysis (MCA) model that can be used, for example, for dimensionality reduction, data visualization, and filtering applications. 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 minor component analyzers (which will be modeled as a mixture of products of constrained Gaussians), and leads to a number of other important advantages of PoCG over its non-probabilistic analogues.

For the case of $(1, m)$ PoPPCA we expected that the model would be equivalent to an m-factor PPCA. The equivalence would mean that the maximum-likelihood solution for the expert weight matrix in $(1, m)$ PoPPCA would correspond to m rotated and scaled principal components of the sample covariance matrix, and that both $(1, m)$ PoPPCA and m-PPCA would have the same representational power.

It was shown analytically and confirmed experimentally that for certain types of data the ML solution for the expert weight matrix in PoPPCA indeed corresponds to the PPCA solution. However, it was also found that for other data sets the unconstrained maximum-likelihood solution gives rise to imaginary expert weights, indicating the limitations of a $(1, m)$ PoPPCA model and its lower representational power compared with an m-PPCA at least for some sample covariance matrices. These problematic data sets may be modeled by increasing the number of non-zero weights in the expert weight matrix. For example, it was shown that one way to model data generated by an arbitrary 1-PPCA is by using a $(1, m, m)$ PoPPCA model with collinear weights, - the "team of experts" solution.

Finally, we have extended the contrastive divergence learning rule (1-step Gibbs sampling for minimizing contrastive divergence) to i steps by using the analytic results of Gibbs sampling for Gaussians. We have also estimated the mean and the variance of the weight update as a function of the number of Gibbs sampling iterations i for a simple example of a 1-dimensional 1-factor linear Gaussian latent variable model.

It was found that for the analyzed model approximation of the absolute value of the log-likelihood in i-step CD learning underestimates the absolute value of the true gradient of the log-likelihood. However, it was shown that the ML solutions for the weights in the CD learning, maximizing approximation of the log-likelihood, are identical to those in standard Boltzmann learning.

We expected that the variance of the update term will increase with the number of Gibbs sampling iterations (i.e. as the negative phase of learning approaches to sampling from the equilibrium distribution). However, the obtained analytic results for the variance indicate that this is not necessarily the case for all model and data distributions. This result requires further understanding and experimental confirmation.

5.4 Assumptions and Limitations

PoE models perform very well in fitting localized patterns in the data, e.g. when the data can be represented as an intersection of data sets generated by expert distributions in the data space or a product of lower-dimensional constraints. However, they are expected to be less efficient than mixture models in modeling smooth distributions and fitting global data patterns (for example, when the data comes from a superposition of wide ridges). While mixture models may result in inaccurate approximation of the data distribution by generating irrelevant data, product of experts may fail to generate some relevant data points.

One way to avoid this problem is by considering a mixture of products of experts, which might be considered to be a probabilistic equivalent of disjunctive normal forms (disjunction of conjunctions) in mathematical logic. It is expected that by adjusting mixing coefficients mixtures of PoE models may be trained to fit both local and global patterns in the data set.

In this work we have considered specific types of PoE models where each expert is a Gaussian. In this case the resulting PoE distribution is also a Gaussian, and cannot be used efficiently for modeling arbitrary probability densities. However, many probabilistic processes arising in practical tasks may be explained by Gaussian-type covariance matrices. We have also shown that products of Gaussians may result in probabilistic analogues of non-probabilistic models (i.e. minor component analysis), what offers a wide range of advantages for their use in practical applications.

For simplification purposes and due to the time constraints, we have made several assumptions in the considered models. First of all, we have considered only 1-dimensional experts. Future studies can probably consider PoE models with experts of different dimensionalities. Nevertheless, the considered model seems to be reasonable if each expert imposes a constraint in a single dimension.

In the discussion of PoCG we have assumed that all the experts have equal means and variances in the directions of elongation of probability contours. In this case contours of equiprobable surfaces of the resulting distributions correspond to the intersection region of disk-shaped Gaussians centered at the same point and having different "thicknesses", but the same "diameters". The natural extension of the model would be allowing different experts to have different means and distortion variances. In this case, the probabilistic MCA is believed to be a special case of a more general ML solution.

The analogous assumption of equal noise variances was also made in PoPPCA. It was shown that in this case PoPPCA models cannot generate all the date generated by a 1-PPCA. However, allowing experts to have different noise variances [in the directions of contraction] is believed to increase the representational power of the models. For example, in fitting a 1-PPCA data with a $(1, m, 1)$ PoPPCA model, the noise variances of $m-1$ discarded experts [i.e. experts weighted by the factor of 0] may be be trained to become close to zero, so that a $(1, m, 1)$ PoPPCA model becomes similar to a $(1, 1, 1)$ PoPPCA and fits the target distribution exactly.

In discussion of Gibbs sampling in contrastive divergence learning, we have considered a simple 1-dimensional 1-factor linear Gaussian model. Due to the limits of time and space, the analytic results have not yet been confirmed experimentally. Dependence of the variance of the update term on the difference between the model and the target variances also requires better understanding. Moreover, the considered toy problem was too simple to have significant practical use. However, the derived results show that in the model we have considered optimization of the contrastive divergence and optimization of the maximum-likelihood lead to similar solutions.

At this stage we should perhaps mention that most of the suggestions of extending the model go beyond the proposed goals of the suggested work. However, they define the direction for the future research in the field.

5.5 Future Work

In the intended future investigation of PoE architectures we are planning to address some of the limitations described in the previous section. First of all, we are planning to confirm by experiments the analytical results of i-step Gibbs sampling for the simple latent variable model considered in chapter 4. We also plan to consider models with a higher number of factors and dimensions and see whether i-step Gibbs sampling is also sufficient in these cases.

To complete investigations of $(1, m)$ PoCG and PoPPCA models, we hope to extend them for the case of non-equal noise variances (see section 5.4). It will also be interesting to continue investigation of the PoPPCA expert weight matrices obtained by applying non-linear optimization techniques for the cases when the sample covariance results in imaginary weights of the analytic ML solution.

Finally, it will be very interesting to consider mixtures of products of Gaussians and other expert architectures [e.g. Kalman filters] that can be used for modeling dynamic data. Extension of PoE models and their mixtures to dynamic tasks will promote their use in prediction problems and is hoped to give rise to a new wide range of their theoretical and practical applications.

Bibliography

Ackley, Hinton and Sejnowski (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, (9):147 – 169.

Bartholomew, D. (1987). *Latent Variable Models and Factor Analysis.* Charles Griffin and Co. Ltd., London.

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

Bishop, C. and Tipping, M. (1999). Latent variable models and data visualization. In Kay, J., editor, *Statistics and Neural Networks: Advances at the Interface*. Oxford University Press.

Bishop, Svensen and Williams (1998). Developments of the Generative Topographic Mapping. *Neurocomputing*, (21).

Brookes, M. (2000). *Matrix Reference Manual.* Electrical and Electronic Engineering Department, Imperial College, http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html.

Chacon, P. and Iranpour, R. (1988). *Basic Stochastic Processes: The Mark Kac Lectures.* Macmillan.

Day, N. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56(3):463 – 474.

Dempster, A. P., Laird, M., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1).

Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis.* John Wiley, New York.

Freund, Y. and Haussler, D. (1992). Unsupervised learning of distributions on binary vectors using two layer networks. In J.E. Moody, S.J. Hanson, R. L., editor, *Advances in Neural Information Processing Systems*, volume 4. Morgan Kaufmann: San Mateo, CA.

Ghahramani, Z. and Hinton, G. (1996). The EM Algorithm for Mixtures of Factor Analyzers. CRG-TR-96-1, Department of Computer Science, University of Toronto.

Haykin, S. (1999). *Neural Networks.* Prentice Hall.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In Jordan, M., editor, *Learning in Graphical Models*, chapter 3. MIT Press.

Hertz, Krogh, and Palmer (1991). *Introduction of the Theory of Neural Computation.* MA: Addison-Wesley Publishing Company.

Hinton, G. (1999). Products of experts. *Proceedings of the Ninth International Conference on Artificial Neural Networks [ICANN 99]*, 1.

Hinton, G. (2000). Training products of experts by minimizing contrastive divergence. GCNU TR 2000-004, Gatsby Computational Neuroscience Unit, UCL.

Hinton, G. and Brown, A. (2000). Spiking Boltzmann Machines. *To appear in Advances in Neural Information Processing Systems*, (12).

Krzanowski, W. and Marriott, F. (1994). *Multivariate Analysis Part II: Distributions, Ordination and Inference.* Edward Arnold, London.

Mackay, D. (1998). Introduction to Monte Carlo methods. In Jordan, M., editor, *Learning in Graphical Models*, chapter 1. MIT Press.

Minka, T. (1999). *Old and New Matrix Algebra Useful for Statistics.* Web article: ftp://vismod.www.media.mit.edu/pub/tpminka/papers/minka-matrix.ps.gz.

Mises, R. V. (1964). *Mathematical Theory of Probability and Statistics.* New York: Academic Press.

Moerland, P. (1998). Localized mixtures of experts. IDIAP-RR 14, IDIAP.

Neal, R. (1993). Probabilistic inference using Markov Chain Monte Carlo methods. CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. (2000). Web resource: www.cs.utoronto.ca/~radford/, Dept. of Computer Science, University of Toronto.

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

Press, Teukolsky, Vetterling, and Flannery (1992). *Numerical Recipes in C.* Cambridge University Press.

Reed and Marks (1999). *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks.* MIT Press, Cambridge, MA.

Rice, J. (1995). *Mathematical Statistics and Data Analysis.* Duxbury Press, California.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach.* Prentice Hall.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D.E. Rumelhart, J. M., editor, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press.

Thompson, P. (1979). An adaptive spectral analysis technique for unbiased frequency estimation in the presence of white noise. *Proceedings of the 13th Asilomar Conference on Circuits, Systems, And Computers*, 1:529 – 533.

Tipping, M. (1999). *Statistical Pattern Analysis.* Unpublished manuscript.

Tipping, M. and Bishop, C. (1997). Probabilistic principal components analysis. NCRG/97/010 TR, NCRG, Dept. of Computer Science and Applied Mathematics, Aston University.

Titterington, S. and Makov (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley, New York.

Tomasi, C. (1999). *Mathematical Methods for Robotics and Vision.* Stanford University, www.stanford.edu/class/cs205/notes/book/book.html.

Weisstein, E. (2000). *World of Mathematics.* Wolfram Research, http://mathworld.wolfram.com.

Williams, C. (1999a). Course material for PMR. Course material, Division of Informatics, University of Edinburgh.

Williams, C. (1999b). Factor analysis and beyond. Web article: www.dai.ed.ac.uk/dai/teaching/modules/pmr/docs/docs/faclex.ps.gz, Division of Informatics, University of Edinburgh.

Williams, C. and Felderhof, S. (2000). Products of tree-structured Gaussian processes. NIPS *00 submission, Division of Informatics, University of Edinburgh.

Appendix A

The Fundamentals

A.1 Basic Definitions

Following is a very brief review of probability theory and definitions, used throughout the thesis. More detailed discussions are available in numerous books on probability theory and statistics. This review is related to (Tipping, 1999, pp.11-15).

Probability:

$$\begin{aligned} \sum_{\mathsf{x}} P(\mathsf{x}) &= 1 - \text{discrete case} \\ \int_{\infty}^{\infty} p(\mathsf{x})\mathrm{d}\mathsf{x} &= 1 - \text{continuous case.} \end{aligned}$$

Sum rule:

$$\begin{aligned} p(\mathsf{x}) &= \sum_{\mathsf{z}} p(\mathsf{x},\mathsf{z}) - \text{discrete case} \\ p(\mathsf{x}) &= \int p(\mathsf{x},\mathsf{z})\mathrm{d}\mathsf{z} - \text{continuous case} \end{aligned}$$

$p(\mathsf{x})$ is referred to as *marginal* density. In general, marginal densities are obtained by integrating over some variables. The process of such integration is referred to as *marginalization*.

Product rule:

$$p(\mathsf{x}, \mathsf{z}) \;=\; p(\mathsf{x}|\mathsf{z})p(\mathsf{z}) = p(\mathsf{z}|\mathsf{x})p(\mathsf{x})$$

$p(\mathsf{x}, \mathsf{z})$ is referred to as *joint* probability density. If z represents latent variables, $p(\mathsf{x}, \mathsf{z})$ is a density over the *complete* data. $p(\mathsf{x}|\mathsf{z})$ and $p(\mathsf{z}|\mathsf{x})$ are *conditional* densities, which will be referred to as *conditional* and *posterior* respectively.

Bayes rule:

$$p(\mathsf{z}|\mathsf{x}) = \frac{p(\mathsf{x}|\mathsf{z})p(\mathsf{z})}{p(\mathsf{x})}$$

Mathematical expectation:

Expressions

$$\begin{aligned} E_{g(\mathsf{x})}[f(\mathsf{x})] &= \sum_{\mathsf{x}} g(\mathsf{x})f(\mathsf{x}) - \text{discrete case} \\ E_{g(\mathsf{x})}[f(\mathsf{x})] &= \int g(\mathsf{x})f(\mathsf{x})\mathrm{d}\mathsf{x} - \text{continuous case} \end{aligned}$$

represent mathematical expectation of $f(\mathsf{x})$ w.r.t. $g(\mathsf{x})$. For example if z are latent and x are visible variables,

$$E_{p(\mathsf{z}|\mathsf{x})}[f(\mathsf{x}, \mathsf{z})] \;=\; \sum_{\mathsf{z}} p(\mathsf{z}|\mathsf{x})f(\mathsf{x}, \mathsf{z})$$

represents expectation of a function of the complete data w.r.t. posterior probabilities (*cf* the E-step of the EM algorithm).

A.2 The Gaussian Distribution

The Gaussian distribution has many convenient properties and is widely used for modeling of many common processes occurring in everyday's life (e.g. distribution of the weight, income, number of children, etc.)

Density function of the Gaussian distribution $N(\mu, \mathbf{C})$ is defined as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{C}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{C}^{-1}(\mathbf{x}-\mu)\right\} \tag{A.1}$$

and has a bell-shaped curve.

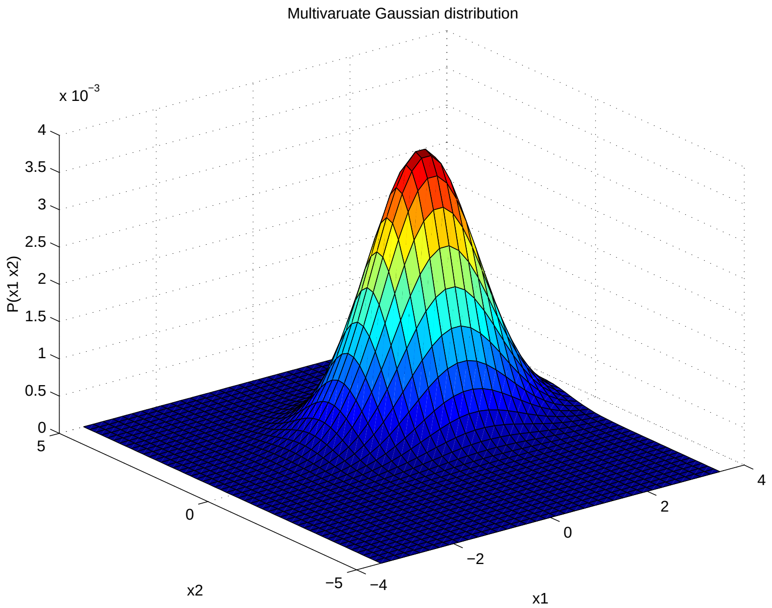

Figure A.1: A two-dimensional Gaussian distribution

Parameters of the Gaussian distribution $N(\mu, \mathsf{C})$ are the mean μ and the covariance matrix C, defined as

$$\mu = E[\mathsf{x}] \tag{A.2}$$

$$\mathsf{C} = E[(\mathsf{x}-\mu)(\mathsf{x}-\mu)^T] \subset \mathcal{R}^{d\times d}. \tag{A.3}$$

Here d is the dimension of the input space, and $E[\ldots]$ denotes mathematical expectation as defined in Appendix A.1.

From (A.3) it is easy to see that $\mathsf{C} = \mathsf{C}^T$, i.e. the covariance matrix of a Gaussian distribution is *symmetric*. Moreover, in order for (A.1) to converge, C^{-1} (and, consequently, C) should be *positive-definite*, i.e.

$$\mathsf{x}\mathsf{C}\mathsf{x}^T > 0$$

should hold for all $\mathsf{x} \in \mathcal{R}^d$. In other words, the covariance matrix should have positive eigenvalues.

This latter fact can be understood from the geometric interpretation of probability contours. Eigenvectors of the sample covariance matrix correspond to directions of the radii of a hyper-ellipsoid, defining probability contours in the data space. Radii of the hyper-ellipsoids equal to multiples of square roots of eigenvalues of the covariance matrix and define the variance in directions of the corresponding eigenvector. The center of the ellipse is defined by the mean vector. The diagram illustrates this for 2 dimensions.

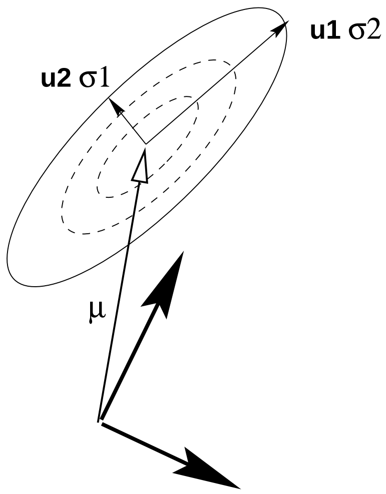

Figure A.2: Contour of a two-dimensional Gaussian distribution.

Since the covariance matrix C is symmetric, eigenvectors corresponding to its non-degenerate eigenvalues are orthogonal. The eigenvalue decomposition[1] of C can be expressed as

$$\mathsf{C} \;=\; \mathsf{U}\Lambda\mathsf{U}^T = \sum_{i=1}^{d} \mathsf{u}_i \mathsf{u}_i^T \lambda_i, \tag{A.4}$$

[1]Eigenvalue decomposition is a special case of singular value decomposition for symmetric matrices. A very good discussion of geometrical properties of the SVD is available online at Tomasi (1999).

where $\mathsf{U} = [\mathsf{u}_1, \mathsf{u}_2, \dots, \mathsf{u}_m] \subset \mathcal{R}^{d \times m}$ is a matrix of eigenvectors of C; $\Lambda = diag(\lambda_1, \lambda_2, \dots \lambda_m) \subset \mathcal{R}^{m \times m}$ is a diagonal matrix of eigenvalues of C. $\lambda_i = \sigma_i^2$ is the variance in the direction defined by u_i.

Properties of Gaussian distributions are described in virtually every book on probability and statistics. See, for example, Rice (1995). Weisstein (2000) also offers very comprehensive on-line reference material.

A.3 Independence of the Hidden States Given the Data in PoE Models

We will show that the hidden states of the PoE model are independent given the data. If $z_1, \dots, z_m$ indicate hidden states of the PoE model, the condition of independence means that

$$p(z_1, \dots, z_m | \mathsf{x}) = \prod_{i=1}^{m} p(z_i | \mathsf{x}) \tag{A.5}$$

and is written as $I(z_1, \dots, z_m | \mathsf{x})$. Derivations of this section are partially based on discussions with Williams.

□ For simplicity of notation we will assume conditioning on parameters of the system and write

$$p(\mathsf{x}) = \frac{1}{Z} \prod_{i=1}^{m} p_i(\mathsf{x}) \tag{A.6}$$

to denote the probability of x under the PoE model. Here Z is some normalizing factor. Probability of the point under each expert is marginalized over the hidden states, i.e.

$$p_i(\mathsf{x}) = \int p(\mathsf{x}|z_i) p(z_i) \mathrm{d}z_i. \tag{A.7}$$

Let us determine how the marginal probability of the model $p(\mathsf{x})$ is expressed from the joint distributions. From the key assumption of the latent variable representation, discussed in section 1.3, observations are independent given the states of the latent variables, and the joint distribution may be written as

$$\begin{aligned} p(\mathsf{x}, z_1, \ldots, z_m) &= \frac{\prod_{i=1}^m p(z_i) p(\mathsf{x}|z_i)}{\acute{Z}} \\ &= p(z_1) p(\mathsf{x}|z_1) \frac{\prod_{i=2}^m p(z_i) p(\mathsf{x}|z_i)}{\acute{Z}}, \end{aligned} \tag{A.8}$$

where $\acute{Z}$ is a normalizing factor of the joint distribution.

Marginalizing over z_1, we obtain

$$\begin{aligned} p(\mathsf{x}, z_2, \ldots, z_m) &= \frac{\prod_{i=2}^m p(z_i) p(\mathsf{x}|z_i)}{\acute{Z}} \int p(z_1) p(\mathsf{x}|z_1) \mathrm{z}_1 \\ &= p_1(\mathsf{x}) \frac{\prod_{i=2}^m p(z_i) p(\mathsf{x}|z_i)}{\acute{Z}}. \end{aligned} \tag{A.9}$$

Analogously, marginalization over $z_2, \ldots, z_m$ leads to the expression for $p(\mathsf{x})$:

$$\begin{aligned} p(\mathsf{x}) &= p_1(\mathsf{x}) \int \cdots \int \prod_{i=2}^m p(z_i) p(\mathsf{x}|z_i) \mathrm{d}z_2 \ldots \mathrm{d}z_m \\ &= \frac{\prod_{i=1}^m p_i(\mathsf{x})}{\acute{Z}}. \end{aligned} \tag{A.10}$$

Comparing (A.10) and (A.6) we see that both the joint and the marginal distributions have the same normalizing factor

$$Z = \acute{Z}. \tag{A.11}$$

Since the posterior $p(z_i|\mathsf{x})$ is expressed as

$$p(z_i|\mathsf{x}) = \frac{p(z_i) p(\mathsf{x}|z_i)}{\int p(z_i) p(\mathsf{x}|z_i) \mathrm{d}z_i}, \tag{A.12}$$

we obtain

$$\begin{aligned} p(z_1, \ \ldots\ , \ z_m|\mathsf{x}) &= \frac{p(z_1, \ldots , z_m, \mathsf{x})}{p(\mathsf{x})} \\ &= \frac{Z}{\acute{Z}} \frac{\prod_{i=1}^m p(z_i)p(\mathsf{x}|z_i)}{\prod_{j=1}^m p_j(\mathsf{x})} \\ &= \frac{\prod_{i=1}^m p(z_i)p(\mathsf{x}|z_i)}{\prod_{i=1}^m \int p(z_i)p(\mathsf{x}|z_i)\mathrm{d}z_i} \\ &= \prod_{i=1}^m p(z_i|\mathsf{x}). \end{aligned} \quad \text{(A.13)}$$

According to the definition of the conditional independence, $I(z_1, \ \ldots\ , \ z_m|\mathsf{x})$ holds. ■

A.4 Maximum Likelihood Expression for the PoE Model

Derivations of this section are based on discussions with Williams.

Let $\tilde{\Theta} = [\Theta_1 \ \Theta_2 \ \ldots \Theta_m]$ be a vector of parameters of a PoE model, with Θ_i denoting parameters of the i^{th} expert. Under the products of m experts the marginal density is expressed as

$$p(\mathsf{x}|\tilde{\Theta}) = \frac{\prod_{i=1}^m p_i(\mathsf{x}|\Theta_i)}{\int \prod_{i=1}^m p_i(\acute{\mathsf{x}}|\Theta_i)\mathrm{d}\acute{\mathsf{x}}}. \quad \text{(A.14)}$$

$p(\mathsf{x}|\Theta_i)$ can be viewed as the probability of point x under the i^{th} expert, m is the number of experts, and $\tilde{\mathsf{x}}$ represents all the possible data points generated by the PoE model.

Then the log-likelihood of a single point x, generated by the model with parameters $\tilde{\Theta}$, is determined as

$$\mathcal{L}_{\mathsf{x}}(\tilde{\Theta}) = \log p(\mathsf{x}|\tilde{\Theta}) = \sum_{i=1}^m \log p_i(\mathsf{x}|\Theta_i) - \log \int \prod_{i=1}^m p_i(\acute{\mathsf{x}}|\Theta_i) \ \mathrm{d}\acute{\mathsf{x}}. \quad \text{(A.15)}$$

Differentiation with respect to the parameter vector Θ_j leads to

$$\frac{\partial \log p(\mathsf{x}|\tilde{\Theta})}{\partial \Theta_j} = \frac{\partial \log p_j(\mathsf{x}|\Theta_j)}{\partial \Theta_j} - \frac{\partial \log \int \prod_{i=1}^{m} p_i(\acute{\mathsf{x}}|\Theta_i)\ \mathrm{d}\acute{\mathsf{x}}}{\partial \Theta_j}. \quad \text{(A.16)}$$

Expanding the right-most term, we obtain

$$\begin{aligned}
\frac{\partial \log \int \prod_{i=1}^{m} p_i(\acute{\mathsf{x}}|\Theta_i)\ \mathrm{d}\acute{\mathsf{x}}}{\partial \Theta_j} &= \frac{1}{\int \prod_{i=1}^{m} p_i(\acute{\mathsf{x}}|\Theta_i)\mathrm{d}\acute{\mathsf{x}}} \cdot \frac{\partial \int \prod_{i=1}^{m} p_i(\acute{\mathsf{x}}|\Theta_i)\mathrm{d}\acute{\mathsf{x}}}{\partial \Theta_j} \\
&= \frac{1}{\int \ldots} \int \prod_{i \neq j}^{m} p_i(\acute{\mathsf{x}}|\Theta_i) \frac{\partial p_j(\acute{\mathsf{x}}|\Theta_j)}{\partial \Theta_j} \mathrm{d}\acute{\mathsf{x}} \\
&= \frac{1}{\int \ldots} \int \frac{\partial p_j(\acute{\mathsf{x}}|\Theta_j)}{p_j(\acute{\mathsf{x}}|\Theta_j)\partial \Theta_j} \prod_{i=1}^{m} p_i(\acute{\mathsf{x}}|\Theta_i)\mathrm{d}\acute{\mathsf{x}} \\
&= \frac{1}{\int \ldots} \int \frac{\partial \log p_j(\acute{\mathsf{x}}|\Theta_j)}{\partial \Theta_{\mathsf{j}}} \prod_{i=1}^{m} p_i(\acute{\mathsf{x}}|\Theta_i)\mathrm{d}\acute{\mathsf{x}} \\
&= E_{p(\acute{\mathsf{x}}|\tilde{\Theta})} \left[\frac{\partial \log p_j(\acute{\mathsf{x}}|\Theta_j)}{\partial \Theta_j} \right], \qquad \text{(A.17)}
\end{aligned}$$

where we used $1/(\int \ldots)$ to denote the leftmost term of the r.h.s. of the first expansion of the expression.

Substituting this result into expression (A.16), we obtain the expression for the likelihood of the point x under the PoE model with parameters $\tilde{\Theta}$:

$$\frac{\partial \log p(\mathsf{x}|\tilde{\Theta})}{\partial \Theta_j} = \frac{\partial \log p_j(\mathsf{x}|\Theta_j)}{\partial \Theta_j} - E_{p(\acute{\mathsf{x}}|\tilde{\Theta})} \left[\frac{\partial \log p_j(\acute{\mathsf{x}}|\Theta_j)}{\partial \Theta_j} \right]. \quad \text{(A.18)}$$

A.5 Gibbs Sampling and MCMC

Gibbs sampling and Markov Chain Monte Carlo are described in numerous statistical sources. This section is influenced by Mackay (1998), Neal (1993), and Hinton (2000).

Many processes occurring in real life lead to highly complicated multi-dimensional distributions of the data and are impossible to model analytically. However, we can examine a sample from a distribution and learn something about its properties. We can also estimate expectations of various functions of random variables by averaging over the distribution of the sampled points:

$$\langle f(\mathsf{x})\rangle = \sum_{\mathsf{x}} f(\mathsf{x}) P(\mathsf{X} = \mathsf{x}) \approx \frac{1}{N} \sum_{i=0}^{N-1} f(\mathsf{x}^{(i)}), \tag{A.19}$$

where N is the sample size, $\mathsf{x}^{(i)}$ is the i^{th} sampled point, and x is distributed according to $P(\mathsf{x})$. However, sampling from the data distribution $P(\mathsf{x})$ may be complicated.

One way to model the data distribution is by using Markov Chains, which are defined by the probability of the initial point $P(\mathsf{x}^{(0)})$ and the *transition* probabilities $P(\mathsf{x}^{(t)}|\mathsf{x}^{(t-1)})$, which are chosen in such a way that $P(\mathsf{x}^{(t)}) \to P(\mathsf{x})$ as the time step t increases. It can be shown [see Neal (1993)] that certain Markov chains converge to a unique invariant distribution[2], referred as the equilibrium distribution. Then a sample from $P(\mathsf{x})$ may be drawn by sampling from the equilibrium distribution.

In some cases it is easier to draw a sample from a conditional distribution than from the joint $P(\mathsf{x})$ [for example, this is often the case for latent variable and graphical models]. In Gibbs sampling each random variable draws its values from a conditional probability with fixed values of other variables, so that for a 1-D 1-factor latent variable model

$$x_0 \leftarrow p(x_0) \tag{A.20}$$

$$z_i \leftarrow p(z|x_i),\ i > 0 \tag{A.21}$$

$$x_{i+1} \leftarrow p(x|z_i),\ i > 0. \tag{A.22}$$

Here z_i, x_i indicate fixed values of the hidden and visible variables on the i^{th} time step of the algorithm. Note that $p(z|x_i)$ can be seen as a density of the hidden variable z given that the visible variable x is fixed at the instance x_i.

[2]The distribution is *invariant* if, when reached, it persists. Usually invariance is a desirable property, showing that no matter how long we sample, the sample still comes from the same distribution(Neal, 1993).

Neal (1993) states that Gibbs sampling process may be seen as creation of a realization of a Markov chain, with the transition probabilities defined in this case as $p(z_i|x_i)$ and $p(x_{i+1}|z_i)$. The illustration of the case is given on figure 4.1 in section 4.3.3. Note that in this case $p(z_i|x_i)$ indicates the density of z_i when the i^{th} reconstruction of the visible variable x_i is fixed at its instance on the i^{th} iteration of the sampling procedure. We can see that in this case

$$\begin{aligned}
p(x_0, z_0, x_1) &= p(x_0)p(z_0|x_0)p(x_1|z_0) &\quad (A.23)\\
p(z_0, x_1|x_0) &= p(z_0|x_0)p(x_1|z_0) &\quad (A.24)\\
p(x_1|x_0) &= \int p(z_0|x_0)p(x_1|z_0)\mathrm{d}z_0 &\quad (A.25)\\
p(x_1) &= \int p(x_0) \int p(z_0|x_0)p(x_1|z_0)\mathrm{d}z_0 \; \mathrm{d}x_0. &\quad (A.26)
\end{aligned}$$

Appendix B

Analytic Solution of the Constrained Gaussian Model

B.1 Covariance of the PoCG model

B.1.1 Covariance in the Visible Space

Covariance in the single expert case

Consider eigenvalue decomposition of the inverse covariance C^{-1} of a single constrained Gaussian model:

$$\mathsf{C}^{-1} = \mathsf{U}\Lambda^{-1}\mathsf{U}^T, \tag{B.1}$$

where $\mathsf{U} = [\mathsf{u}_1, \dots, \mathsf{u}_{d-1}, \hat{\mathsf{w}}] \subset \mathcal{R}^{d\times d}$ is the matrix of normalized eigenvectors of the covariance C, $\Lambda = diag(\lambda_1, \lambda_2, \dots \lambda_m) \subset \mathcal{R}^{d\times d}$ is the diagonal matrix of corresponding eigenvalues.

The inverse covariance C^{-1} can be written explicitly as

$$\mathsf{C}^{-1} = \sum_{i=1}^{d-1} \mathsf{u}_i\mathsf{u}_i^T\beta_0 + \hat{\mathsf{w}}\hat{\mathsf{w}}^T\beta_{\hat{w}}, \tag{B.2}$$

where $\beta_0 = \sigma_0^{-2}, \beta_{\hat{w}} = \sigma_{\hat{w}}^{-2}$, $\mathsf{u}_1, \dots, \mathsf{u}_{d-1}$ are directions of elongation, $\hat{\mathsf{w}}$ is the direction of shrinkage.

Then (B.2) can be re-written in a more convenient form not involving the summations:

$$\begin{aligned} \mathsf{C}^{-1} &= \beta_0 \mathsf{I}_d + (\beta_{\hat{w}} - \beta_0)\hat{\mathsf{w}}\hat{\mathsf{w}}^T \\ &= \beta_0 \mathsf{I}_d + \hat{\beta}\hat{\mathsf{w}}\hat{\mathsf{w}}^T, \end{aligned} \tag{B.3}$$

where

$$\hat{\beta} = \beta_{\hat{w}} - \beta_0. \tag{B.4}$$

The expression follows from the fact that for any matrix $\mathsf{V} = [\mathsf{v}_1, \ldots, \mathsf{v}_d] \subset \mathcal{R}^{d \times d}$, whose columns constitute an orthonormal basis,

$$\sum_{i=1}^{d} \mathsf{v}_i \mathsf{v}_i^T \cdot 1 = \mathsf{V}\mathsf{V}^T = \mathsf{I}_d. \tag{B.5}$$

Thus the summation of all the outer product of columns of U, weighted by the constant variance β_0, yields the multiple of the identity matrix $\beta_0 \mathsf{I}_d$.

If we now make a substitution

$$\mathsf{w} = \hat{\mathsf{w}}\sqrt{\hat{\beta}}, \tag{B.6}$$

we see that (1.23) can be re-written as

$$\mathsf{C}^{-1} = \beta_0 \mathsf{I}_d + \mathsf{w}\mathsf{w}^T. \tag{B.7}$$

Notice that according to the constraint considerations (1.22) $\beta_0 < \beta_{\hat{w}} \Rightarrow \mathsf{w} \subset \mathcal{R}^d$, i.e. weights w are real-valued.

Covariance in the Multiple Expert Case

In order to find covariance C_Σ from (2.4), we can apply the Woodbury transform, also known as the Matrix Inversion Lemma (see, for example, Press et al. (1992)):

$$(\mathsf{A} + \mathsf{BDC})^{-1} = \mathsf{A}^{-1} - \mathsf{A}^{-1}\mathsf{B}(\mathsf{D}^{-1} + \mathsf{CA}^{-1}\mathsf{B})^{-1}\mathsf{CA}^{-1}. \tag{B.8}$$

From (B.3), the resulting covariance matrix for m experts can be expressed as

$$\begin{aligned} \mathsf{C}_\Sigma^{-1} = \sum_i \mathsf{C}_i^{-1} &= \beta_\Sigma \mathsf{I}_d + \sum_{i=1}^{m} \hat{\beta}_i \hat{\mathsf{w}}_i \hat{\mathsf{w}}_i^T \\ &= \beta_\Sigma \mathsf{I}_d + \mathsf{W}\mathsf{W}^T \subset \mathcal{R}^{d\times d}, \end{aligned} \tag{B.9}$$

where

$$\beta_\Sigma = \sum_{i=1}^{m} \beta_{0_i} \tag{B.10}$$

$$\hat{\beta}_i = \beta_{\hat{w}_i} - \beta_{0i} \tag{B.11}$$

$$\mathsf{W} = [\sqrt{\hat{\beta}_1}\hat{\mathsf{w}}_1 \ \sqrt{\hat{\beta}_2}\hat{\mathsf{w}}_2 \ \ldots \ \sqrt{\hat{\beta}_m}\hat{\mathsf{w}}_m] \subset \mathcal{R}^{d\times m}. \tag{B.12}$$

If we make an assumption that **all m experts have equal variances β_0 in the directions of elongation** then

$$\beta_\Sigma = m\beta_0 \tag{B.13}$$

$$\hat{\beta}_i = \beta_{\hat{w}_i} - \beta_0. \tag{B.14}$$

It is easy to see that C_Σ^{-1} is symmetric. We can further notice that since $\beta_\Sigma > 0$, C_Σ^{-1} is also positive-definite.

□ Indeed, consider $\mathsf{x} \subset \mathcal{R}^d$, $\mathsf{x} \neq \mathsf{0}$.

$$\begin{aligned} \mathsf{x}^T \mathsf{C}_\Sigma^{-1} \mathsf{x} &= \beta_\Sigma \mathsf{x}^T\mathsf{x} + \mathsf{x}^T \hat{\mathsf{W}}\hat{\mathsf{W}}^T \mathsf{x} \\ &= \beta_\Sigma ||\mathsf{x}||^2 + ||\hat{\mathsf{W}}^T\mathsf{x}||^2 > 0. \end{aligned} \tag{B.15}$$

■

Therefore, C_Σ is a valid covariance of the Gaussian distribution.

B.1.2 Hidden Variable Representation

Let us consider a single constrained Gaussian distribution. By analogy with the results obtained in appendix B.1.1, we can find the expression of the covariance matrix

$$\begin{aligned}\mathsf{C} &= \sigma_0^2 \mathsf{I}_d - (\sigma_0^2 - \sigma_{\hat{w}}^2)\hat{\mathsf{w}}\hat{\mathsf{w}}^T \\ &= \sigma_0^2 \mathsf{I}_d - \hat{\sigma}^2 \hat{\mathsf{w}}\hat{\mathsf{w}}^T,\end{aligned} \tag{B.16}$$

where $\hat{\sigma}^2 = \sigma_0^2 - \sigma_{\hat{w}}^2$. Note that for the covariance C to be positive definite, we have a constraint on $\hat{\sigma}$:

$$0 < \hat{\sigma}^2 < \sigma_0^2. \tag{B.17}$$

It is always satisfied for our definition of the variances.

To find a linear latent variable model resulting in this covariance in the visible space, we need to assume a particular mapping from the hidden to the visible space and determine variances of the hidden variables and the noise term.

Let us consider a linear mapping

$$\mathsf{x} = z_1 \hat{\mathsf{w}} + \mathsf{z}_2, \tag{B.18}$$

where $z_1 \sim N(0, \sigma_{z1}^2)$, $\mathsf{z}_2 \sim N(0, \mathsf{C}_2)$. Expression (B.18) leads to $\mathsf{x} \sim N(0, \sigma_{z1}^2 \hat{\mathsf{w}}\hat{\mathsf{w}}^T + \mathsf{C}_2)$.

Then, the covariances of z_2 and the variance of z_1 can be found from

$$\sigma_{z1}^2 \hat{\mathsf{w}}\hat{\mathsf{w}}^T + \mathsf{C}_2 \equiv \sigma_{\hat{w}}^2 \hat{\mathsf{w}}\hat{\mathsf{w}}^T - \sigma_0^2 \hat{\mathsf{w}}\hat{\mathsf{w}}^T + \sigma_0^2 \mathsf{I}_d. \tag{B.19}$$

We can obtain the possible expressions for the covariance matrix in the visible space if we assume

$$\begin{aligned}\sigma_{z_1}^2 &= \sigma_{\hat{w}}^2 \\ \mathsf{C}_2 &= \sigma_0^2 \mathsf{I}_d - \sigma_0^2 \hat{\mathsf{w}}\hat{\mathsf{w}}^T.\end{aligned} \tag{B.20}$$

Note that this is not the only solution. Other values of $\sigma_{z_1}^2$ and C_2, satisfying (B.19), are also valid[1].

[1]Notice, however, that the uncorrelated noise term $\mathsf{C}_2 = \sigma_0^2 \mathsf{I}_d$ leads to the non-Gaussian distribution of the latent variables with $\sigma_{z_1}^2 < 0$.

We see that unlike the PPCA and the FA latent variable representations, the PoCG may lead to the correlated noise (z_2 term), and does not result in a decrease in the number of free parameters.

B.2 Derivation of the ML Equations for the Parameter Estimates

In sections B.2.1 and B.2.2 we derive ML equations for parameters of a PoCG model. Section B.2.1 makes a simplifying assumption of independence of the expert weights W from the log-noise term x. In section B.2.2 we show that even without the simplifying assumption, the equations result in the same solutions.

B.2.1 ML Equations Under the Assumption of Independence of W and β_0

Here we make a simplifying assumption that the expert weight matrix W is independent of the inverse log-noise component x (and, consequently, of β_0). This is not entirely true, since from (B.10) and (B.14) the weight W depends on β_0, and we need to perform an equality constraint optimization to estimate optimal parameters. However, later we will show that this simplification leads to the same ML estimates as the constrained case.

By analogy with (1.8), we can specify ML conditions for each of the model parameters:

$$\frac{\partial \mathcal{L}}{\partial \mathsf{W}} = \frac{N}{2}\left[\frac{\partial \ln |\mathsf{C}_\Sigma^{-1}|}{\partial \mathsf{W}} - \frac{\partial \mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S})}{\partial \mathsf{W}}\right] = 0 \tag{B.21}$$

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{N}{2}\left[\frac{\partial \ln |\mathsf{C}_\Sigma^{-1}|}{\partial x} - \frac{\partial \mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S})}{\partial x}\right] = 0, \tag{B.22}$$

As derived, for example, in Krzanowski and Marriott (1994), or can be obtained by applying matrix differentiation rules (Minka, 1999), this leads to

$$\frac{\partial \mathcal{L}}{\partial \mathsf{W}} = \mathrm{tr}\left[\mathsf{C}_\Sigma \frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial \mathsf{W}}\right] - \mathrm{tr}\left[\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial \mathsf{W}} \mathsf{S}\right] = 0. \tag{B.23}$$

[The expression for x is totally analogous].

To obtain the ML equation for x and W, we need to estimate derivatives of determinants and traces in (B.21), (B.22). In order to do it, we can compute differentials of the log-likelihood w.r.t. parameters we need to optimize, and transform the results into canonical forms for matrix derivatives (see Minka (1999) for matrix differentiation rules).

We will write d_X to denote a differential with respect to the matrix X.

This results in the following expressions:

$$\begin{aligned}
\mathrm{d}_\mathsf{W} \ln|\mathsf{C}_\Sigma^{-1}| &= \mathrm{tr}(\mathsf{C}_\Sigma \mathrm{d}_\mathsf{W} \mathsf{C}_\Sigma^{-1}) \\
&= \mathrm{tr}(\mathsf{C}_\Sigma (\mathrm{d}_\mathsf{W} \mathsf{W}) \mathsf{W}^T + \mathsf{C}_\Sigma \mathsf{W} (\mathrm{d}_\mathsf{W} \mathsf{W})^T) \\
&= 2\mathrm{tr}(\mathsf{W}^T \mathsf{C}_\Sigma (\mathrm{d}_\mathsf{W} \mathsf{W})); \qquad (B.24)
\end{aligned}$$

$$\begin{aligned}
\mathrm{d}_\mathsf{W} \mathrm{tr}(\mathsf{C}_\Sigma^{-1} S) &= \mathrm{tr}((\mathrm{d}_\mathsf{W} \mathsf{C}_\Sigma^{-1}) \mathsf{S}) \\
&= \mathrm{tr}((\mathrm{d}_\mathsf{W} \mathsf{W}) \mathsf{W}^T \mathsf{S} + \mathsf{W} (\mathrm{d}_\mathsf{W} \mathsf{W})^T \mathsf{S}) \\
&= 2\mathrm{tr}(\mathsf{W}^T \mathsf{S} (\mathrm{d}_\mathsf{W} \mathsf{W})); \qquad (B.25)
\end{aligned}$$

$$\begin{aligned}
\mathrm{d}_x \ln|\mathsf{C}_\Sigma^{-1}| &= \mathrm{tr}(\beta_0 m \mathsf{C}_\Sigma \ \mathrm{d}_x x) \qquad (B.26) \\
\mathrm{d}_x \mathrm{tr}(\mathsf{C}_\Sigma^{-1} S) &= \mathrm{tr}(\beta_0 m \mathsf{S} \ \mathrm{d}_x x). \qquad (B.27)
\end{aligned}$$

Substituting into (B.21) and (B.22) and differentiating with respect to W and x we obtain expressions for the derivatives of the log-likelihood $\mathcal{L}$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathsf{W}} &= N(\mathsf{C}_\Sigma \mathsf{W} - \mathsf{S}\mathsf{W}) \qquad (B.28) \\
\frac{\partial \mathcal{L}}{\partial x} &= Nm\mathrm{tr}(\beta_0(\mathsf{C}_\Sigma - \mathsf{S})). \qquad (B.29)
\end{aligned}$$

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

$$\mathsf{C}_\Sigma \mathsf{W} - \mathsf{S}\mathsf{W} = 0 \tag{B.30}$$

$$\mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}) = 0. \tag{B.31}$$

Note that in order to find the maximum-likelihood solution for the weight matrix W and the distortion term β_0, we need to solve (B.30) and (B.31) simultaneously. In order to simplify derivations of explicit expressions for the model parameters, we can once again make use of the assumption of independence of W from β_0. Then we can first solve (B.30) for W, and further express β_0^{ML} by solving $\partial\mathcal{L}(\mathsf{W}_{ML})/\partial x = 0$.

B.2.2 Justification of the Simplifying Assumptions *

Here we will briefly show that even without making the simplifying assumption of Appendix B.2.1, the resulting ML equations will lead to the same solutions as (B.30) and (B.31). This section may be skipped without any loss of understanding of the key results derived in the thesis.

Consider optimization of a function of several variables. Generally, in optimization tasks we cannot introduce new variables, imposing some constraints on several of the old ones, substitute these variables into the expression of the function, and derive extremums assuming condition of independence of old and new parameters. In general, we will need to use methods of constrained optimization [most notably, based on the Lagrange's principle for equality constraints]. Note that in the previous section we have carried out unconstrained optimization on W and the log x of the inverse noise term β_0, though elements of W were in fact functions of β_0 [see (B.6), (B.14)].

Here we will perform unconstrained optimization on the log-likelihood, where the covariance C_Σ is expressed as a function of independent variables $\hat{\mathsf{W}}$ and x.

From (B.9), the inverse covariance of the PoCG model C_Σ^{-1} can be re-written as

$$\mathsf{C}_\Sigma^{-1} = \beta_0 \mathsf{I}_d + \hat{\mathsf{W}}\mathsf{B}\hat{\mathsf{W}}^T, \tag{B.32}$$

Here columns of $\hat{\mathsf{W}}$ constitute directions of shrinkage for each expert, and

$$\mathsf{B} = diag(\beta_{w_1}, \dots, \beta_{w_m}) - \beta_0 \mathsf{I}_d. \tag{B.33}$$

Then (B.32) implies

$$\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial x} = \beta_0(\mathsf{I}_d - \hat{\mathsf{W}}\hat{\mathsf{W}}^T) \tag{B.34}$$

Once again applying matrix differentiation rules as shown in the previous section, we see that

$$\frac{\partial \mathcal{L}}{\partial \hat{\mathsf{W}}} = \frac{N}{2}(\mathsf{C}_\Sigma - \mathsf{S})\hat{\mathsf{W}}(\mathsf{B} + \mathsf{B}^T) = 0 \tag{B.35}$$

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{N}{2}\mathrm{tr}((\mathsf{C}_\Sigma - \mathsf{S})(m\mathsf{I}_d - \hat{\mathsf{W}}\hat{\mathsf{W}}^T)\beta_0) = 0. \tag{B.36}$$

Since B is a non-singular matrix, (B.35) implies

$$\mathsf{C}\hat{\mathsf{W}} = \mathsf{S}\hat{\mathsf{W}}, \tag{B.37}$$

what corresponds to the ML equation of the previous section [see (B.30)].

Since $\hat{\mathsf{W}}$ is independent of β_0, we can indeed first find the ML estimate for the experts' weights, and further express x_{ML} by solving $\partial\mathcal{L}(\mathsf{W}_{ML})/\partial x = 0$.

This will lead to the ML estimates that are identical to the ones in the previous section. Throughout the thesis, we will use the notation of the previous section for better clarity.

B.3 ML Solutions for the Parameter Estimates

There are two possible cases of non-zero maximum-likelihood solutions for the the PoCG weight matrix W. In the first case, the model covariance equals to the sample covariance, i.e. $\mathsf{S} = \mathsf{C}_\Sigma$, and C_Σ is *exact*. In the second case, $\mathsf{C}_\Sigma \mathsf{W} = \mathsf{S}\mathsf{W}, \mathsf{W} \neq 0, \mathsf{S} \neq \mathsf{C}_\Sigma$, and the model covariance C_Σ will be to as *approximate*.

By analogy with Tipping and Bishop (1997), we will consider singular value decomposition of the weight matrix, and try to establish dependencies that may exist between left singular values of W and eigenvectors of the sample covariance S.

B.3.1 Exact Model Covariance

The compact form of the singular value decomposition of the weight matrix W can be expressed as follows:

$$\mathsf{W} = \mathsf{U}\mathsf{L}\mathsf{R}^T, \tag{B.38}$$

where $\mathsf{U} \subset \mathcal{R}^{d\times m}$ is a matrix of orthonormal column vectors, $\mathsf{L} \subset \mathcal{R}^{m\times m}$ is a diagonal, and $\mathsf{R} \subset \mathcal{R}^{m\times m}$ is a rotation matrix. (This representation follows from singular value decomposition with $\hat{\mathsf{U}} = [\mathsf{U}\ 0_{d\times(d-m)}]$ as an orthogonal and $\hat{\mathsf{L}} = [\mathsf{L}\ 0_{m\times(d-m)}]^T$ as a pseudo-diagonal matrix. $\hat{\mathsf{U}}\hat{\mathsf{L}}$ is a product of vectors of matrices, leading to (B.38)).

Considering non-singularity of S and C_Σ,

$$\mathsf{S} = \mathsf{C}_\Sigma \Rightarrow \mathsf{S}^{-1} = {\mathsf{C}_\Sigma}^{-1}. \tag{B.39}$$

Then, substituting (B.38) into the expression for the inverse covariance (B.9), we can rewrite the r.h.s. of (B.39) as

$$\mathsf{U}(\beta_\Sigma \mathsf{I}_m + \mathsf{L}^2)\mathsf{U}^T \quad = \quad \mathsf{U}_S \Lambda_S^{-1} \mathsf{U}_S^T, \tag{B.40}$$

where U_S, Λ_S correspond to eigenvector and eigenvalue matrices of S.

Moreover, since $\beta_\Sigma \mathsf{I}_m + \mathsf{L}^2$ is a diagonal, both parts of the equation represent eigenvalue decomposition of the same matrix. Then, from the definition of eigenvectors and eigenvalues, $\mathsf{U} = \mathsf{U}_S$. Therefore, the weight matrix for an

m-expert PoCG model corresponds to m scaled and rotated eigenvectors of S. Using the notations defined above, the weight matrix W is expressed as

$$\mathsf{W} = \mathsf{U}_S(\Lambda_S^{-1} - \beta_\Sigma \mathsf{I}_m)^{1/2}\mathsf{R}^T. \tag{B.41}$$

B.3.2 Approximate Model Covariance

Applying SVD to W and substituting the result and the expression of the PoCG covariance (2.6) into (B.30), we get

$$\begin{aligned}
\mathsf{C}_\Sigma\mathsf{UL} &= (\beta_\Sigma^{-1}\mathsf{I}_d - \beta_\Sigma^{-1}\mathsf{W}(\beta_\Sigma + \mathsf{W}^T\mathsf{W})^{-1}\mathsf{W}^T)\mathsf{UL} \\
&= (\beta_\Sigma^{-1}\mathsf{I}_d - \beta_\Sigma^{-1}\mathsf{ULR}^T(\beta_\Sigma\mathsf{I}_m + \mathsf{RL}^2\mathsf{R}^T)^{-1}\mathsf{RLU}^T)\mathsf{UL} \\
&= \mathsf{U}(\beta_\Sigma^{-1}\mathsf{I}_m - \beta_\Sigma^{-1}\mathsf{LR}^T(\beta_\Sigma\mathsf{I}_m + \mathsf{RL}^2\mathsf{R}^T)^{-1}\mathsf{RL})\mathsf{L} \\
&= \mathsf{U}(\beta_\Sigma^{-1}\mathsf{I}_m - \beta_\Sigma^{-1}(\beta_\Sigma\mathsf{L}^{-2} + \mathsf{I}_m)^{-1})\mathsf{L}.
\end{aligned} \tag{B.42}$$

where $\mathsf{U} = [\mathsf{u}_1\mathsf{u}_2 \dots \mathsf{u}_m] \subset \mathcal{R}^{d\times m}$ is a matrix of left singular vectors of W with the columns constituting an orthonormal basis, $\mathsf{L} = diag(l_1 l_2 \dots l_m) \subset \mathcal{R}^{m\times m}$ is a matrix of singular values of W, and $\mathsf{R} \subset \mathcal{R}^{m\times m}$ defines an arbitrary rigid rotation of W.

The solution of (B.30) leads to

$$\mathsf{C}_\Sigma\mathsf{W} = \mathsf{SW} \Rightarrow \mathsf{C}_\Sigma\mathsf{UL} = \mathsf{SUL}, \tag{B.43}$$

from where we can see that

$$\mathsf{SUL} = \mathsf{U}(\beta_\Sigma^{-1}\mathsf{I}_d - \beta_\Sigma^{-1}(\beta_\Sigma\mathsf{L}^{-2} + \mathsf{I}_m)^{-1})\mathsf{L}. \tag{B.44}$$

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 (B.44) is a diagonal matrix (i.e, just a scaling factor of U). Then from the definition of eigenvectors and eigenvalues, (B.44) defines the matrix form of the eigenvector equation[2], with both sides post-multiplied by the diagonal matrix L.

[2]Consider $\mathsf{A} \subset \mathcal{R}^{d\times d}$, $\mathsf{U}_A \subset \mathcal{R}^{d\times m}$, and $\Lambda_A \subset \mathcal{R}^{m\times m}$. If $\mathsf{AU_A} = \mathsf{U_A}\Lambda_\mathsf{A}$ and Λ_A is a diagonal matrix, then U_A and Λ_A give rise to $\mathsf{Au}_i = \lambda_i\mathsf{u}_i$ for $i = 1\dots m$ and therefore correspond to the matrices of m eigenvectors and eigenvalues of A respectively.

Obviously, if $l_i = 0$ then u_i is arbitrary. As it follows from the expansion of the SVD, this situation corresponds to the case when the dimension defined by the vector u_i is *discarded* in the projection of an m-dimensional unit hyper-sphere from the hidden space onto a d-dimensional hyper-ellipse in the visible space, with the projection defined by W. Otherwise, the dimension defined by u_i is *retained*, and

$$\begin{aligned} \mathsf{C}_\Sigma \mathsf{u}_i &= \mathsf{S}\mathsf{u}_i = \mathsf{u}_i \lambda_i, && \text{(B.45)} \\ \lambda_i &= \beta_\Sigma^{-1}(1 - (\beta_\Sigma l_i^{-2} + 1)^{-1}), && \text{(B.46)} \end{aligned}$$

where u_i is an eigenvector of S, and λ_i is its corresponding eigenvalue.

From the expression for eigenvalues (B.46), the scaling factor of the i^{th} expert can be expressed as

$$l_i = (\lambda_i^{-1} - \beta_\Sigma)^{1/2} \Rightarrow \lambda_i \leq \frac{1}{m\beta_0}. \tag{B.47}$$

This result resembles the solution for the scaling factors in the PPCA case [*cf* Tipping and Bishop (1997)]. However, in contrast to the PPCA solution where $l_i = (\lambda_i - \sigma^2)^{1/2}$, we see that λ_i and σ^2 are replaced by their inverses. In addition, in the PoCG case there is a factor of the number of experts m in the expansion of β_Σ.

Thus possible values of each scaling factor are

$$l_i = \begin{cases} (\lambda_i^{-1} - \beta_\Sigma)^{1/2} & \text{if } \mathsf{u}_i \text{ is retained;} \\ 0 & \text{if } \mathsf{u}_i \text{ is discarded.} \end{cases} \tag{B.48}$$

Equivalently, W_{ML} can be re-written as

$$\mathsf{W}_{ML} = \mathsf{U}_S(\tilde{\Lambda}_S - \beta_\Sigma \mathsf{I}_m)^{1/2}\mathsf{R}^T, \tag{B.49}$$

where $\mathsf{R} \subset \mathcal{R}^{m\times m}$ is an arbitrary rotation matrix, $\mathsf{U}_S = \mathsf{U} = [\mathsf{u}_1\mathsf{u}_2\ldots\mathsf{u}_m] \subset \mathcal{R}^{d\times m}$ is a matrix whose columns to m eigenvectors of S, and

$$\tilde{\Lambda}_S = diag(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_m) \subset \mathcal{R}^{m\times m},$$

such that

$$\tilde{\lambda}_i = \begin{cases} \lambda_i^{-1} & \text{if } \mathsf{u}_i \text{ is retained;} \\ \beta_\Sigma & \text{if } \mathsf{u}_i \text{ is discarded.} \end{cases} \tag{B.50}$$

B.4 Investigation of Parameter Properties

Some of the derivations of this section are based on Tipping and Bishop's solution for the PPCA model [see pp. 9-10 Tipping and Bishop (1997)]. However, we explain the intuition behind the influence of permutations of discarded and retained eigenvalues on changes in the log-likelihood, and the effects of retaining a smaller number of eigenvalues.

B.4.1 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 (Tomasi, 1999), 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 the eigenvectors retained in W_{ML}, we will express the log-likelihood (1.8) of a Gaussian through eigenvectors and eigenvalues of its covariance matrix.

From expression of the PoCG weights (B.49) and the form of the model's inverse covariance (B.9), the components of (1.8) can be expressed as follows:

$$\mathsf{C}_\Sigma^{-1} = \sum_{i=1}^{r} \mathsf{u}_i \mathsf{u}_i^T \lambda_i^{-1} + \sum_{i=r+1}^{d} \mathsf{u}_i \mathsf{u}_i^T \beta_\Sigma. \tag{B.51}$$

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

$$\ln |\mathsf{C}_\Sigma|^{-1} = \ln\left(\beta_\Sigma^{(d-r)} \prod_{i=1}^{r} \lambda_i^{-1}\right) = -\sum_{i=1}^{r} \ln(\lambda_i) + (d-r)\ln\beta_\Sigma \quad \text{(B.52)}$$

$$\mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S}) = r + \beta_\Sigma \sum_{i=r+1}^{d} \lambda_i. \quad \text{(B.53)}$$

By substituting these expressions into the form of $\mathcal{L}$ for a Gaussian (1.8), we get

$$\mathcal{L}(\mathsf{W}_{ML}) = -\frac{N}{2}\left[d\ln(2\pi) + \sum_{i=1}^{r}\ln(\lambda_i) - (d-r)\ln\beta_\Sigma\right.$$
$$\left. + \quad r + \beta_\Sigma \sum_{i=r+1}^{d}\lambda_i\right]. \quad \text{(B.54)}$$

As we have assumed in Appendix B.2.1 and later verified in Appendix B.2.2, we may consider W to be independent of the noise parameter β_0. Then (B.54) leads to

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \frac{\partial \mathcal{L}(\mathsf{W}_{\mathcal{ML}})}{\partial \beta_0} = 0 \quad \Rightarrow \quad -\frac{N}{2}\left[-\frac{d-r}{\beta_0 m}m + m\sum_{i=r+1}^{d}\lambda_i\right] = 0, \quad \text{(B.55)}$$

what results in

$$\beta_0^{ML} = \frac{d-r}{m\sum_{i=r+1}^{d}\lambda_i}. \quad \text{(B.56)}$$

We can notice that 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

$\mathcal{L}(\mathsf{W}_{ML}, \beta_0^{ML}) =$

$$= -\frac{N}{2}\left[\sum_{i=1}^{r}\ln(\lambda_i) + (d-r)\ln\frac{\sum_{i=r+1}^{d}\lambda_i}{(d-r)} + r + (d-r)\right] + c \quad \text{(B.57)}$$

$$= -\frac{N}{2}\left[(d-r)\ln\frac{\sum_{j=r+1}^{d}\lambda_j}{(d-r)} - \sum_{j=r+1}^{d}\ln(\lambda_j)\right] + \acute{c} \quad \text{(B.58)}$$

where $c, \acute{c}$ are constants.

Here we have used the fact that

$$\sum_{i=1}^{d} \ln(\lambda_i) = ln|\mathsf{S}| \tag{B.59}$$

is a constant.

Let A be the term in the square brackets of (B.58). It is easy to see that $\mathcal{L}$ is maximized with respect to the λ terms when A is minimized. We will further investigate how flipping a retained and a discarded eigenvector can influence a change in the log-likelihood.

Tipping and Bishop claim that if λ_i is a retained eigenvalue (i.e. it corresponds to a retained eigenvector with the scaling factor $l_i \neq 0$), λ_k corresponds to a discarded eigenvector, $\bar{\lambda}$ is the mean of the discarded eigenvalues, and

$$\lambda_k \leq \lambda_i \leq \bar{\lambda}, \tag{B.60}$$

then flipping λ_i and λ_k (discarding λ_i and retaining λ_k) will result in an increase in $\mathcal{L}$ (decrease in A).

In other words, if a retained eigenvalue λ_i lies between discarded eigenvalues, the likelihood of the solution can be increased by flipping λ_i with the lower of the discarded eigenvalues. Let us show it formally.

At first, let λ_k be a discarded, and λ_i - a retained eigenvalue. A, the term in the brackets of (B.58), is

$$\begin{aligned} A &= (d-r)\ln\frac{\sum_{j=r+1}^{d}\lambda_j}{(d-r)} - \sum_{j=r+1}^{d}\ln(\lambda_j) && \text{(B.61)} \\ &= \ln\frac{\bar{\lambda}^{(d-r)}}{\lambda_k\prod_{j=r+1,j\neq k}^{d}\lambda_j}. && \text{(B.62)} \end{aligned}$$

After flipping λ_i and λ_k, the new mean of discarded eigenvalues is

$$\hat{\bar{\lambda}} = \bar{\lambda}\left(1 + \frac{\lambda_i - \lambda_k}{(d-r)\bar{\lambda}}\right). \tag{B.63}$$

If $\hat{A}$ is a new value of A after flipping λ_i and λ_k, then the difference between $\mathcal{L}(\hat{\mathcal{A}})$ and $\mathcal{L}(\mathcal{A})$ is proportional to

$$\hat{A} - A = \ln\left[\left(1 + \frac{\lambda_i - \lambda_k}{(d-r)\bar{\lambda}}\right)^{(d-r)} \frac{\lambda_k}{\lambda_i}\right]. \tag{B.64}$$

Let us denote

$$\begin{aligned} \lambda_i &= \lambda_k x, \;\; x \geq 1 \\ \bar{\lambda} &= \lambda_k z, \;\; z \geq 1 \\ n &= d - r > 0. \end{aligned} \tag{B.65}$$

To prove correctness of the claim, we need to show that

$$\forall n, x, z.\; n > 1.1 \leq x \leq z \Rightarrow \ln\left[\left(1 + \frac{x-1}{zn}\right)^n \frac{1}{x}\right] \leq 0. \tag{B.66}$$

(B.60), (B.66) $\Rightarrow$

$$\ln\left[\left(1 + \frac{x-1}{zn}\right)^n \frac{1}{x}\right] \leq \max_x \left(\ln\left[\left(1 + \frac{x-1}{xn}\right)^n \frac{1}{x}\right]\right) \tag{B.67}$$

For $n > 1$, the right-most term has the only maximum at $x = 1$ (i.e. $\lambda_i = \lambda_k$) and equals to 0 at that point. Thus, (B.65) $\Rightarrow \hat{A} - A \leq 0$, and the gap monotonically increases when $x > 1$.

Like in Tipping and Bishop's solution for PPCA, this implies that the retained eigenvalues should be continuous within a sorted multi-set of all the eigenvalues of the sample covariance S.

This result, together with the constraint condition on the retained eigenvalues (B.47) and the noise parameter (B.55), yields to

$$\forall i = 1 \ldots r.\; \lambda_i \leq \frac{\sum_{j=r+1}^{d} \lambda_j}{d - r}, \tag{B.68}$$

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

B.4.2 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 $\mathsf{W} \subset \mathcal{R}^{d \times m}$ will result in a higher value of $\mathcal{L}$. To prove this, we need to show that the term in the brackets of (B.57) decreases with a decrease in the number of the discarded eigenvectors $(d-r)$.

Let the term in the brackets of (B.57) be denoted by B_n for $n = d - r$ discarded eigenvectors. Then

$$\begin{aligned} B_n &= \sum_{i=1}^{r} \ln(\lambda_i) + (d-r) \ln \frac{\sum_{i=r+1}^{d} \lambda_i}{(d-r)} + r + (d-r) \\ &= n \left(\ln \bar{\hat{\lambda}}(n) - \ln(\prod_{i=1}^{n} \hat{\lambda}_i^{1/n}) \right) + d + \ln |\mathsf{S}|, \end{aligned} \quad \text{(B.69)}$$

where $\hat{\lambda}_i = \lambda_{n+i}$ (i.e. $\hat{\lambda}_i$ is the ith discarded eigenvalue), and $\bar{\hat{\lambda}}(n)$ is the mean of n discarded eigenvalues.

Note that since the sample covariance S of a Gaussian distribution is positive definite, $\ln |\mathsf{S}|$ is some constant.

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

$$B_n - B_{n+1} \leq 0. \quad \text{(B.70)}$$

We will simplify this inequality, and check whether it holds for all $n \in \mathcal{N}, n < d$ (i.e. with at least one vector retained).

Substituting B_n and B_{n+1} into (B.70), we obtain

$$B_n - B_{n+1} = \ln \frac{\bar{\hat{\lambda}}(n)^n \hat{\lambda}_{n+1}}{\bar{\hat{\lambda}}(n+1)^{n+1}} \quad \text{(B.71)}$$

$\Rightarrow$

$$\frac{\bar{\hat{\lambda}}(n)^n \hat{\lambda}_{n+1}}{\bar{\hat{\lambda}}(n+1)^{n+1}} \leq 0 \;\; \Rightarrow \;\; B_n - B_{n+1} \leq 0. \quad \text{(B.72)}$$

Thus, in order to prove (B.70) it is only sufficient to show that the inequality on the l.h.s. of (B.72) holds.

As shown in the previous section, discarding n eigenvalues in the PoCG model corresponds to discarding n principal components of S. Since $\hat{\lambda}_{n+1}$ corresponds to the $n+1$st 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

$$0 < \hat{\lambda}_{n+1} \leq \bar{\hat{\lambda}}(n+1) \leq \bar{\hat{\lambda}}(n). \tag{B.73}$$

Let $\bar{\hat{\lambda}}(n) = z\bar{\hat{\lambda}}(n+1)$, $\bar{\hat{\lambda}}_{n+1} = x\hat{\lambda}_{n+1}$. Then (B.73) $\Rightarrow$

$$1 \leq x \leq z. \tag{B.74}$$

From (B.73) and (B.74), the l.h.s. of (B.72) can be rewritten as

$$\ln \frac{\bar{\hat{\lambda}}(n)^n \hat{\lambda}_{n+1}}{\bar{\hat{\lambda}}(n+1)^{n+1}} \quad \Leftrightarrow \quad \ln \frac{z^n}{x^{n+1}} \leq 0. \tag{B.75}$$

Note that

$$\bar{\hat{\lambda}}(n+1) \quad = \quad \frac{\bar{\hat{\lambda}}(n)n + \hat{\lambda}_{n+1}}{n+1} \Rightarrow x = \frac{nz+1}{n+1}. \tag{B.76}$$

From (B.72) and by substitution of (B.76) into (B.75), we see that

$$\Delta(z) \geq 0 \Leftrightarrow \ln \frac{z^n}{x^{n+1}} \leq 0 \quad \Rightarrow \quad B_n - B_{n+1} \leq 0, \tag{B.77}$$

where

$$\Delta(z) = (n+1)\ln(nz+1) - n\ln z - (n+1)\ln(n+1) \geq 0 \tag{B.78}$$

is a function of z defined for all $n = 1, \ldots, d-1$.

It can be easily shown by means of analysis that

$$\forall z \geq 1. \Delta(z) \geq \Delta|_{z=1} = 0. \tag{B.79}$$

Thus, (B.70) holds, and retaining all m eigenvalues of S in the weight matrix of the *rotcaf* model leads to the same or higher likelihood $\mathcal{L}$ as retaining a smaller number of them. ■

Appendix C

Analytic Solution for the PoPPCA model

C.1 Verification of PPCA

We have seen that the inverse covariance of the PoE model equals to the summation of inverse covariances of each individual expert. Here we will find the derivatives of the log-likelihood $\mathcal{L}$, expressed through the inverse covariance C^{-1}, with respect to elements of the weight vector. We confirm that the resulting ML equation for w corresponds to Tipping and Bishop's results for 1-factor PPCA described in Tipping and Bishop (1997).

Consider expression (1.6) for the log-likelihood $\mathcal{L}$. From (1.8), the derivative of $\mathcal{L}$ w.r.t. the i^{th} element w_i of the weight vector of a 1-factor 1-expert PoPPCA is expressed as

$$\frac{\partial \mathcal{L}}{\partial w_i} = -\frac{N}{2}\frac{\partial \ln|\mathsf{C}|}{\partial w_i} - \frac{N}{2}\frac{\partial \mathrm{tr}[\mathsf{C}^{-1}\mathsf{S}]}{\partial w_i} = 0. \tag{C.1}$$

Applying differentiation rules for determinants and traces (Minka, 1999), (Krzanowski and Marriott, 1994), we get the system of equations for maximum-likelihood estimates of weight elements w_i:

$$\begin{cases} \mathrm{tr}[\mathsf{C}^{-1}\frac{\partial \mathsf{C}}{\partial w_1}] - \mathrm{tr}[\mathsf{C}^{-1}\frac{\partial \mathsf{C}}{\partial w_1}\mathsf{C}^{-1}\mathsf{S}] = 0 \\ \cdots \\ \mathrm{tr}[\mathsf{C}^{-1}\frac{\partial \mathsf{C}}{\partial w_d}] - \mathrm{tr}[\mathsf{C}^{-1}\frac{\partial \mathsf{C}}{\partial w_d}\mathsf{C}^{-1}\mathsf{S}] = 0 \end{cases}, \tag{C.2}$$

which should hold for all integer indices i from 1 to d.

It can be shown that elements of the matrices in the arguments of the traces on the l.h.s. of each equation in (C.2) may be expressed as

$$\left[\mathsf{C}^{-1}\right]_{kl} = \beta\delta_{kl} - \beta\gamma w_k w_l \tag{C.3}$$

$$\left[\frac{\partial \mathsf{C}}{\partial w_i}\right]_{kl} = w_k\delta_{li} + w_l\delta_{ki} \tag{C.4}$$

$$\left[\mathsf{C}^{-1}\mathsf{S}\right]_{kl} = \beta S_{kl} - \beta\sum_{p=1}^{d}\gamma w_k w_p S_{pl} = \beta\left(S_{kl} - \gamma w_k \mathsf{w}^{\mathrm{T}}\mathsf{s}_\mathsf{l}\right), \tag{C.5}$$

where β and γ are defined as in (3.4) and (3.5), δ_{ij} is Kronecker delta, and s_i is a column vector such that $\mathsf{S} = [\mathsf{s}_1\ \mathsf{s}_2\ \cdots\ \mathsf{s}_\mathsf{d}]$. Here and further $[\mathsf{A}]_{kl}$ denotes the element in the k^{th} row and l^{th} column of matrix A.

From (C.3) and (C.4),

$$\left[\mathsf{C}^{-1}\frac{\partial \mathsf{C}}{\partial w_i}\right]_{kl} = -\beta\gamma w_k w_l w_i + \beta w_l\delta_{ki} + \beta w_k\delta_{li} - \beta\gamma\|\mathsf{w}\|^2 w_k\delta_{li}. \tag{C.6}$$

Substituting these results into the system of equations (C.2), we obtain

$$\frac{\partial \ln|\mathsf{C}|}{\partial w_i} = 2\gamma w_i = 2\left[\mathsf{C}^{-1}\mathsf{S}\right]_i; \tag{C.7}$$

$$\frac{\partial \mathrm{tr}[\mathsf{C}^{-1}\mathsf{S}]}{\partial w_i} = -2\beta\gamma\mathsf{w}^T\left(\mathsf{s}_\mathsf{i} - \gamma w_i\sum_{k=1}^{d} w_k\mathsf{s}_\mathsf{k}\right) = -2\left[\mathsf{C}^{-1}\mathsf{S}\mathsf{C}^{-1}\mathsf{w}\right]_i. \tag{C.8}$$

The r.h.s. terms of (C.7) and (C.8) are due to

$$\begin{aligned} 2\gamma w_i = 2w_i(\beta - \beta\gamma\|\mathsf{w}\|^2) &= 2\sum_{f=1}^{d}(\beta\delta_{if} - \beta\gamma \mathsf{w}_i \mathsf{w}_f)\mathsf{w}_f && \text{(C.9)} \\ &= 2\left[\mathsf{C}^{-1}\mathsf{S}\right]_i, \\ -2\beta\gamma \mathsf{w}^T\left(\mathsf{s_i} - \gamma w_i \sum_{k=1}^{d} w_k \mathsf{s_k}\right) &= -2\left[\sum_{k=1}^{d}\beta\gamma\mathsf{w}^T(\delta_{ik}\mathsf{s}_k - w_i w_k \mathsf{s}_k)\right]_i && \text{(C.10)} \\ &= -2\sum_{k=1}^{d}\left[\mathsf{C}^{-1}\right]_{ik}\gamma\mathsf{w}^T\mathsf{s}_i \\ &= -2\sum_{k=1}^{d}\left[\mathsf{C}^{-1}\right]_{ik}(\beta - \beta\gamma\|\mathsf{w}\|^2)\mathsf{w}^T\mathsf{s}_i \\ &= -2\left[\mathsf{C}^{-1}\mathsf{S}\mathsf{C}^{-1}\mathsf{w}\right]_i. \end{aligned}$$

Then, solving (C.2) w.r.t. w_i and generalizing for w we can see that

$$\mathsf{C}^{-1}\mathsf{S}\mathsf{C}^{-1}\mathsf{w} = \mathsf{C}^{-1}\mathsf{w} \Rightarrow \mathsf{w} = \mathsf{S}\mathsf{C}^{-1}\mathsf{w}. \qquad \text{(C.11)}$$

Equation (C.11) corresponds to the one obtained in Tipping and Bishop (1997).

The same result could be obtained without explicit expression of matrix elements (C.3), (C.4), etc. by applying matrix differentiation rules as described in Minka (1999). For example,

$$\begin{aligned} d\ln|\mathsf{C}| &= \mathrm{tr}(\mathsf{C}^{-1}d\mathsf{C}) = \mathrm{tr}((\beta - \beta\gamma\mathsf{w}\mathsf{w}^T)d(\mathsf{w}\mathsf{w}^T)) && \text{(C.12)} \\ &= \mathrm{tr}[(\beta - \beta\gamma\mathsf{w}\mathsf{w}^T)((d\mathsf{w})\mathsf{w}^T)] + \mathrm{tr}[(\beta - \beta\gamma\mathsf{w}\mathsf{w}^T)\mathsf{w}(d\mathsf{w})^T] \\ &= 2(\beta\mathsf{w}^T - \beta\gamma\|\mathsf{w}\|^2\mathsf{w}^T). \end{aligned}$$

Then

$$\frac{d\ln|\mathsf{C}|}{d\mathsf{w}} = 2(\beta\mathsf{w}^T - \beta\gamma\|\mathsf{w}\|^2\mathsf{w}^T)^T = 2\gamma\mathsf{w}$$

[compare with (C.7)]. However, the derivative of the log-likelihood obtained in this way does not directly indicate how to get the r.h.s. parts of the equalities (C.9) and (C.10), which simplify the derivation of (C.11).

C.2 Equal Experts for Modeling 1-factor PPCA Data

We have shown [see section (3.3)] that some types of data, generated by the simplest (1-factor) PPCA model, cannot be modeled by an $(1, m)$ PoPPCA with a single retained weight. Here we show that this data can be modeled by m experts with equal weight vectors.

Let σ^2 be the noise variance of the data-generating PPCA model, and $\sigma_\Sigma^2 = \beta^{-1}$ be the noise variance of the $(1, m)$ PoPPCA model. Let us further assume $\mathsf{w}_1 = \mathsf{w}_2 = \ldots = \mathsf{w}_m$. Then, according to (3.5), $\gamma_1 = \gamma_2 = \ldots = \gamma_m$, and

$$\begin{aligned} \mathsf{C}_\Sigma^{-1} &= \beta m \mathsf{I}_d - m\beta\gamma_1 \frac{\mathsf{w}_1\mathsf{w}_1^T}{\|\mathsf{w}_1\|^2}\|\mathsf{w}_1\|^2 \\ &= \beta m \mathsf{I}_d - m\beta\gamma_1\|\mathsf{w}_1\|^2\mathsf{u}_1\mathsf{u}_1^T - \sum_{i=2}^{m}\mathsf{u}_i\mathsf{u}_i^T \cdot 0 \\ &= m\beta\sum_{i=2}^{m}\mathsf{u}_i\mathsf{u}_i^T + m\beta\gamma_1\|\mathsf{w}_1\|^2\mathsf{u}_1\mathsf{u}_1^T, \end{aligned} \tag{C.13}$$

where $[\mathsf{u}_1, \mathsf{u}_2, \ldots, \mathsf{u}_m]$ form an orthonormal basis of eigenvalues of S, and

$$\mathsf{u}_1 = \frac{\mathsf{w}_1}{\|\mathsf{w}_1\|}. \tag{C.14}$$

From (C.13) it is easy to see that the eigenvalues of C_Σ^{-1} are $\beta m - m\beta\gamma_1\|\mathsf{w}_1\|^2$, βm, $\ldots$, βm. If λ_Σ denotes an eigenvalue of C_Σ, then

$$\begin{aligned} \lambda_\Sigma^{-1} &= \beta m(1 - \|\mathsf{w}_1\|^2\gamma_1) \\ &\Rightarrow \gamma_1 = \left(1 - \frac{\lambda_\Sigma^{-1}}{\beta m}\right)\frac{1}{\|\mathsf{w}_1\|^2} \\ &\Rightarrow \gamma_1^{-1} = \left(\frac{\beta m - \lambda_\Sigma^{-1}}{\beta m\|\mathsf{w}_1\|^2}\right)^{-1}. \end{aligned} \tag{C.15}$$

On the other hand, (3.5) implies

$$\gamma_1^{-1} = \|\mathsf{w}_1\|^2 + \beta^{-1}. \tag{C.16}$$

Then the norm of the retained weight is expressed from

$$\beta m \|\mathsf{w}_1\|^2 = (\|\mathsf{w}_1\|^2 + \beta^{-1})(\beta m - \lambda_\Sigma^{-1}). \tag{C.17}$$

Then (C.17) $\Rightarrow$

$$\|\mathsf{w}_1\|^2 \;\; = \;\; m\lambda_\Sigma - \sigma_\Sigma^2. \tag{C.18}$$

Considering (3.11), the parameters of $(1,m)$ PoPPCA modeling 1-factor PPCA data exactly can be expressed as

$$\begin{cases} \mathsf{w}_1 = \ldots = \mathsf{w}_m = \sqrt{m(\lambda - \sigma^2)}\mathsf{u}_1/\|\mathsf{u}_1\| \\ \sigma_\Sigma^2 = m\sigma^2, \end{cases} \tag{C.19}$$

where λ is the largest eigenvalue of C and u_1 is its corresponding eigenvector. From the first equation of the system we can see that since $\lambda > \sigma^2$, the weights remain real-valued and have no constraints on their magnitudes.

Moreover, if the rotation factor is ignored in ML solution for the weight matrix W in PPCA, then each retained weight is collinear with one of the eigenvectors of S. For 1-factor PPCA this case results in collinearity of eigenvector u_1 of the PPCA model covariance C and eigenvector $\tilde{\mathsf{u}}_1$ of S. Then (C.19) can be re-written as

$$\begin{cases} \mathsf{w}_1 = \ldots = \mathsf{w}_m = \sqrt{m(\lambda - \sigma^2)}\tilde{\mathsf{u}}_1/\|\tilde{\mathsf{u}}_1\| \\ \sigma_\Sigma^2 = m\sigma^2. \end{cases} \tag{C.20}$$

An $(1,m)$ PoPPCA model with parameters (C.20) is one way of modeling a 1-factor PPCA exactly. We expect that non-linear optimization algorithms should not diverge from the found point in the parameter space of $(1,m)$ PoPPCA.

C.3 ML Parameter Estimators for PoPPCA with Implicitly Scaled Weights

C.3.1 ML Equations for Implicitly Scaled Weights

Here we will consider ML equations for the model with implicitly scaled weights, confirm that the covariance C_Σ remains positive-definite, and find an analytic expression for the original weights of the PoPPCA model.

The differentials of the determinant and the trace in expression (1.8) for derivatives of the log-likelihood $\mathcal{L}$ w.r.t. parameters can be expressed by applying rules for matrix differentials, described in Minka (1999):

$$\begin{aligned}
\mathrm{d}_{\tilde{\mathsf{W}}} \ln|\mathsf{C}_\Sigma^{-1}| &= -\mathrm{tr}(\mathsf{C}_\Sigma(\mathrm{d}_{\tilde{\mathsf{W}}}\tilde{\mathsf{W}})\tilde{\mathsf{W}}^T - \mathsf{C}_\Sigma\tilde{\mathsf{W}}\mathrm{d}_{\tilde{\mathsf{W}}}(\tilde{\mathsf{W}})^T) \\
&= -2\mathrm{tr}(\tilde{\mathsf{W}}^T\mathsf{C}_\Sigma(\mathrm{d}_{\tilde{\mathsf{W}}}\tilde{\mathsf{W}})); && \text{(C.21)} \\
\mathrm{d}_{\tilde{\mathsf{W}}}\ \mathrm{tr}(\mathsf{C}_\Sigma^{-1}\mathsf{S}) &= -\mathrm{tr}((\mathrm{d}_{\tilde{\mathsf{W}}}\tilde{\mathsf{W}})\tilde{\mathsf{W}}^T\mathsf{S} - \tilde{\mathsf{W}}(\mathrm{d}_{\tilde{\mathsf{W}}}\tilde{\mathsf{W}})^T\mathsf{S}) && \text{(C.22)} \\
&= -2\mathrm{tr}(\tilde{\mathsf{W}}^T\mathsf{S}(\mathrm{d}_{\tilde{\mathsf{W}}}\tilde{\mathsf{W}})); && \text{(C.23)} \\
\mathrm{d}_x \ln|\mathsf{C}_\Sigma^{-1}| &= -\mathrm{tr}(\beta m\mathsf{C}_\Sigma\ \mathrm{d}_x x); && \text{(C.24)} \\
\mathrm{d}_x\ \mathrm{tr}(\mathsf{C}_\Sigma^{-1}S) &= -\mathrm{tr}(\beta m\mathsf{S}\ \mathrm{d}_x x). && \text{(C.25)}
\end{aligned}$$

As in (B.2.2), we make a simplifying assumption that $\tilde{\mathsf{W}}$ is independent of x. Differentiating with respect to $\tilde{\mathsf{W}}$ and x, we obtain expressions for the derivatives of the log-likelihood $\mathcal{L}$:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\tilde{\mathsf{W}}} &= N(\mathsf{S}\tilde{\mathsf{W}} - \mathsf{C}_\Sigma\tilde{\tilde{\mathsf{W}}}) && \text{(C.26)} \\
\frac{\partial\mathcal{L}}{\partial x} &= Nm\beta\mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}). && \text{(C.27)}
\end{aligned}$$

Now we can drop the constant factors and obtain the ML equations for the expert weights W and the inverse noise variance β.

$$\begin{aligned}
\mathsf{S}\tilde{\mathsf{W}} - \mathsf{C}_\Sigma\tilde{\mathsf{W}} &= 0 && \text{(C.28)} \\
\mathrm{tr}(\mathsf{C}_\Sigma - \mathsf{S}) &= 0. && \text{(C.29)}
\end{aligned}$$

Note the similarity of these equations with (B.26) and (B.27) obtained for the *rotcaf* model. However, as we will show in the following sections, even though PoCG and PoPPCA result in identical equations for the ML weight matrices, the analytic solutions of the latter are more difficult due to constraints on the weight magnitudes.

C.3.2 ML Solutions for Implicitly Scaled Weights

As in PoCG, there are two possible cases of non-zero maximum-likelihood solutions for the the PoPPCA weight matrix W with exact and approximate model covariance matrices C_Σ.

Applying SVD for the scaled weights $\tilde{\mathsf{W}} = \mathsf{ULR}^T$ and performing manipulations similar to the ones described in section B.3, we obtain

$$\begin{aligned} \mathsf{C}_\Sigma \mathsf{UL} = \mathsf{SUL} &= \beta_\Sigma^{-1}(\mathsf{I}_d + \mathsf{ULR}^T(\beta_\Sigma \mathsf{I}_m - \mathsf{RL}^2\mathsf{R}^T)^{-1}\mathsf{RLU}^T)\mathsf{UL} \\ &= \mathsf{U}(\beta_\Sigma^{-1}\mathsf{I} + \beta_\Sigma^{-1}\mathsf{LR}^T(\beta_\Sigma \mathsf{I}_m - \mathsf{RL}^2\mathsf{R}^T)^{-1}\mathsf{RL})\mathsf{L} \\ &= \mathsf{U}(\beta_\Sigma \mathsf{I}_d - \mathsf{L}^2)^{-1}\mathsf{L}, \end{aligned} \tag{C.30}$$

where $\beta_\Sigma = \beta m$ is the inverse noise variance for $(1, m)$ PoPPCA model. The last step of (C.30) is due to the reverse application of the Woodbury transform (B.8) with $\mathsf{A} = \beta_\Sigma \mathsf{I}$, $\mathsf{B} = \mathsf{LR}^T$, $\mathsf{D} = \mathsf{I}_m$, $\mathsf{C} = -\mathsf{RL}$.

Diagonality of $\beta_\Sigma \mathsf{I}_d - \mathsf{L}^2$ and non-singularity of L imply

$$\mathsf{U} = \mathsf{U}_S \tag{C.31}$$

$$\tilde{\mathsf{W}} = \mathsf{U}_S(\beta_\Sigma \mathsf{I}_m - \tilde{\Lambda})^{1/2}\mathsf{R}^T. \tag{C.32}$$

U_S is composed of eigenvectors of the sample covariance S, R is a rotation matrix, $\tilde{\Lambda}_i = diag(\tilde{\lambda}_1, \dots, \tilde{\lambda}_m)$ is a diagonal matrix, such that

$$\tilde{\lambda}_i = \begin{cases} \lambda_i^{-1} & \text{if } \mathsf{u}_i \text{ is retained} \\ \beta_\Sigma & \text{if } \mathsf{u}_i \text{ is discarded,} \end{cases} \tag{C.33}$$

and λ_i is an eigenvalue of S, corresponding to the column vector u_i of U_S.

From expression (3.15) for the inverse covariance, C_Σ^{-1} can be expressed as

$$\begin{aligned}
\mathsf{C}_\Sigma^{-1} &= \beta_\Sigma \mathsf{I}_d - \mathsf{U}_S\mathsf{U}_S^T\beta_\Sigma + \mathsf{U}_S\tilde{\Lambda}\mathsf{U}_S^T \\
&= \sum_{i=1}^{d} \mathsf{u}_i\mathsf{u}_i^T\beta_\Sigma + \sum_{i=1}^{r} \mathsf{u}_i(\lambda_i^{-1} - \beta_\Sigma)\mathsf{u}_i^T + \sum_{i=r+1}^{d} \mathsf{u}_i\mathsf{u}_i^T \cdot 0 \\
&= \sum_{i=1}^{r} \mathsf{u}_i\mathsf{u}_i^T\lambda_i^{-1} + \sum_{i=r+1}^{d} \mathsf{u}_i\mathsf{u}_i^T\beta_\Sigma, \qquad \text{(C.34)}
\end{aligned}$$

where r is the number of eigenvectors of S retained in the left singular vectors of the scaled weight matrix $\tilde{\mathsf{W}}$. Note the correspondence of this identity to (B.51) for the PoCG model.

We can also notice that since the eigenvalues of the sample covariance $\lambda_i > 0$, and the noise parameter $\beta_\Sigma > 0$, both C_Σ^{-1} and C_Σ are positive-definite and symmetric. Thus, C_Σ is a valid covariance of a Gaussian distribution, and no explicit constraint inequality optimization on $\tilde{\mathsf{W}}$ is necessary.

Analogously to Appendix B.3.2 and Appendix B.4, we can verify that the ML solution for the scaled weight matrix $\tilde{\mathsf{W}}$ corresponds to the solution for the factor loading matrix in the PPCA model (Tipping and Bishop, 1997). This implies that the likelihood is maximized when the weight matrix $\tilde{\mathsf{W}}$ consists of m scaled and rotated eigenvectors of S. The expert weight matrix can be rewritten as

C.3.3 Conditions on Real Weights of the PoPPCA Model

Since Γ is a non-singular matrix, the expression for the implicitly scaled weights (3.16) implies

$$\begin{aligned}
\mathsf{W} &= \tilde{\mathsf{W}}\Gamma^{-1/2}\beta^{-1/2} \Rightarrow \qquad \text{(C.35)} \\
\mathsf{w}^{(i)} &= \tilde{\mathsf{w}}^{(i)}\beta^{-1/2}\gamma_i^{-1/2}, \qquad \text{(C.36)}
\end{aligned}$$

where each element of Γ^{-1} can be found from (3.5) as $\gamma_i^{-1} = \beta^{-1} + ||\mathsf{w}^{(i)}||^2$.

To express the weight vector of the i^{th} expert $\mathsf{w}^{(i)}$ through the scaled weight $\tilde{\mathsf{w}}^{(i)}$, found in the previous sections, we express the norm $||\mathsf{w}^{(i)}||^2$ and the scaling factor γ_i^{-1} through $\tilde{\mathsf{w}}^{(i)}$, substitute the result into (C.36), and renormalize. This leads to

$$\begin{aligned} \mathsf{w}^{(i)} &= \tilde{\mathsf{w}}^{(i)}(\beta^{-1} + ||\mathsf{w}^{(i)}||^2)^{1/2}\beta^{-1/2} \Rightarrow \\ ||\mathsf{w}^{(i)}||^2 &= ||\tilde{\mathsf{w}}^{(i)}||^2(\beta^{-1} + ||\mathsf{w}^{(i)}||^2)\beta^{-1}. \qquad \text{(C.37)} \end{aligned}$$

Note that if $\beta = ||\tilde{\mathsf{w}}_i||^2$, then (C.37) $\Rightarrow$

$$\begin{aligned} ||\mathsf{w}^{(i)}||^2 &= \beta(\beta^{-1} + ||\mathsf{w}^{(i)}||^2)\beta^{-1} \qquad \text{(C.38)} \\ &= \frac{1}{||\tilde{\mathsf{w}}^{(i)}||^2} + ||\mathsf{w}^{(i)}||^2 \qquad \text{(C.39)} \\ &\Rightarrow \lim_{\beta \to ||\mathsf{w}^{(i)}||^2} ||\mathsf{w}^{(i)}||^2 \to \infty \\ &\Rightarrow \beta \to \infty \Rightarrow \sigma^2 \to 0. \qquad \text{(C.40)} \end{aligned}$$

We see that (C.39) does not have finite solutions on the complex plane. Thus, we will consider the cases when $\sigma^2 \neq 0$, i.e. $\beta \neq ||\mathsf{w}^{(i)}||^2$.

From (C.37) we can find that

$$\begin{aligned} ||\mathsf{w}^{(i)}||^2 &= \frac{\beta^{-2}||\tilde{\mathsf{w}}^{(i)}||^2}{1 - \beta^{-1}||\tilde{\mathsf{w}}^{(i)}||^2} \qquad \text{(C.41)} \\ \gamma_i^{-1} &= \beta^{-1} + \frac{\beta^{-2}||\tilde{\mathsf{w}}^{(i)}||^2}{1 - \beta^{-1}||\tilde{\mathsf{w}}^{(i)}||^2} = \frac{\beta^{-1}}{1 - \beta^{-1}||\tilde{\mathsf{w}}^{(i)}||^2} \qquad \text{(C.42)} \end{aligned}$$

Then the unscaled weight of the i^{th} expert of the PoPPCA is expressed as

$$\mathsf{w}^{(i)} = \frac{\tilde{\mathsf{w}}^{(i)}\beta^{-1}}{\sqrt{1 - \beta^{-1}||\tilde{\mathsf{w}}^{(i)}||^2}}. \qquad \text{(C.43)}$$

Note that the scaling factor of the i^{th} expert is real when

$$||\tilde{\mathsf{w}}^{(i)}||^2 < \beta \qquad \text{(C.44)}$$

for each retained eigenvalue λ_i of the sample covariance S.

In case when $\mathsf{R} = \mathsf{I}$, this implies that the following conditions must hold for $\mathsf{w}^{(i)}$ to be real:

$$diag(\tilde{\mathsf{W}}^T\tilde{\mathsf{W}}) - \mathsf{I}_m\beta < 0 \Rightarrow \tag{C.45}$$

$$\forall i = 1, \ldots, m. \ \frac{1}{\lambda_i} > (m-1)\beta. \tag{C.46}$$

Substituting β_{ML} from (3.21), we obtain the range for eigenvalue λ_i of the sample covariance resulting in a real-valued weight $\mathsf{w}^{(i)}$:

$$\begin{cases} \lambda_i > 0 & \text{if } m = 1 \\ 0 < \bar{\lambda} \leq \lambda_i < m\bar{\lambda}/(m-1) & \text{if } m \geq 2 \end{cases} \tag{C.47}$$

where λ_i is the i^{th} retained eigenvalue, $\bar{\lambda}$ is the mean of the discarded eigenvalues, m is the number of experts. In other cases, $\mathsf{w}^{(i)}_{ML}$ is an imaginary weight in d dimensions.

The lower bound of the last condition of (C.47) follows from the fact that the ML solution for the PoPPCA model implies that each retained eigenvalue is greater than the mean of the discarded eigenvalues [the proof is similar to the one in Appendix B.4.2].

C.4 ML Parameter Estimators for PoPPCA with Unscaled Weights

C.4.1 Expressions of Derivatives of the Log-Likelihood

In this section we derive expressions for derivatives of the log-likelihood $\mathcal{L}$ with respect to parameters W and β of the experts' distribution. Expanding the expression for the inverse covariance (3.6),

$$\left[\mathsf{C}_\Sigma^{-1}\right]_{kl} = m\beta\delta_{kl} - \beta\sum_{f=1}^{m}\gamma_f w_k^{(f)} w_l^{(f)}. \tag{C.48}$$

Now we will express the derivatives with respect to the weights and the noise parameter.

Derivation of $\partial\mathcal{L}/\partial\mathsf{W}$

To find the expression for $\partial\mathcal{L}/\partial\mathsf{W}$, we will estimate it with respect to each weight of each expert. For the p^{th} expert

$$\begin{aligned}
\frac{\partial \ln|\mathsf{C}_\Sigma|}{\partial w_i^{(p)}} &= \operatorname{tr}\left[{\mathsf{C}_\Sigma}^{-1}\frac{\partial \mathsf{C}_\Sigma}{\partial w_i^{(p)}}\right] = -\operatorname{tr}\left[\mathsf{C}_\Sigma^{-1}\mathsf{C}_\Sigma\frac{\partial {\mathsf{C}_\Sigma}^{-1}}{w_i^{(p)}}\mathsf{C}_\Sigma\right] \\
&= -\operatorname{tr}\left[\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial w_i^{(p)}}\mathsf{C}_\Sigma\right] \qquad \text{(C.49)}\\
\frac{\partial \operatorname{tr}[{\mathsf{C}_\Sigma}^{-1}\mathsf{S}]}{\partial w_i^{(p)}} &= \operatorname{tr}\left[\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial w_i^{(p)}}\mathsf{S}\right]. \qquad \text{(C.50)}
\end{aligned}$$

Then the derivative of the log-likelihood $\mathcal{L}$ with respect to the i^{th} weight of the p^{th} expert can be expressed as

$$\frac{\partial\mathcal{L}}{\partial w_i^{(p)}} = \frac{N}{2}\left(\operatorname{tr}\left[\frac{\partial {\mathsf{C}_\Sigma}^{-1}}{w_i^{(p)}}\mathsf{C}_\Sigma\right] - \operatorname{tr}\left[\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial w_i^{(p)}}\mathsf{S}\right]\right). \qquad \text{(C.51)}$$

Since γ_p is a function of w_p, the derivative $\partial\gamma_p/\partial\mathsf{w}_p$ is generally non-zero. Then, the elements of the derivative of the inverse covariance with respect to w_p are expressed as follows:

$$\begin{aligned}
\left[\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial w_i^{(p)}}\right]_{kl} &= -\beta\frac{\partial w_k^{(p)} w_l^{(p)}\gamma_p}{\partial w_i^{(p)}} \\
&= \beta\gamma_p\left[2\gamma_p w_i^{(p)} w_k^{(p)} w_l^{(p)} - [w_k^{(p)}\delta_{li} + w_l^{(p)}\delta_{ki}]\right]. \qquad \text{(C.52)}
\end{aligned}$$

Transforming to the vector form, the diagonal elements are

$$\left[\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial w_i^{(p)}}\mathsf{C}_\Sigma\right]_{kk} = (2\beta\gamma_p^2 w_i^{(p)} w_k^{(p)}\mathsf{w}^{(p)T} - \beta\gamma_p w_k^{(p)}\Delta_i^T - \beta\gamma_p\delta_{ki}\mathsf{w}^{(p)T})\mathsf{c}_k, \qquad \text{(C.53)}$$

where $\mathsf{C}_\Sigma = [\mathsf{c}_1\ \mathsf{c}_2\ \dots\ \mathsf{c}_d]$, and $\Delta_i = [\delta_{1i}\ \delta_{2i}\ \dots\ \delta_{di}]^T$.

After analogous transformation for the derivative of the traces (C.50), the system of ML expression for the weights (C.2) is transformed to

$$\begin{cases} \partial\mathcal{L}/\partial w_1^{(p)} = 2\beta\gamma_p[\beta\gamma_p w_1^{(p)}\mathsf{w}^{(p)T}\sum_{k=1}^d w_k^{(p)}\mathsf{c}_k - \beta\mathsf{w}^{(p)T}\mathsf{c}_1 - \\ \qquad \beta\gamma_p w_1^{(p)}\mathsf{w}^{(p)T}\sum_{k=1}^d w_k^{(p)}\mathsf{s}_k + \beta\mathsf{w}^{(p)T}\mathsf{s}_1] \\ \qquad\qquad \cdots \\ \partial\mathcal{L}/\partial w_d^{(p)} = 2\beta\gamma_p[\beta\gamma_p w_d^{(p)}\mathsf{w}^{(p)T}\sum_{k=1}^d w_k^{(p)}\mathsf{c}_k - \beta\mathsf{w}^{(p)T}\mathsf{c}_d - \\ \qquad \beta\gamma_p w_d^{(p)}\mathsf{w}^{(p)T}\sum_{k=1}^d w_k^{(p)}\mathsf{s}_k + \beta\mathsf{w}^{(p)T}\mathsf{s}_d] \end{cases} \tag{C.54}$$

for all experts $p \in [1, m]$.

To see whether this system leads to the same result as system (C.2) in case of a single expert, we can notice that

$$m = 1 \Rightarrow \mathsf{c}_i = w_i\mathsf{w} + \beta^{-1}\Delta_\mathsf{i},$$

and, by substitution, obtain the left and the right hand sides of (C.2) in the l.h.s. and the r.h.s of (C.54). Thus, the expression of the derivative of the log-likelihood $\mathcal{L}$ w.r.t. to the weights in the multi-expert PoPPCA model is consistent with the correspondent expression for PPCA.

Since for any matrix S and vector w

$$\sum_{k=1}^d \mathsf{s}_k w_k^{(p)} = \mathsf{S}\mathsf{w}^{(p)}, \tag{C.55}$$

each equation in (C.54) can be re-written as

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial w_i^{(p)}} &= 2\beta\gamma_p\mathsf{w}^{(p)T}\left(\gamma_p w_i^{(p)}\mathsf{C}_\Sigma\mathsf{w}^{(p)} - \mathsf{c}_i\right) \\ &- 2\beta\gamma_p\mathsf{w}^{(p)T}\left(\gamma_p w_i^{(p)}\mathsf{S}\mathsf{w}^{(p)} - \mathsf{s}_i\right). \end{aligned} \tag{C.56}$$

This is the expression of the derivative of log-likelihood $\mathcal{L}$ w.r.t. the i^{th} weight of the p^{th} expert at each point of the $\mathcal{R}^{d\times m}$ space.

Derivation of $\partial\mathcal{L}/\partial\beta$

To complete expression of the ML solution for the PoPPCA model, we need to express the derivative of the log-likelihood with respect to the noise parameter

β. Since $\beta = 1/\sigma^2 > 0$, we need to ensure that $\beta_{ML} > 0$. In order to do this, we introduce a new parameter x, such that

$$\beta = e^x. \tag{C.57}$$

Then, due to the monotonic growth of the ln function,

$$\beta_{ML} = e^{x_{ML}}.$$

Similarly to what has been discussed in section B.2.2, we may assume independence of W and the noise parameter β. Thus

$$\begin{aligned} \frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial x} &= e^x \mathsf{W}\Gamma\mathsf{W}^T + \beta\mathsf{W}\frac{\partial\Gamma}{\partial x}\mathsf{W}^T \\ &= e^x m \mathsf{I}_d - \mathsf{W}(e^x\Gamma + \Gamma^2)\mathsf{W}^T \\ &= \mathsf{C}_\Sigma^{-1} - \mathsf{W}\Gamma^2\mathsf{W}^T. \end{aligned} \tag{C.58}$$

By analogy with the expression of the ML solution (C.2) and (C.51), we see that

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x} &= \mathrm{tr}\left[\frac{\partial {\mathsf{C}_\Sigma}^{-1}}{\partial x}\mathsf{C}_\Sigma\right] - \mathrm{tr}\left[\frac{\partial \mathsf{C}_\Sigma^{-1}}{\partial x}\mathsf{S}\right] = 0 \\ \Rightarrow \frac{\partial \mathcal{L}}{\partial x} &= \mathrm{tr}[\mathsf{I}_d - \mathsf{C}_\Sigma\mathsf{W}\Gamma^2\mathsf{W}^T - \mathsf{C}_\Sigma^{-1}\mathsf{S} + \mathsf{W}\Gamma^2\mathsf{W}^\mathrm{T}\mathsf{S}] = 0. \end{aligned} \tag{C.59}$$

Substituting the inverse covariance matrix C_Σ^{-1} from (3.6) into (C.59), we obtain the expression for the noise parameter β. Analytical solution of (C.59) w.r.t. β is non-trivial, and it is not necessary to derive it at this stage. However, the solution can be estimated numerically [e.g. by performing gradient ascent on $\mathcal{L}$ with the gradients expressed by (C.59) and (C.56)].

C.4.2 ML Equations for Unscaled Weights

In this section we will find the ML equation for the weight matrix of the PoPPCA model. As before, in order to derive ML equations for the expert weight matrix, we will set all derivatives of $\mathcal{L}$ w.r.t. its parameters to zero.

To simplify the form of the ML equation for the weight matrix W, we will first represent $\partial\mathcal{L}/\partial w_i^{(p)}$, derived in Appendix C.4.1, in the matrix form. Then we will introduce an auxiliary variable simplifying the form of the ML equation, investigate the number of possible solutions, and finally obtain the ML equation for W.

Derivation of the Auxiliary ML Equation

From (C.54) and (C.56) it can be shown that

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial w_i^{(p)}} &= 0 \\ &\Rightarrow \gamma_p^2 w_i^{(p)} \mathsf{w}^{(p)T}\mathsf{C}_\Sigma \mathsf{w}^{(p)} - \gamma_p \mathsf{c}_i^T \mathsf{w}^{(p)} = \\ &\quad\; \gamma_p^2 w_i^{(p)} \mathsf{w}^{(p)T}\mathsf{S}\mathsf{w}^{(p)} - \gamma_p \mathsf{s}_i^T \mathsf{w}^{(p)} && \text{(C.60)} \\ &\Rightarrow \gamma_p(\gamma_p \mathsf{w}^{(p)}\mathsf{w}^{(p)T} - \mathsf{I}_d)(\mathsf{C}_\Sigma - \mathsf{S})\mathsf{w}^{(p)} = 0. && \text{(C.61)} \end{aligned}$$

In the matrix form this can be written as

$$\mathsf{W}diag(\Gamma\mathsf{W}^T(\mathsf{C}_\Sigma - \mathsf{S})\mathsf{W}\Gamma) - (\mathsf{C}_\Sigma - \mathsf{S})\mathsf{W}\Gamma = 0. \tag{C.62}$$

Expression (C.62) can be rewritten as

$$\mathsf{W}\Gamma diag(\mathsf{W}^T\mathsf{Q}) - \mathsf{Q} \;=\; 0, \tag{C.63}$$

where $\mathsf{Q} = [\mathsf{q}^{(1)} \ldots \mathsf{q}^{(m)}] = (\mathsf{C}_\Sigma - \mathsf{S})\mathsf{W}\Gamma$.

If we write (C.63) in the vector form, we can see that it leads to

$$\boxed{\forall p = 1, \ldots, m.\ (\gamma_p \mathsf{w}^{(p)}\mathsf{w}^{(p)T} - \mathsf{I}_d)\mathsf{q}^{(p)} = \mathsf{0}}\,, \tag{C.64}$$

which may be viewed as a system of linear equations for $\mathsf{q}^{(p)}$. From Cramer's rule, it has non-degenerate solutions for $\mathsf{q}^{(p)}$ only when

$$|\gamma_p \mathsf{w}^{(p)}\mathsf{w}^{(p)T} - \mathsf{I}_d| = 0. \tag{C.65}$$

Investigation of the Number of Possible Solutions

To solve the system of linear equations (C.64) for all elements of $\mathsf{q}^{(p)}$, we will first estimate the number of possible solutions of (C.64). This number is equal to the rank of the $d \times d$ matrix $\gamma_p \mathsf{w}^{(p)} \mathsf{w}^{(p)T} - \mathsf{I}_d$ (in this case, it is the number of meaningful equations in the system).

The null-vector columns do not contribute to an increase in the rank of a matrix, since they are linearly dependent on any other vector, This can be shown by the following lemma:

Lemma 1: $\forall \mathsf{a} \subset \mathcal{R}^d$, a *and* $\mathsf{0} \in \mathcal{R}^d$ *are linearly dependent.*

□ $\forall \mathsf{a} \subset \mathcal{R}^d$, a null vector $\mathsf{0}$ and a accept 0 as their linear combination, i.e.

$$\exists \lambda_1, \lambda_2 \in \mathcal{R}. \lambda_1 \neq 0.\ \lambda_1 \mathsf{0} + \lambda_2 \mathsf{a} = 0.$$

Therefore, from the definition of linear dependence $\mathsf{0}$ and a are linearly dependent. ■

To estimate the number of solutions of the system, let us denote the leftmost term in the brackets of (C.64) as A:

$$\mathsf{A} = [\mathsf{a}_1 \ldots \mathsf{a}_d] = \gamma_p \mathsf{w}^{(p)} \mathsf{w}^{(p)T} \subset \mathcal{R}^{d \times d}. \tag{C.66}$$

Let us introduce an auxiliary matrix $\tilde{\mathsf{A}}$ such that

$$\tilde{\mathsf{A}} = [\tilde{\mathsf{a}}_1, \ldots, \tilde{\mathsf{a}}_{d-l}] = \gamma_p \tilde{\mathsf{w}} \tilde{\mathsf{w}}^{\mathsf{T}} \subset \mathcal{R}^{(d-l)\times(d-l)}, \tag{C.67}$$

where l is the number of $\mathsf{0}$-elements in $\mathsf{w}^{(p)}$, and $\tilde{\mathsf{w}}$ is constructed from the expert weight $\mathsf{w}^{(p)}$ by eliminating l of its $\mathsf{0}$-elements. From lemma 1,

$$rank(\mathsf{A}) = rank(\tilde{\mathsf{A}}) \leq d - l. \tag{C.68}$$

We can see that every column of A is expressed as

$$\forall i \in [1, d-l].\ \tilde{\mathsf{a}}_i = \gamma_p \tilde{w}_i{}^{(p)} \tilde{\mathsf{w}}^{(p)} = \begin{bmatrix} \tilde{w}_i \tilde{w}_1 \gamma_p \\ \vdots \\ \tilde{w}_i \tilde{w}_d \gamma_p \end{bmatrix}. \tag{C.69}$$

Since $\tilde{w}_1 \neq 0, \ldots, \tilde{w}_{d-l} \neq 0$, this can be re-written as

$$\tilde{\mathsf{a}}_i = \frac{\tilde{w}_i}{\tilde{w}_k} \tilde{\mathsf{a}}_k. \tag{C.70}$$

Thus, any column a_i of $\tilde{\mathsf{A}}$ is linearly dependent on some other column a_k, therefore

$$rank(\mathsf{A}) = rank(\tilde{\mathsf{A}}) = 1. \tag{C.71}$$

In other words, A has only 1 non-zero singular value, and system (C.64) may have at most one non-degenerate solution for $\mathsf{q}^{(p)}$.

Derivation of the ML Equation for W

Here we will expand (C.64) and derive the ML equation for W.

It is known that rank of a matrix is the number of its non-zero singular vectors. Then, applying SVD to A we get

$$\mathsf{A} = \mathsf{U}_A \Lambda_A \mathsf{U}_A^T \quad = \quad [\mathsf{r}_1 \mathsf{r}_2 \ldots \mathsf{r}_{d-1} \mathsf{u}] \begin{bmatrix} 0 & \ldots & 0 \\ 0 & \cdots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & \lambda \end{bmatrix} \begin{bmatrix} \mathsf{r}_1{}^T \\ \vdots \\ \mathsf{r}_{\mathsf{d}-1}{}^T \\ \mathsf{u}^T \end{bmatrix}, \tag{C.72}$$

where $\mathsf{U}_A \subset \mathcal{R}^{d \times d}$ is an orthogonal, $\Lambda_A \subset \mathcal{R}^{d \times d}$ is a diagonal matrices, and $\mathsf{r}_1, \ldots, \mathsf{r}_{d-1}$ are vectors in the orthonormal basis of u in $\mathcal{R}^d$. They are scaled by the factor of 0 and, provided the condition of orthonormality holds, can be arbitrary.

Then,

$$\mathsf{A} \quad = \quad \mathsf{w}^{(p)} \mathsf{w}^{(p)T} \gamma_p = \mathsf{u}\mathsf{u}^T \lambda \subset \mathcal{R}^{d \times d}. \tag{C.73}$$

Orthonormality of U_A implies

$$\mathsf{w}^{(p)} = \mathsf{u}\sqrt{\frac{\lambda}{\gamma_p}} \tag{C.74}$$

$$\lambda = \|\mathsf{w}^{(p)}\|^2\gamma_p. \tag{C.75}$$

Thus, condition (C.65) for non-degenerate solutions for $\mathsf{q}^{(p)}$ can be written as

$$\begin{aligned} |\gamma_p \mathsf{w}^{(p)}\mathsf{w}^{(p)T} - \mathsf{I}_d| &= |\mathsf{U}_A| \begin{vmatrix} -\mathsf{I}_{d-1} & \mathsf{0}_{(d-1)\times 1} \\ \mathsf{0}_{1\times(d-1)} & \gamma_p\|\mathsf{w}^{(p)}\|^2 - 1 \end{vmatrix} |\mathsf{U}_A^T| \\ &= (\gamma_p\|\mathsf{w}^{(p)}\|^2 - 1)(-1)^{(d-1)} = 0. \end{aligned} \tag{C.76}$$

However, $\gamma_p\|\mathsf{w}^{(p)}\|^2 - 1 \neq 0$ for $\sigma^2 \neq 0$, i.e. system (C.64) only has the degenerate solution $\mathsf{q}^{(p)} = 0$, i.e.

$$(\mathsf{C}_\Sigma - \mathsf{S})\mathsf{W}\Gamma = 0. \tag{C.77}$$

C.4.3 ML Solutions for Parameter Estimates

From the definition (C.63) of Q and non-singularity of the scaling matrix Γ, all the possible answers of (C.64) can be expressed as

$$\begin{aligned} \forall p = 1, \dots, m.\ \mathsf{q}^{(p)} = 0 \ &\Leftrightarrow \ \mathsf{Q} = 0 \Leftrightarrow \\ (\mathsf{C} - \mathsf{S})\mathsf{W}\Gamma = 0. \ &\Rightarrow \left[\begin{array}{l} \mathsf{SW} = \mathsf{C}_\Sigma\mathsf{W}, \mathsf{W} \neq 0, \mathsf{S} \neq \mathsf{C}_\Sigma \\ \ \mathsf{S} \ = \qquad \mathsf{C}_\Sigma \\ \ \mathsf{W} \ = \qquad 0 \end{array} \right. \end{aligned} \tag{C.78}$$

Notice that these are exactly the same classes of solutions as the one obtained in section 2.2.2 for the PoCG model. However, we will show that the solution of (C.78) is generally more difficult due to the scaling factor in the expression for C_Σ in PoPPCA. As for the PoCG case, we can consider two non-zero classes of solutions.

Exact Model Covariance

If C_Σ is exact, then due to non-singularity of C_Σ and S,

$$\mathsf{C}_\Sigma^{-1} = \mathsf{S}^{-1}. \tag{C.79}$$

From the definition of the inverse covariance (3.6) we see that this implies

$$\tilde{\mathsf{W}}\tilde{\mathsf{W}}^T = \beta_\Sigma \mathsf{I}_d - \mathsf{S}^{-1}. \tag{C.80}$$

This has a known solution

$$\tilde{\mathsf{W}} = \mathsf{U}_S(\beta_\Sigma \mathsf{I}_m - \Lambda_S^{-1})^{1/2}\mathsf{R}^T. \tag{C.81}$$

Approximate Model Covariance

In this case

$$\mathsf{SW} \;=\; \mathsf{C}_\Sigma \mathsf{W}, \mathsf{W} \neq 0, \mathsf{S} \neq \mathsf{C}_\Sigma. \tag{C.82}$$

From the compact form of SVD[1], we can express the matrix W as

$$\mathsf{W} = \mathsf{ULR}^T, \tag{C.83}$$

where $\mathsf{U} \subset \mathcal{R}^{d\times m}$ is a matrix of orthonormal column vectors, $\mathsf{L} \subset \mathcal{R}^{m\times m}$ is a diagonal, and $\mathsf{R} \subset \mathcal{R}^{m\times m}$ is a rotation matrix. Substituting into (C.82), this implies

$$\mathsf{SUL} \;=\; \mathsf{C}_\Sigma \mathsf{UL}. \tag{C.84}$$

Note that the previous equation can be rewritten as

$$[\mathsf{Su}_1 l_1 \ldots \mathsf{Su}_m l_m] \;=\; [\mathsf{C}_\Sigma \mathsf{u}_1 l_1 \ldots \mathsf{C}_\Sigma \mathsf{u}_m l_m], \tag{C.85}$$

from where it can be clearly seen that if $l_i = 0$, the corresponding left singular vector u_i is arbitrary.

[1]See the discussion in Appendix B.3.1.

In order to simplify further derivations, we may want to consider the situation when the scaling factor L is non-singular, in which case both parts of (C.84) can be post-multiplied by L^{-1}. We will show that if r $[r \leq m]$ is the number of eigenvectors retained in W, then (C.84) reduces from the $\mathcal{R}^{d\times m}$ to the $\mathcal{R}^{d\times r}$-dimensional space with the non-singular scaling factor L_1.

To confirm it, assume without any loss of generality that

$$L = diag(l_1, l_2, \dots, l_r, 0, \dots 0) \subset \mathcal{R}^{m\times m}.$$

This assumption is valid due to invariance of SVD under simultaneous permutations of the left and right singular vectors of a matrix. Then

$$\mathsf{L} = \begin{bmatrix} \mathsf{L}_{1\ r\times r} & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(m-r)} \end{bmatrix}, \tag{C.86}$$

where $\mathsf{L}_1 = diag(l_1, l_2, \dots, l_r)$, and $\forall i = 1, \dots, m.\ l_i \neq 0$.

If $\mathsf{U} = [\mathsf{U}_{1\ d\times r}\mathsf{U}_{2\ d\times(m-r)}]$ and $\mathsf{R} = [\mathsf{R}_{1\ m\times r}\mathsf{R}_{2\ m\times(m-r)}]$, then

$$\begin{aligned} \mathsf{W} &= \begin{bmatrix} \mathsf{U}_{1\ d\times r} & \mathsf{U}_{2\ d\times(m-r)} \end{bmatrix} \begin{bmatrix} \mathsf{L}_{1\ r\times r} & 0_{r\times(m-r)} \\ 0_{(m-r)\times r} & 0_{(m-r)\times(m-r)} \end{bmatrix} \begin{bmatrix} \mathsf{R}^T_{1\ m\times r} \\ \mathsf{R}^T_{2\ m\times(m-r)} \end{bmatrix} \\ &= \mathsf{U}_1\mathsf{L}_1\mathsf{R}_1^T \subset \mathcal{R}^{d\times m}, \end{aligned} \tag{C.87}$$

where $\mathsf{U}_1 \subset \mathcal{R}^{d\times r},\ \mathsf{L}_1 \subset \mathcal{R}^{r\times r},\ r = rank(\mathsf{L}) \leq m.$

Since L_1 is a non-singular, both parts of (C.85) can be post-multiplied by L_1^{-1} and rewritten as

$$\mathsf{S}\mathsf{U}_1 = \mathsf{C}_\Sigma \mathsf{U}_1. \tag{C.88}$$

To investigate a relationship between the weight matrix W, the sample covariance S and covariance C_Σ, we will substitute expression (3.7) for the covariance C_Σ into equation (C.88):

$$\begin{aligned}
\mathsf{C}_\Sigma \mathsf{U}_1 &= \beta^{-1} m^{-1} \mathsf{U}_1 [\mathsf{I}_r + \mathsf{L}_1 \mathsf{R}_1^T (m\Gamma^{-1} - \mathsf{R}_1 \mathsf{L}_1^2 \mathsf{R}_1^T)^{-1} \mathsf{R}_1 \mathsf{L}_1] \\
&= \beta^{-1} m^{-1} \mathsf{U}_1 [\mathsf{I}_r + [m \mathsf{L}_1^{-1} \mathsf{R}_1^T \Gamma^{-1} \mathsf{R}_1 \mathsf{L}_1^{-1} - \mathsf{I}_r]^{-1}] \\
&= \mathsf{U}_1 [\beta m \mathsf{I}_r - \beta \mathsf{L}_1 \mathsf{R}_1^T \Gamma \mathsf{R}_1 \mathsf{L}_1]^{-1} \qquad (C.89) \\
&= \mathsf{U}_1 [\mathsf{U}_1^T \mathsf{C}_\Sigma^{-1} \mathsf{U}_1]^{-1} \qquad (C.90) \\
&= \mathsf{U}_1 \tilde{\mathsf{C}}_\Sigma \subset \mathcal{R}^{r\times r},
\end{aligned}$$

where $\tilde{\mathsf{C}}_\Sigma^{-1}$ is obtained from C_Σ^{-1} by discarding $d-r$ of its dimensions:

$$\tilde{\mathsf{C}}_\Sigma^{-1} = \beta m \mathsf{I}_r - \beta \mathsf{L}_1 \mathsf{R}_1^{\mathrm{T}} \Gamma \mathsf{R}_1 \mathsf{L}_1 \subset \mathcal{R}^{r\times r}. \qquad (C.91)$$

[We can also notice that in general (C.90) holds only for full ranked matrices $\mathsf{U}_1 \subset \mathcal{R}^{d\times d}$].

To simplify the further discussion, let P denote the right-most term in the brackets of (C.89):

$$\begin{aligned}
\mathsf{P} &= \mathsf{L}_1 \mathsf{R}_1^{\mathrm{T}} \Gamma \mathsf{R}_1 \mathsf{L}_1 \subset \mathcal{R}^{r\times r}, \qquad (C.92) \\
\mathsf{P}_{kl} &= l_k l_l \mathsf{r}_k^T \Gamma \mathsf{r}_l. \qquad (C.93)
\end{aligned}$$

Notice that columns of $\mathsf{R}_1 \subset \mathcal{R}^{m\times r}$ correspond to r columns of R and form an orthonormal basis. Therefore, R_1 may be thought of as a projection matrix.

From the identities above, (C.84) can be re-written for singular values $l_i \neq 0$:

$$\mathsf{S}\mathsf{U}_1 = \mathsf{U}_1 (\beta m \mathsf{I}_r - \mathsf{P})^{-1}. \qquad (C.94)$$

It may be found that solutions of (C.94) are:

$$\begin{cases} \mathsf{U}_1 = \mathsf{U}_S \Psi^T \subset \mathcal{R}^{d\times r}, \\ (\beta m \mathsf{I}_r - \mathsf{P})^{-1} = \Psi \Lambda_S \Psi^T \subset \mathcal{R}^{r\times r} \end{cases} . \qquad (C.95)$$

Here $\Psi \subset \mathcal{R}^{r\times r}$ is an arbitrary orthogonal matrix, $\mathsf{U}_S \subset \mathcal{R}^{d\times r}, \Lambda_S \subset \mathcal{R}^{r\times r}$ are r eigenvectors and corresponding eigenvalues of S.

From (C.83) and (C.92) system (C.95) can be re-written as

$$\begin{cases} \mathsf{W} = \mathsf{U}_S \Psi^T \mathsf{L}_1 \mathsf{R}_1^T \\ \beta m \mathsf{I}_r - \mathsf{L}_1 \mathsf{R}_1^T \Gamma \mathsf{R}_1 \mathsf{L}_1 = \Psi \Lambda_S \Psi^T. \end{cases} \tag{C.96}$$

Again, Ψ is an arbitrary rotation matrix and R_1 is a projection matrix.

C.4.4 Special Cases of Solutions for Parameter Estimates

From (C.94) it may be clearly seen that if P is not a diagonal matrix, then the $d \times r$ matrix U_1 does not have to be a matrix of eigenvectors of S. In this case, the analytic solution for the weight matrix W of the PoPPCA model appears to be difficult. However, if P is a diagonal matrix, then solutions for the expert weights W may be shown to correspond to rotated and scaled principal eigenvectors of S. Here we will comment on both of these cases.

P is a Diagonal Matrix

If $(\beta m \mathsf{I}_r - \mathsf{P})$ is a diagonal then

$$[\mathsf{P}]_{kl} \;\; = \;\; l_k l_l \mathsf{r}_k^T \Gamma \mathsf{r}_l \delta_{kl}, \tag{C.97}$$

It is possible to show that $(\beta m \mathsf{I}_r - \mathsf{P}_1)^{-1}$ is a diagonal if

$$\forall i, l, k = 1 \ldots m. \exists x \in \mathcal{R}. \sum_{f \neq i}^{m} [\mathsf{R}]_{fk} [\mathsf{R}]_{fl} (\gamma_f - \gamma_i) l_k l_l \equiv x \delta_{kl}, \tag{C.98}$$

where $\mathsf{R} \subset \mathcal{R}^{m \times m}$ is orthogonal. The identity holds when $\mathsf{R} = \mathsf{I}_m$, or $\Gamma = \alpha \mathsf{I}$, i.e. when all expert weights have identical weight magnitudes or are collinear with the eigenvectors of S.

- **Equal Experts' Norms:**

If $\Gamma = \alpha\mathsf{I}$, C_Σ and $\mathcal{L}$ are invariant under rotations of W, then from (3.6) we obtain

$$\begin{aligned}\mathsf{C}_\Sigma &= \beta^{-1}m^{-1}(\mathsf{I}_d + \mathsf{ULR}^T(\alpha m\mathsf{RR}^T - \mathsf{RL}^2\mathsf{R}^T)^{-1}\mathsf{RLU}^T)\\ &= \beta^{-1}m^{-1}(\mathsf{I}_d + \mathsf{UL}(\alpha m\mathsf{I}_m - \mathsf{L}^2)^{-1}\mathsf{LU}^T)\\ &= \beta^{-1}m^{-1}(\mathsf{I}_d + \tilde{\mathsf{W}}\tilde{\mathsf{W}}^T), \end{aligned} \tag{C.99}$$

where

$$\tilde{\mathsf{W}} = \mathsf{UL}(\alpha m\mathsf{I}_m - \mathsf{L}^2)^{-1/2}\mathsf{R}^T \subset \mathcal{R}^{d\times m}, \tag{C.100}$$

$\mathsf{R} \subset \mathcal{R}^{m\times m}$ is an orthogonal matrix.

Then, according to Tipping and Bishop (1997), $\tilde{\mathsf{W}}$ corresponds to scaled and rotated principal components of the sample covariance S. W also corresponds to rotated and scaled principal components of S, but with the different scaling factor. Now, in order to estimate W from $\tilde{\mathsf{W}}$, we may express the matrix L of singular values of W through the singular values of $\tilde{\mathsf{W}}$. However, it can be shown that expansion of (C.100) results in a higher-order equation for factor L.

- **Unit Rotation:**

Analogously, if $\mathsf{R} = \mathsf{I}_m$, then the covariance matrix can be expressed as

$$\mathsf{C}_\Sigma = \beta^{-1}m^{-1}(\mathsf{I}_d + \tilde{\mathsf{W}}\tilde{\mathsf{W}}^T), \tag{C.101}$$

where

$$\tilde{\mathsf{W}} = \mathsf{UL}(m\Gamma - \mathsf{L}^2)^{-1/2}\mathsf{V}^T, \tag{C.102}$$

$\mathsf{V} \subset \mathcal{R}^{m\times m}$ is an orthogonal matrix. Again, from Tipping and Bishop (1997) the solution for $\tilde{\mathsf{W}}$ still corresponds to rotated and scaled principal components of S.

P is Not a Diagonal Matrix

If $(\beta m \mathsf{I}_r - \mathsf{P}_1)$ is not a diagonal matrix, $\tilde{\mathsf{C}}_\Sigma$ is a block-diagonal matrix. In this case W corresponds to an arbitrary rotation (by Ψ), scaling, and another rotation (by R) of eigenvalues of the sample covariance matrix. As we have noted above, the left singular vectors of W do not need to correspond to eigenvectors of S.

The analytic solution for W in this case is complicated. We will limit our discussion to showing that the value of the log-likelihood $\mathcal{L}$ is not invariant under arbitrary orthogonal transformation of eigenvectors U_S. In order to show it, consider the following lemma:

Lemma 2: If $\mathsf{U}_m \subset \mathcal{R}^{d\times m}$, $\mathsf{B}_m \subset \mathcal{R}^{m\times m}$, $d > m$, *then* $|\mathsf{U}_{\mathrm{m}}\mathsf{B}_{\mathrm{m}}{\mathsf{U}_{\mathsf{m}}}^T| = 0$.

$\Box$ Let $\mathsf{U} = [\mathsf{U}_m \mathsf{0}_{d\times(d-m)}] \subset \mathcal{R}^{d\times d}$, and

$$\mathsf{B} = \begin{bmatrix} \mathsf{B}_{m\times m} & \mathsf{0}_{m\times(d-m)} \\ \mathsf{0}_{(d-m)\times m} & \mathsf{0}_{(d-m)\times(d-m)} \end{bmatrix} = \begin{bmatrix} \mathsf{B}_{m\times d} \\ \mathsf{0}_{(d-m)\times d} \end{bmatrix} \subset \mathcal{R}^{d\times d},$$

where subscripts indicate dimensions of the corresponding matrices. Then it is clear that

$$\mathsf{U}\mathsf{B} = \mathsf{U}_m\mathsf{B}_{m\times d} + \mathsf{0}_{d\times(d-m)}\mathsf{0}_{(d-m)\times d} = \begin{bmatrix} \mathsf{U}_{d\times m}\mathsf{B}_{m\times m} & \mathsf{0}_{d\times(d-m)} \end{bmatrix}.$$

Therefore,

$$\mathsf{U}_m\mathsf{B}_m\mathsf{U}_m^T = \mathsf{U}\mathsf{B}\mathsf{U}^T \tag{C.103}$$

$$|\mathsf{U}_m\mathsf{B}_m\mathsf{U}_m^T| = |\mathsf{U}\mathsf{B}\mathsf{U}^T| = 0. \tag{C.104}$$

■

To show that in general $\mathcal{L}$ is not invariant to rotation of W, consider the difference of $\mathcal{L}(\tilde{\mathsf{U}})$ and $\mathcal{L}(\tilde{\mathsf{U}}\Psi^T)$, where Ψ is an arbitrary orthogonal matrix, $\tilde{\mathsf{U}}$ - a eigenvector-solution for (C.84). Computing the difference in values of the log-likelihood, we get

$$\mathcal{L}(\tilde{\mathsf{U}}\Psi^T) - \mathcal{L}(\tilde{\mathsf{U}}) = \mathrm{tr}(\tilde{\mathsf{U}}(\mathsf{P} - \Psi^T\mathsf{P}\Psi)\tilde{\mathsf{U}}\mathsf{S}) + \ln\frac{|\beta m\mathsf{I} - \tilde{\mathsf{U}}\mathsf{P}\tilde{\mathsf{U}}^T|}{|\beta m\mathsf{I} - \tilde{\mathsf{U}}\Psi^T\mathsf{P}\Psi\tilde{\mathsf{U}}^T|}, \tag{C.105}$$

where $\mathsf{P} = \mathsf{L}\mathsf{R}^T\Gamma\mathsf{R}\mathsf{L} \subset \mathcal{R}^{m\times m}$ is a symmetric matrix. From (C.103), the rightmost term of the r.h.s. of the above expression is 0. However in general

$$\mathrm{tr}(\tilde{\mathsf{U}}(\mathsf{P} - \Psi^T\mathsf{P}\Psi)\tilde{\mathsf{U}}\mathsf{S}) \neq 0. \tag{C.106}$$

Unlike the PCA case when any solution leads to the same value of the optimality measure, and where it is sufficient to consider the simplest solution of (C.84) [e.g. when $\Psi = \mathsf{I}_m$], it is not guaranteed that the simplest solution in the PoPPCA model will lead to the highest value of the likelihood.

Appendix D

Analysis of CD Learning: Auxiliary Derivations

The appendix offers some of the auxiliary derivations supporting the result (4.33) for the second moment of the i^{th} reconstruction of the visible variable $\langle x_i^2 \rangle$ [see section 4.4.3].

$\square$ To express $\langle x_i^2 \rangle$ through the initial data point x_0 we make use of the recurrent relation (4.30):

$$\begin{cases} z_i = \sigma x_i + \tau u_i, \; i \geq 0 \\ x_j = \sigma z_{j-1} + \tau v_j, \; j \geq 1. \end{cases} \tag{D.1}$$

From (D.1) we can easily see that

$$\begin{aligned} z_0 &= x_0\sigma + \tau u_0 && \text{(D.2)} \\ x_1 &= x_0\sigma^2 + \tau u_0\sigma + \tau v_1 && \text{(D.3)} \\ z_1 &= x_0\sigma^3 + \tau u_0\sigma^2 + \tau v_1\sigma + \tau u_1 && \text{(D.4)} \\ x_2 &= x_0\sigma^4 + \tau u_0\sigma^3 + \tau v_1\sigma^2 + \tau u_1\sigma\tau v_2 && \text{(D.5)} \\ &\ldots \end{aligned}$$

$$
\begin{aligned}
z_2 &= x_0\sigma^5 + \tau u_0\sigma^4 + \tau v_1\sigma^3 + \tau u_1\sigma^2\tau v_2\sigma + \tau u_2 && \text{(D.6)}\\
x_3 &= x_0\sigma^6 + \tau u_0\sigma^5 + \tau v_1\sigma^4 + \tau u_1\sigma^3\tau v_2\sigma^2 + \tau u_2\sigma + \tau v_3 && \text{(D.7)}
\end{aligned}
$$

Continuing in the similar manner we can notice that

$$
\begin{aligned}
x_i &= x_0\sigma^{2i} + (\tau u_0)\sigma^{2i-1} + (\tau v_1)\sigma^{2i-2} + \ldots + (\tau u_{i-1})\sigma + \tau v_i\sigma^0\\
&= x_0\sigma^{2i} + \tau\sum_{k=0}^{i-1} u_k\sigma^{2i-1-2k} + \tau\sum_{k=1}^{i} v_k\sigma^{2i-2k}\\
&= x_0\sigma^{2i} + \tau\sum_{k=1}^{i}\sigma^{2i-2k}(u_{k-1}\sigma + v_k). && \text{(D.8)}
\end{aligned}
$$

Then by taking a square of (D.8) we see that

$$
\begin{aligned}
x_i^2 &= [x_0\sigma^{2i} + \tau\sum_{k=1}^{i}\sigma^{2i-2k}(\sigma u_{k-1} + v_k)]^2\\
&= \sigma^{4i}[x_0^2 + 2x_0\tau\sum_{k=1}^{i} C_k\sigma^{-2k} + \tau^2(\sum_{k=1}^{i}\sigma^{-2k}C_k)^2], && \text{(D.9)}
\end{aligned}
$$

where $C_k = \sigma u_{k-1} + v_k$.

Since $u_k,\ v_k \sim N(0,1)$ we obtain $\langle C_k\rangle = 0$ and

$$
\langle x_i^2\rangle = \sigma^{4i}\left[\langle x_0^2\rangle + \tau^2\langle(\sum_{k=1}^{i}\sigma^{-2k}C_k)^2\rangle\right]. \qquad \text{(D.10)}
$$

Considering that $\langle u_i v_k\rangle = 0,\ \langle u_i u_k\rangle = \langle v_i v_k\rangle = \delta_{ik}$, where δ_{ik} is the Kronecker delta, and performing some algebraic manipulations with (D.10), we can derive that

$$
\langle x_i^2\rangle = \sigma^{4i}\left[x_0^2 + (\sigma^2+1)\tau^2\sum_{k=1}^{i}\sigma^{-4k}\right]. \qquad \text{(D.11)}
$$

Notice that the sum of the geometric series in the term in brackets of the r.h.s. of (D.11) can be written as

$$
\sum_{k=1}^{i}\sigma^{-4k} = \frac{\sigma^{-4}(1-\sigma^{-4i})}{1-\sigma^{-4}} = \frac{\sigma^{4i}-1}{\sigma^{4i}(\sigma^4-1)}. \qquad \text{(D.12)}
$$

Substituting this expression into (D.11) we obtain

$$\begin{aligned}\langle x_i^2 \rangle &= \sigma^{4i} x_0^2 + \frac{\sigma^{4i}(\sigma^2+1)\tau^2(\sigma^{4i}-1)}{\sigma^{4i}(\sigma^4-1)} \\ &= \sigma^{4i} x_0^2 + \frac{\tau^2(\sigma^{4i}-1)}{\sigma^2-1} \\ &= \sigma^{4i}(x_0^2 - a) + a. \qquad \text{(D.13)}\end{aligned}$$

■