2022-11-04

Probabilistic Nonlinear Dimensionality Reduction

Adams, Matthew

UNIVERSITY OF CALGARY

Probabilistic Nonlinear Dimensionality Reduction

by

Matthew Adams

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

NOVEMBER, 2022

UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled "Probabilistic Nonlinear Dimensionality Reduction" submitted by Matthew Adams in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

Dr. Cristian Rios
Supervisor
Department of Mathematics and Statistics

Dr. Matthew Greenberg
Supervisory Committee Member
Department of Computer Science

Dr. Tony Ware
Supervisory Committee Member
Department of Mathematics and Statistics

Date

Abstract

High-dimensional datasets are present across scientific disciplines. In the analysis of such datasets, dimensionality reduction methods which provide clear interpretations of their model parameters are required. Principal components analysis (PCA) has long been a preferred method for linear dimensionality reduction, but is not recommended for data lying on or near low-dimensional nonlinear manifolds. On the other hand, neural networks have been used for dimension reduction but the associated model parameters have no clear interpretation. The main contribution of the current work is the introduction of probabilistic piecewise PCA, an interpretable model for approximating nonlinear manifolds embedded in high-dimensional space. Probabilistic piecewise PCA serves as a bridge between linear PCA and highly nonlinear neural network approaches to dimensionality reduction. Our model is an extension of probabilistic PCA and may be used when assuming any member of the natural exponential family of distributions on the observations. The model is explicitly defined for Gaussian and Poisson distributions, and posterior distributions for prediction and sampling are computed. A full comparative study of probabilistic piecewise PCA and existing dimensionality reduction methods is presented with a real-world bibliometric dataset.

Acknowledgements

This work has greatly benefited from formal and informal conversations with many talented and supportive people. Firstly, I must thank my PhD supervisor Cristian Rios for his continued interest, experienced mentorship, and mathematical support of my work. Our conversations have been both enlightening and challenging as he has encouraged me to chase down every detail.

I extend my heartfelt thanks to my supervisory committee members, Tony Ware and Matthew Greenberg. My early conversations with Tony were greatly encouraging and grounding in selecting the PhD path, and my meetings with Matthew were immensely helpful in tying the work to real-world thinking and mathematical rigor.

I am also extremely grateful for the support of the Alberta Children's Hospital Research Institute (ACHRI). In particular, Benedikt Hallgrimsson has been both a collaborator and a mentor as we have explored the fascinating world of bibliometrics together. I also thank Christina Hirota and Susa Benseler for the opportunity to work with ACHRI and their insightful input and questions into the bibliometric study portion of this work. I also extend my gratitude for the data support and patience I received from Kristy James and Kyle Demes at Elsevier.

I could not have completed my PhD work without the financial assistance of Alberta Innovates. I am incredibly grateful that such funding agencies exist and that they saw fit to take on my project.

I would also be grossly remiss if I failed to acknowledge the enduring support of my loving wife, Robyn, and my three sons, Wesley, Everett, and Quinton. Thank you for enthusiastically cheering me on through this adventure.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations, and Nomenclature

Symbol	Definition
$\mathbb{R}^n_+$	$\{\boldsymbol{x} \in \mathbb{R}^n : x_i > 0, 1 \leq i \leq n\}$
δ_{jk}	1 if $j = k, 0$ otherwise (Dirac delta)
$E_p(x)$	Expected value of the variable x with respect to the distribution p
$\odot$	Element-wise product of vectors or matrices (Hadamard product)
$\otimes$	Product of measures
$\|\|\cdot\|\|_2$	Vector 2-norm
$\|\|\cdot\|\|_{\mathcal{F}}$	Frobenius matrix norm
$\langle\cdot,\cdot\rangle$	Inner product
$\text{erf}(x)$	Gaussian error function with real argument x
$\text{erfc}(x)$	Gaussian complementary error function with real argument x
$Pr(\cdot)$	Probability of observing an event
iid	Independently and identically distributed
WLOG	Without Loss Of Generality

Introduction

High dimensional data are frequently encountered across scientific disciplines. In many data analysis settings, the goal is to extract information about the structure underlying the data. When the data lie on or near lower-dimensional manifolds, dimensionality reduction schemes may be used to map the data to a dimension of more manageable dimensionality for further analysis or visualization. In such cases, we say that there are 'latent' variables in a lower-dimensional space which can adequately explain the observations. A wide suite of techniques to achieve dimensionality reduction has emerged since the introduction of principal components analysis (PCA), a method which uses a single linear mapping to reduce dimensions. The linear mapping in PCA is defined as the projection onto the first few eigenvectors corresponding to the largest eigenvalues of the covariance matrix of the observed data. PCA has been extended by introducing probability distributions on the data and on its lower-dimensional representation. Within this probabilistic framework, the PCA model becomes generative: new points may be generated in either the original data space or the latent space by sampling from the appropriate distribution. Reciprocally, probabilistic models provide a way of determining how well new data fit into the initial dataset. Since the latent variables are not observed, PCA and its variants belong to the class of 'unsupervised' machine learning models.

In this work, we consider the problem of extending the probabilistic PCA model to nonlinear datasets in a piecewise manner. Until now, data lying on nonlinear manifolds embedded in high-dimensional space have required the use of either kernel methods or autoencoders for dimensionality reduction. In the case of kernel methods, expert knowledge is required to design the particular kernel which then defines the kernel covariance matrix to be diagonalized. Furthermore, kernel methods require the projection of the data into a possibly infinite dimensional Hilbert space before a finite subset is selected for the lower-dimensional representation. In the case of autoencoders, which are neural networks with a dimension-reducing architecture, the dimension

reduction model is highly overparametrized to provide maximum flexibility. Despite their flexibility, autoencoders are prone to overfitting - the resulting model does not always generalize well to new datasets. In addition, the parameters in the autoencoder model have no interpretation in terms of their contribution to the model. Two autoencoder models may be fitted to the same dataset with the same architecture and result in similar latent space representations, but the two models may have completely different parameters.

Our method for probabilistic piecewise PCA is guided by the linearity of classical PCA with the addition of probability assumptions on the data. In a high-dimensional observation space, we assume the data have a unimodal probability distribution across a lower-dimensional nonlinear manifold. We further assume that the data manifold may be well approximated by a collection of hyperplanes. In the simplest case, imagine a hinge embedded in three dimensions and suppose that we are given observations distributed across the surface of the hinge. The goal in this example is to 'flatten' the hinge while respecting the distribution across its surface. Classical PCA will lead to distortions in the latent space since it can only fit a single hyperplane to the data. Kernel PCA built on Gaussian kernels will similarly lead to distortions in latent space, as the Euclidean distance used in the Gaussian kernel would incorrectly group points which are close to the apex of the hinge but on separate sides. Autoencoders could fit the data perfectly, but the resulting latent representation and the model parameters have no real physical interpretation. Our method provides a solution to this problem by first finding a series of hyperplanes which can approximate the data. We then use the basis vectors for these hyperplanes to find optimal affine mappings to the latent space while respecting the relative ordering of the hyperplanes. A full probability model is applied to our method to map the unimodal distribution of the observations to the latent space. We have computed the probability models for both multivariate Gaussian and Poisson distributions, and lay the groundwork for further extensions to the general natural exponential family of distributions. The main benefits of our approach are that the model parameters are fully interpretable and that the model itself is generative.

Overview

The theoretical background prerequisites for this work are found in Chapter 0, wherein we collect basic definitions and general results from probability and measure theory, statistical learning, and optimization theory.

Bayesian inference is used extensively throughout this work in fitting parameters for various probabilistic models. Therefore, a brief introduction to the principles behind Bayesian inference, the general Expectation-Maximization (EM) algorithm, and variational inference are presented in Chapter 1.

We then proceed to Chapter 2, which provides a detailed look at the current methods available for dimensionality reduction. Two equivalent formulations of PCA are presented from the literature and related to the later probabilistic PCA. In addition, probabilistic PCA with the added assumption of Poisson distributions on the observations is presented and compared to PCA. Kernel PCA is also introduced and contrasted with standard affine PCA methods. Finally, this chapter introduces the structure of feedforward neural networks and the architecture of dimension-reducing autoencoders. Variational autoencoders are presented as the probabilistic extension of autoencoders and are related to probabilistic PCA.

Probabilistic piecewise PCA is defined in Chapter 3. After establishing the model assumptions and probabilistic framework, we discuss the options for maximizing the associated likelihood function. A variational approach is taken for model inference, where the log-likelihood function of the model is approximated through its variational lower bound. Positions in the latent space are estimated by their variational counterparts and resulting latent spaces are shown for synthetic datasets. Similarly to classical PCA, we calculate the proportion of the sample variance explainable by each of the latent dimensions produced by the piecewise PCA model. Though the main focus of Chapter 3 is on model inference, posterior distributions are provided for sampling and prediction. Finally, a criterion is established for estimating the unknown dimensionality of the latent space based on the log-likelihood of the piecewise PCA model.

Chapter 4 presents a comparison of dimension reduction methods applied to bibliometric data. Such datasets are records of the citation and publication trends of scholars across scientific disciplines. In our study, we consider the problem of summarizing several bibliometric variables into a two-dimensional scholarly profile for scholars affiliated with children's hospital research institutes across Canada. This chapter features the novel application of our probabilistic piecewise PCA model to a real dataset. For comparison, linear PCA models, kernel PCA, and autoencoders are also fitted to the bibliometrics data. When possible, the influence of each bibliometric variable on the resulting latent dimensions and the proportion of sample variance accounted for by the model are computed for each of these models. For the generative models, predictions are also generated from latent points sampled from the simulated posterior distributions. The resulting latent spaces and predictions are graphically compared between all models. From the results in this chapter, piecewise PCA is shown to outperform all other models both in terms of the proportion of explained sample variance and in the accuracy of its predictions.

Data Sources

The dataset used in the bibliometric study in Chapter 4 was downloaded using the Scopus API between December 14, 2021 and Janurary 11, 2022 via http://api.elsevier.com.ezproxy.lib.ucalgary.ca and http://www-scopus-com.ezproxy.lib.ucalgary.ca.

The bibliometric study was completed in support of the University of Calgary's 2022 Canada First Research Excellence Fund (CFREF) application. All code and data files are available at

`https://github.com/uniformlyMatt/compare-dim-reduction`.

A Python implementation of probabilistic piecewise PCA can be found at

`https://github.com/uniformlyMatt/piecewise-pca`.

Novel Contributions

The unique contributions of this work are twofold. First, we establish probabilistic piecewise PCA. To our knowledge, this is the first model that can achieve nonlinear dimension reduction by using piecewise affine functions within a probabilistic framework. This method allows data analysts to use familiar dimension reduction ideas with unimodal data distributed across the surface of a manifold. Our method also comes with an estimated proportion of variance explained by the model and an estimation of the dimensionality of the underlying latent space. The probabilistic piecewise PCA model provides a bridge between the affine model of probabilistic PCA and the highly nonlinear architecture of autoencoders, and may provide future insights to the interpretation the latent spaces computed by autoencoders. Furthermore, the addition of affine pieces to our model only linearly increases the number of model parameters.

Second, we present the first application of dimension reduction methods to bibliometric count data. Previous attempts have been made at aggregating bibliometric variables into lower-dimensional structures, but our application is the first that takes both discrete and continuous variables into account in the model. We also apply our novel piecewise PCA model to the real-world data and show its superior performance when compared to existing methods.

Chapter 0

Preliminaries

0.1 Probability Theory

Throughout this work we will be studying possible outcomes of randomly observed natural phenomena. The definitions in this section are from [7]. We will gather all possible observable outcomes in a nonempty set Ω and refer to this set as the *sample space*. A single point $\omega \in \Omega$ will be called a *sample point*. We will generally concern ourselves with computing the probability of observing subsets of Ω and we will gather sets of $\omega \in \Omega$ for which such probability statements are valid into the σ-algebra $\mathfrak{F}$, also known as the *event space*. Finally, we will equip $(\Omega, \mathfrak{F})$ with a measure P such that $P(\Omega) = 1$ and define the probability space $(\Omega, \mathfrak{F}, P)$.

Definition 0.1.1. *A real-valued random variable (RV) is a function $Y : \Omega \to \mathbb{R}$ such that $Y^{-1}(A) \in \mathfrak{F}$ for all $A \in \mathfrak{B}(\mathbb{R})$.*

A k-dimensional random vector X is then similarly defined as a function $X : \Omega \to \mathbb{R}^k$ such that $Y^{-1}(A) \in \mathfrak{F}$ for all $A \in \mathfrak{B}(\mathbb{R}^k)$.

We will take for granted the basic definitions of probability distribution, probability density functions (pdfs), and independence as presented in [7].

Definition 0.1.2. *(Conditional probability) Let Y, W be random variables with pdfs p_Y, p_W, and joint pdf $p_{Y,W}$. The conditional pdf of Y given $W = w$ is written as*

$$p_{Y|W=w}(y) = \frac{p_{Y,W}(y,w)}{p_W(w)}. \tag{0.1}$$

Theorem 1. *(Bayes' Theorem)*

Let Y, W be random variables with density functions p_Y and p_W. Then

$$p_{Y|W=w}(y) = \frac{p_{W|Y=y(w)}(w)p_Y(y)}{p_W(w)}. \tag{0.2}$$

Remark. Except in cases where it may cause ambiguity, we will refer to the pdf of Y as $p(y)$. We will also often interchange the terms 'probability density function' and 'distribution'.

Definition 0.1.3. *(Marginal pdf)*

Let Y, W be random variables with joint probability density function $p(y,w)$. The marginal pdfs of Y and W are defined as

$$p(y) = \int p(y,w)dw, \tag{0.3}$$

$$p(w) = \int p(y,w)dy. \tag{0.4}$$

Definition 0.1.4. *(Expected Value and Variance)*

For a random variable Y with pdf $p(y)$, the expected value or expectation of Y is defined as

$$E(Y) = \int yp(y)dy, \tag{0.5}$$

as long as the integral is well-defined.

The variance of Y is defined as

$$Var(Y) = E(Y - E(Y))^2 = \int (y - E(Y))^2 p(y)dy, \tag{0.6}$$

as long as $E(Y^2) < \infty$.

Definition 0.1.5. *(Convex Function)*

A function $f : (a,b) \to \mathbb{R}$ is called convex if for all $0 \leq \lambda \leq 1$ and $a < x \leq y < b$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y). \tag{0.7}$$

Theorem 2. *(Jensen's Inequality) Let Y be a random variable with $P(Y \in (a,b)) = 1$ for some interval $(a,b) \subset \mathbb{R}$, and let $f : (a,b) \to \mathbb{R}$ be a convex function. Then*

$$f\left(E(Y)\right) \leq E\left[f(Y)\right], \tag{0.8}$$

as long as $E|Y| < \infty$ and $E|f(Y)| < \infty$.

Definition 0.1.6. *(Likelihood function)*

Let $Y = \{Y_i\}_{i=1}^N$ be a set of N independent random variables, each with a one-parameter pdf $p(y_i|\theta_i)$. Collect all distribution parameters θ_i into the parameter vector θ. We define the likelihood function as the likelihood of observing the dataset Y given the parameter vector θ,

$$p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta_i). \tag{0.9}$$

Due to the properties of the logarithm, it is often more convenient to work with the log-likelihood given by

$$\log p(y|\theta) = \sum_{i=1}^{N} \log p(y_i|\theta_i). \tag{0.10}$$

The following theorem, proven in [9], will be useful when we discuss Expectation-Maximization (EM) algorithms for computing principal components.

Theorem 3. *Let Y, W be random variables with joint pdf $p(y,w|\theta)$. Let $q(w)$ be any distribution over W. Then the log-likelihood has the decomposition*

$$\log p(y|\theta) = \mathcal{L}(q,\theta) + KL(q||p), \tag{0.11}$$

where

$$\mathcal{L}(q,\theta) := \int q(w) \log\left(\frac{p(y,w|\theta)}{q(w)}\right) dw, \tag{0.12}$$

and

$$KL(q||p) := -\int q(w) \log\left(\frac{p(w|y,\theta)}{q(w)}\right) dw \tag{0.13}$$

Theorem 4. *Let Y be a continuous random variable with pdf $p(y)$. Suppose that $q(y)$ is another density function. In the decomposition* (0.11) *from Theorem 3, $KL(q||p) \geq 0$, with equality if and only if $p = q$.*

0.2 General Results

The following theorem is taken from [7], where its proof may be found.

Theorem 5. *(Fubini Theorem)*

Let $(\Omega_1, \mathfrak{F}_1, \mu_1)$ and $(\Omega_2, \mathfrak{F}_2, \mu_2)$ be σ-finite measure spaces and let f be an integrable function on the product space $(\Omega_1 \times \Omega_2, \mathfrak{F}_1 \times \mathfrak{F}_2, \mu_1 \times \mu_2)$. Then:

$$\begin{aligned}\int f(\omega_1,\omega_2) d(\mu_1 \times \mu_2) &= \int \left[\int f(\omega_1,\omega_2) d\mu_2(\omega_2)\right] d\mu_1(\omega_1) \\ &= \int \left[\int f(\omega_1,\omega_2) d\mu_1(\omega_1)\right] d\mu_2(\omega_2).\end{aligned} \tag{0.14}$$

Definition 0.2.1. *(Gradient Descent)*

Let f be a differentiable function on $\mathbb{R}^n$ with gradient ∇f, and let $\alpha > 0$. The gradient descent method for calculating the minima of f is defined by the iterative sequence

$$\begin{cases} x_0, \quad \textit{initial guess}, \\ x_{j+1} = x_j + \alpha \nabla f(x_j), \quad \forall j \geq 1. \end{cases} \tag{0.15}$$

0.3 Special Functions

Definition 0.3.1. *(Gamma Function)*

For $x \in \mathbb{R}$, the gamma function is denoted by $\Gamma(\cdot)$ and defined as:

$$\Gamma(x) = \int_0^\infty t^{x-1} exp\,(-t)\,dt. \tag{0.16}$$

Definition 0.3.2. *(Error Function and Complementary Error Function)*

For $x \in \mathbb{R}$, the error function is denoted by $erf(x)$ and defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x exp\left(-\frac{t^2}{2}\right) dt. \tag{0.17}$$

The complementary function is denoted by $erfc\,(x)$ and defined as:

$$erfc\,(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty exp\left(-\frac{t^2}{2}\right) dt. \tag{0.18}$$

0.4 Kernel Theory

Definition 0.4.1. *(Kernel Function)*

A kernel k is a function $k : \Omega \times \Omega \to \mathbb{R}$. For a finite set $\{x_i\}_{i=1}^m \subset \Omega$, we associate to the kernel its kernel matrix $K = [k(x_i, x_j)]_{ij} \in \mathbb{R}^{m \times m}$.

Definition 0.4.2. *(Positive Definite Matrix)*

Let $q \in \mathbb{R}$. A symmetric matrix $A \in \mathbb{R}^{q \times q}$ is called positive definite if, for any non-zero vector $x \in \mathbb{R}^q$,

$$x^T A x \geq 0. \tag{0.19}$$

Remark. In some texts, positive definiteness requires that the quadratic form $x^T A x$ be strictly greater than zero. In this work, it makes no substantial difference.

Definition 0.4.3. *(Positive Definite Symmetric Kernel)*

A kernel is positive definite symmetric (PDS) if, for any finite set $\{x_i\}_{i=1}^m \subset \Omega$, the associated kernel matrix K is symmetric positive definite.

Definition 0.4.4. *(Reproducing Kernel Hilbert Space)*

A Hilbert space $\mathbb{H}$ has the reproducing property if all evaluation functionals in $\mathbb{H}$ are bounded and linear.

The next theorem is proven in [68].

Theorem 6. *(Representer Theorem)*

Let X be a non-empty set. For every PSD kernel $k : X \times X \to \mathbb{R}$ there exists a reproducing kernel Hilbert space $\mathbb{H}$ and a mapping $\Phi : X \to \mathbb{H}$ such that

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle, \quad \text{for all } x,y \in X. \tag{0.20}$$

0.5 The Natural Exponential Family of Distributions

Since this project is partially motivated by the analysis of count data, we will introduce appropriate probability distributions for discrete count data. Throughout this work, we will concern ourselves mainly with distributions from the Natural Exponential Family (NEF). The following definitions and theorems are from [21].

Definition 0.5.1. *(One-parameter natural exponential family)*

Let $Y = \{Y_1, \ldots, Y_N\} \subset \mathbb{R}^p$ be random variables distributed according to p_λ, where $\lambda \in \mathcal{L} \subset \mathbb{R}$. If we can write the pdf p_λ in the canonical form

$$f(y|\lambda) = h(y) exp\left(\lambda T(y) - \psi(\lambda)\right), \tag{0.21}$$

with $T(y), \psi(\lambda)$ real-valued, $h(y) \geq 0$, and

$$\lambda \in \mathcal{L} := \left\{ \lambda : exp\left(\psi(\lambda)\right) = \int_{\mathbb{R}^d} h(y) exp\left(\lambda T(y)\right) dy < \infty \right\}, \tag{0.22}$$

then we say that the family of distributions $\{p_\lambda, \lambda \in \mathcal{L}\}$ belongs to the one-parameter natural exponential family (NEF).

Definition 0.5.2. *(Canonical Link Function)*

For one-parameter NEF distributions of the form (0.21)*, the canonical link function g is defined as the function such that*

$$g(\psi'(\lambda)) = \lambda. \tag{0.23}$$

Theorem 7. *Let* $\lambda \in \mathcal{L}$ *as in* (0.22)*. Then the maximum likelihood estimate of* $T(Y)$ *is given by* $\psi'(\lambda)$*. That is,*

$$E_\lambda(T(Y)) = \psi'(\lambda). \tag{0.24}$$

0.5.1 Gaussian Distributions in Canonical Form

Several frequently encountered distributions are included in the one-parameter NEF. For example, the pdf for a Gaussian distributed random variable with known variance σ^2 and single parameter μ is often written as

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{0.25}$$

In the literature, the terms Gaussian and 'normal' are used interchangeably. In this work, we will use the name Gaussian, and will denote the Gaussian distribution with mean μ and variance σ^2 by $\mathcal{N}(\mu,\sigma^2)$. In higher dimensions, the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ will be written as $\mathcal{N}(\boldsymbol{\mu},\Sigma)$.

We can rearrange the terms in (0.25) to see the Gaussian distribution in its canonical form.

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)\exp\left(\mu\frac{x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) = h(x)\exp\left(\mu T(x) - \psi(\mu)\right), \tag{0.26}$$

where we denote

$$\begin{aligned} h(x) &= (2\pi)^{-1/2}\exp\left(-x^2/2\sigma^2\right), \\ T(x) &= x/\sigma^2, \text{ and} \\ \psi(\mu) &= \frac{\mu^2}{2\sigma^2}. \end{aligned} \tag{0.27}$$

In this case, $\mathcal{L} \equiv \mathbb{R}$. By Definition 0.5.2, the canonical link function for the univariate Gaussian is given by

$$g(\psi'(t)) = t \implies g(t) = \sigma^2 t. \tag{0.28}$$

The multidimensional Gaussian distribution with known covariance matrix Σ can also be written as a one-parameter NEF. Let X be a p-dimensional Gaussian random variable and denote its mean vector as $\boldsymbol{\mu}$. To write this distribution in the NEF canonical form, we will consider the entire mean vector as a 'single' parameter. Then,

$$\begin{aligned} f(\boldsymbol{x}|\boldsymbol{\mu}) &= (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right) \\ &= (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right) \exp\left(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right) \\ &= h(\boldsymbol{x})\exp\left(T(\boldsymbol{x})\cdot\boldsymbol{\mu} - \psi(\boldsymbol{\mu})\right), \end{aligned} \tag{0.29}$$

where we have denoted

$$\begin{aligned} h(\boldsymbol{x}) &= (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right), \\ T(\boldsymbol{x}) &= \boldsymbol{x}^T\Sigma^{-1}, \text{ and} \\ \psi(\boldsymbol{\mu}) &= \frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}. \end{aligned} \tag{0.30}$$

The canonical link function for the multivariate Gaussian with known covariance matrix Σ is given by

$$g(\psi'(\boldsymbol{\tau})) = \boldsymbol{\tau} \implies g(\boldsymbol{\tau}) = \Sigma\boldsymbol{\tau}. \tag{0.31}$$

Remark. For a p-dimensional Gaussian distribution with isotropic covariance matrix $\sigma^2 I$, the NEF canonical form can be written as

$$h(\boldsymbol{x})\exp\left(T(\boldsymbol{x})\cdot\boldsymbol{\mu}-\psi(\boldsymbol{\mu})\right), \tag{0.32}$$

where

$$\begin{aligned} h(\boldsymbol{x}) &= (2\pi\sigma^2)^{-\frac{p}{2}}\exp\left(-\frac{1}{2\sigma^2}x^T x\right), \\ T(\boldsymbol{x}) &= \sigma^{-2}\boldsymbol{x}^T, \\ \psi(\boldsymbol{\mu}) &= \frac{1}{2\sigma^2}\boldsymbol{\mu}^T\boldsymbol{\mu}, \text{ and} \\ g(\boldsymbol{\mu}) &= \sigma^2\boldsymbol{\mu}. \end{aligned} \tag{0.33}$$

0.5.2 Poisson Distributions in Canonical Form

Similarly, the pdf for the Poisson distribution with single parameter λ is often written as

$$f(x|\lambda) = \frac{1}{x!}\lambda^x\exp(-\lambda). \tag{0.34}$$

This pdf can be rewritten in the form (0.21) with the canonical parameter $\eta = \log\lambda$.

$$f(x|\lambda) = \frac{1}{x!}\exp(x\log\lambda-\lambda) = h(x)\exp(\eta T(x)-\psi(\eta)) =: F(x|\eta), \tag{0.35}$$

where

$$\begin{aligned} h(x) &= \frac{1}{x!}, \\ T(x) &= x, \text{ and} \\ \psi(\eta) &= \exp(\eta). \end{aligned} \tag{0.36}$$

The canonical link function for the univariate Poisson distribution is given by:

$$g(\psi'(t)) = t \implies g(t) = \log t. \tag{0.37}$$

Remark. In this work, we will often write the Poisson distribution with canonical parameter η in the notational shorthand $\mathcal{P}(\exp(\eta))$.

Many versions of the multidimensional Poisson distribution exist [41]. For a finite set of RVs $\{Y_i\}_{i=1}^N$, the simplest, most convenient formulation of a multivariate Poisson distribution would be to consider each independent RV as a univariate Poisson distribution with canonical parameter λ_i and with no correlation structure between variables. This multivariate Poisson distribution with independent marginal Poisson distributions was proven to be the limit of multivariate binomial distributions in [45]. For the present work, we adopt this simplified multivariate Poisson distribution, noting that we cannot capture correlations between variables with such a model.

Let $Y \in \mathbb{R}^p$ be a random vector with independent marginal Poisson distributions. That is, for every component Y_j of Y,

$$p(y_j) = h(y_j)\exp\left(y_j\eta_j - \exp\left(\eta_j\right)\right), \tag{0.38}$$

where $\lambda_j = \exp\left(\eta_j\right)$ is the canonical parameter. Then, the canonical form of the density function for the random vector Y with independent components Y_j is given by

$$\begin{aligned}
p(y) &= \prod_{j=1}^{N} p(y_j) \\
&= \prod_{j=1}^{N} h(y_j)\exp\left(y_j\eta_j - \exp\left(\eta_j\right)\right) \\
&= \exp\left(\log\left[\prod_{j=1}^{N} h(y_j)\exp\left(y_j\eta_j - \exp\left(\eta_j\right)\right)\right]\right) \\
&= \exp\left(\sum_{j=1}^{p} \log h(y_j) + y_j\eta_j - \exp\left(\eta_j\right)\right) \\
&= \exp\left(\sum_{j=1}^{p} \log\left(\frac{1}{y_j!}\right)\right)\exp\left(\sum_{j=1}^{p} y_j\eta_j - \exp\left(\eta_j\right)\right) \\
&= \exp\left(-\sum_{j=1}^{p}\sum_{k=2}^{y_j} \log k\right)\exp\left(y^T\boldsymbol{\eta} - \sum_{j=1}^{p}\exp\left(\eta_j\right)\right) \\
&= H(y)\exp\left(y^T\boldsymbol{\eta} - \Psi(\boldsymbol{\eta})\right),
\end{aligned} \tag{0.39}$$

where

$$H(y) = \exp\left(-\sum_{j=1}^{p}\sum_{k=2}^{y_j}\log k\right)$$

$$\boldsymbol{\eta} = [\eta_1, \ldots, \eta_j, \ldots, \eta_p]^T, \text{ and}$$

$$\Psi(\boldsymbol{\eta}) = \sum_{j=1}^{p}\exp\left(\eta_j\right). \tag{0.40}$$

Now, since $\nabla\psi(\boldsymbol{\eta}) = [\exp(\eta_1), \ldots, \exp(\eta_p)]^T$, the canonical link function is given by:

$$g(\nabla\psi(\boldsymbol{\tau})) = \boldsymbol{\tau} \implies g(\boldsymbol{\tau}) = [\log\tau_1, \ldots, \log\tau_p]^T = \log\boldsymbol{\tau}. \tag{0.41}$$

Chapter 1

Bayesian Inference

1.1 Introduction

Given a probability model for an observed dataset, Bayesian Inference is a method used for the estimation of the model parameters. The method is called 'Bayesian' since it prominently uses Bayes' Theorem (Theorem 1). Problems in Bayesian Inference are generally posed with a complete probability model for all of the variables involved in the problem. We will encounter cases where the problem includes both observed and unobserved, or 'latent', variables. Throughout this work we will adopt the Bayesian framework for probabilistic approaches to dimensionality reduction problems. Parameters for the probabilistic models in this chapter are found by maximizing the associated likelihood functions. When the direct maximization of a given likelihood function is too difficult, the Expectation-Maximization (EM) algorithm is often used as an alternative. The EM algorithm, which was shown to converge to local maxima in [22][1], is introduced in § 1.3.

1.2 Latent Variable Models

We begin with two sample spaces $\mathcal{Y} \subset \mathbb{R}^p$ and $\mathcal{W} \subset \mathbb{R}^q$. At present, we will not impose any restrictions on the dimensions of the two respective spaces: p is not necessarily equal to q. We assume a functional relationship between the two spaces such that for $Y \in \mathcal{Y}$, $Y = Y(W)$. The difference between these two spaces is that random variables $Y \in \mathcal{Y}$ are observable, while variables $W \in \mathcal{W}$ are not. For this reason, W are called 'latent' variables. Our goal is to gain information from the observations that we can use to describe the latent space.

We can put this more intuitively as follows: we are presented with a set of observations Y, but

[1] Further conditions for the convergence of the EM algorithm, such as when the prior distribution comes from the NEF distributions, were given in [84].

we suspect that there is some unseen mechanism determining the distribution of Y. We give this unseen set of variables a name, W, and hypothesize that they are distributed according to some density function $p(W)$. Since $p(W)$ is assumed at the outset, we call it the *prior* distribution. Any density function on the observations needs to take the prior distribution on the latent variables into account. In probabilistic terms, the observations are conditioned on the latent variables. Therefore, we define the data distribution as $p(y|w,\theta)$, where θ are the unknown model parameters.

In order to find the model parameters, we need to take away the dependence of the observations on the latent variables. We can do this using Definition 0.1.2 and the prior distribution to find the marginal data distribution:

$$p(y|\theta) = \int p(y,w|\theta)dw = \int p(y|w,\theta)p(w)dw. \tag{1.1}$$

Now, we may optimize the likelihood of observing our dataset given the model parameters. That is, we want to solve

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\ p(y|\theta). \tag{1.2}$$

Most often we are given a finite set of iid observed variables $\{Y_i\}_{i=1}^N$. In this case, we associate to each observation a latent variable W_i. The independence of the observations means that the likelihood function factors, and is given by

$$p(y|\theta) = \prod_{i=1}^{N} p(y_i|\theta). \tag{1.3}$$

A simplification occurs by taking the logarithm to transform the product into a sum as in (0.10) in Definition 0.1.6. We then have the log-likelihood function

$$\log p(y|\theta) = \sum_{i=1}^{N} \log p(y_i|\theta), \tag{1.4}$$

where

$$p(y_i|\theta) = \int p(y_i,w_i|\theta)dw. \tag{1.5}$$

Whether we use (1.1) or (1.5) to solve the problem given by (1.2), we are faced with the integration[2] of the joint distribution inside the logarithm.

[2]When W are discrete random variables, the integral is replaced by a summation.

Before we present a systematic way around this difficult integration via the EM algorithm in § 1.3, we will briefly discuss the use of Bayesian methods for prediction. While it is not the focus of our current work, prediction is one of the main advantages to the Bayesian perspective. Our overall goal is to describe the latent space. So, we compute an expression for the distribution of the latent variables conditioned on the observations and the model parameters using Bayes' Theorem (Theorem 1):

$$p(w|y,\theta) = \frac{p(y,w|\theta)}{p(y|\theta)} = \frac{p(w)p(y|w,\theta)}{p(y|\theta)}. \tag{1.6}$$

This distribution is called the *posterior* distribution, since it is calculated after both the model parameters have been estimated and the marginal data distribution $p(y|\theta)$ has been computed.

Once we have computed the marginal data distribution, estimated the model parameters, and calculated the posterior distribution, our probabilistic model is now 'generative'. That is, new observations or latent points can be easily generated by sampling from the appropriate distribution. We can then check the fit of the model by sampling from the posterior distribution and comparing the result of $p(y|w,\theta)$ to the observations using various statistics.

1.3 The Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is a optimization technique for parametrized probabilistic models, and was formalized in [22]. We present the EM algorithm in its general form, and will show specific applications of the EM algorithm and its variational counterpart in Chapter 3.

Denote by Y the set of all observations and by W the set of all latent variables with associated prior distribution $q(w)$, where $q(w) > 0$ almost everywhere. Define the joint density function $p(y,w|\theta)$ with parameters θ. The log-likelihood of the dataset is given by

$$\log p(y|\theta) = \log\left(\int p(y,w|\theta)dw\right) \tag{1.7}$$

Our goal is to find values of θ that maximize this log-likelihood function. The EM algorithm is applied to instances where it is easier to maximize the joint distribution $\log p(y,w|\theta)$ rather than

the marginal distribution $\log p(y|\theta)$.

From Theorem 3 we have the decomposition of the log-likelihood

$$\begin{aligned}\log p(y|\theta) &= \mathcal{L}(q,\theta) + KL(q||p) \\ &= E_{q(w)} \log\left(\frac{p(y,w|\theta)}{q(w)}\right) - E_{q(w)} \log\left(\frac{p(w|y,\theta)}{q(w)}\right),\end{aligned} \tag{1.8}$$

where we have used the notation E_q to indicate taking the expectation with respect to the distribution q. When using the EM algorithm, we only deal directly with the term $\mathcal{L}(q,\theta)$. Since $KL(q||p) \geq 0$ by Theorem 4, $\mathcal{L}(q,\theta)$ serves as a lower bound for the log-likelihood $\log p(y|\theta)$. That is,

$$\log p(y|\theta) \geq \mathcal{L}(q,\theta). \tag{1.9}$$

The EM algorithm is composed of two alternating iterative steps. In the expectation step (E-step), we fix the model parameters θ_j and maximize the lower bound $\mathcal{L}(q,\theta_j)$ with respect to the distribution q. Given that $KL(q||p) = 0$ if and only if $q = p$, it is clear that the lower bound is maximized when $q = p$. In the context of the decomposition (1.8), this amounts to setting $q(w) = p(w|y,\theta_j)$. The convenience in setting the posterior distribution to $q(w)$ is fully realized by Theorem 3, which states that the decomposition (1.8) holds for any choice of prior distribution q. We will take full advantage of this fact in § 1.3.1. At present, we will set $p(w|y,\theta) = q(w)$ so that $\log p(y|\theta_j) = \mathcal{L}(q,\theta_j)$, and the E-step is completed by simply evaluating the log-likelihood.

To recap, the E-step involves fixing the model parameters θ_j and evaluating the log-likelihood $\log p(y|\theta)$. In the subsequent maximization step (M-step), we update the model parameters by maximizing the lower bound $\mathcal{L}(q,\theta_j)$ with respect to θ_j while keeping the distribution q fixed. In the case of the NEF distributions, the logarithm in $\mathcal{L}(q,\theta_j)$ cancels the exponential function, which leads to a much simpler optimization problem than that of maximizing $\log p(y,w|\theta_j)$ [9].

Iterations of alternating E- and M- steps continue until a desired convergence criterion is reached, either in terms of the log-likelihood or the parameter values themselves. In this work, the EM algorithm continues until the absolute difference between the updated log-likelihood and

its previous value drops below a prescribed tolerance $\varepsilon > 0$. A complete view of the general EM algorithm is shown in Algorithm 1.

Data:
  Observations Y;
  Model parameters θ^{old};
  Latent prior distribution $q(w)$;
  Convergence tolerance $\varepsilon > 0$;
Result: θ^{opt};
Set $|\mathcal{L}(q, \theta^{\text{old}}) - \mathcal{L}(q, \theta^{\text{new}})| = 1$;
while $|\mathcal{L}(q, \theta^{old}) - \mathcal{L}(q, \theta^{new})| > \varepsilon$ **do**
  $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$;
  E-step:
    Evaluate $\mathcal{L}(q, \theta^{\text{old}})$;
  M-step:
    $\theta^{\text{new}} \leftarrow \text{argmax}_{\ \theta^{\text{old}}} \mathcal{L}(q, \theta^{\text{old}})$;
end
$\theta^{\text{opt}} \leftarrow \theta^{\text{new}}$;

Algorithm 1: The general EM algorithm.

1.3.1 Variational Approximation

The decomposition of the log-likelihood function using the marginal data distribution comes with a particularly useful advantage. Recall that the decomposition (1.8) holds for any choice of prior distribution q. Now, in the E-step of the EM algorithm, we usually set $q(w) = p(w|y, \theta)$ so that $KL(q||p) = 0$. This step may be seen as approximating the posterior distribution with the prior distribution, which makes intuitive sense when we view $KL(q||p)$ as a measure of the discrepancy between the distributions q and p. Since the decomposition holds for any other choice of prior, we introduce the approximate distribution $\tilde{p}$. The advantage of this choice becomes clear when we choose $\tilde{p}$ to be a distribution that easily factors.

Let $\tilde{p} = \otimes_{i=1}^{N} \tilde{p}_i$ be a product distribution with marginal probability distributions $\tilde{p}_i$. When this

product distribution $\tilde{p}$ is applied to the lower bound of the log-likelihood from (1.8), we obtain:

$$\begin{aligned}
\mathcal{L}(\tilde{p},\theta) &= \int \tilde{p}(w)\log\left(\frac{p(y,w|\theta)}{\tilde{p}(w)}\right)dw \\
&= \int \prod_{i=1}^{N}\tilde{p}_i(w_i)\log\left(\frac{\prod_{i=1}^{N}p(y_i,w_i|\theta)}{\prod_{i=1}^{N}\tilde{p}_i(w_i)}\right)dw_i \\
&= \int \prod_{i}\tilde{p}_i(w_i)\left(\sum_i \log p(y_i,w_i|\theta) - \sum_i \log\tilde{p}_i(w_i)\right)dw_i \\
&= \int \prod_{i}\tilde{p}_i(w_i)\left(\sum_i \log p(y_i|w_i,\theta) + \log p(w_i) - \log\tilde{p}_i(w_i)\right)dw_i
\end{aligned} \tag{1.10}$$

Now, using Lemma 2 with the measurable function $f := \sum_i \log p(y_i|w_i,\theta) + \log p(w_i) - \log\tilde{p}_i(w_i)$, we get an approximation to the lower bound in the general form:

$$\begin{aligned}
\mathcal{L}(\tilde{p},\theta) &= \int \prod_{i}\tilde{p}_i(w_i)\left(\sum_i \log p(y_i|w_i,\theta) + \log p(w_i) - \log\tilde{p}_i(w_i)\right)dw_i \\
&= \sum_{i=1}^{N}\int \tilde{p}_i(w_i)\left[\log p(y_i|w_i,\theta) + \log p(w_i) - \log\tilde{p}_i(w_i)\right]dw_i.
\end{aligned} \tag{1.11}$$

Thus, by redefining the EM algorithm in terms of the easily factored product distribution $\tilde{p}$, we end up with simpler terms to optimize. Since $\tilde{p}$ is an approximation to the distribution inside an integral, we say that it is a 'variational' approximation. Such variational approximations allow us to maximize far simpler lower bounds for the log-likelihood, particularly when $\tilde{p}_i$ are chosen as members of the NEF distributions. The associated variational lower bound (VLB) is used in many modern probabilistic models, such as Poisson log-Normal principal components analysis (PLNPCA) (see § 2.5.1) and in the piecewise probabilistic principal components analysis model defined in Chapter 3.

Chapter 2

Dimension Reduction

2.1 Introduction

Data analysis is a wide-ranging collection of techniques for summarizing data and exploring the relationships between random variables. In this work, we focus on the task of dimension reduction. Datasets may contain either discrete or continuous variables and be either finite or infinite, but in this work we will adopt the convention that datasets are finite sets composed of real-valued random vectors living in some linear space $\mathbb{R}^p$. There is not a standard definition of 'high' dimensionality, so we will adopt the convention that data has relatively high dimensionality when the number of observed random variables exceeds four. High dimensional data is frequently encountered in many modern applications, such as bibliometrics, ecology, and image processing.

Throughout this work, we will associate various multivariate probability distributions to the data that we encounter. Probabilistic dimension reduction techniques are used to map high dimensional datasets to lower dimensional spaces where probability distributions are assigned to both spaces. In this section, we first present a problem that occurs when analysts attempt to describe the spread of high dimensional data around the center of the distribution. We then introduce several dimension reduction techniques, beginning with linear methods and progressing to nonlinear methods such as autoencoders. Not all of the covered dimension reduction methods come with a probabilistic perspective, which leads to disadvantages for some methods and advantages for others. In the interest of brevity, a full study of the comparative performance of dimension reduction methods in real-world applications is postponed until Chapter 4. A summary table showing the relative strengths and deficiencies of all of the considered techniques is presented in Table 2.2.

When discussing the various models presented in this chapter, we will often refer to the interpretability of the model parameters. Throughout this work, we will adopt the following definition

of an 'interpretable' model:

> A dimension reduction model is interpretable if the effect of each model parameter on the input can be easily followed through the model's computations, and the corresponding model output may be computed *a priori*.

With this definition in mind, it is immediately clear that linear and affine models are interpretable. On the other hand, post-hoc explanations of model parameters do not constitute an interpretable model within this context.

2.2 The Curse of High-dimensional Data

High dimensional datasets are attractive since they can richly describe a set of observations using several variables. However, data residing in higher dimensional spaces displays some non-intuitive behavior. As the number of dimensions increases, the amount of data needed to describe the joint probability distribution of the high dimensional variables increases exponentially. In this section, we will show an example of how high dimensionality leads to sparsely distributed points where we might expect the opposite. The example in this section is adapted from [21].

Suppose that $Y \in \mathbb{R}^p$ is a random vector such that Y has independent uniform marginal distributions in each of its p dimensions. That is, $p(y_j) = \frac{1}{2^p}$, $1 \leq j \leq p$. To illustrate the behavior of this distribution in high dimensions, we will compute the probability of observing a point inside the p-dimensional unit sphere given by $|y|^2 \leq 1$ for various values of p.

The probability of a single point appearing inside the unit sphere is given by

$$\int_{|y|^2 \leq 1} \frac{1}{2^p} dy = \frac{1}{2^p} \left(\frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} \right), \tag{2.1}$$

where $\Gamma(\cdot)$ is the gamma function (see Definition 0.3.1). As in [21], we present probabilities for various values of p in Table 2.1, where we also record the number of data points needed to observe at least one point inside the unit p-sphere.

Table 2.1: Probability of observing a random vector with uniform marginal distributions in the p-dimensional sphere for exponentially increasing values of p. The column N is the number of data points required to observe at least one point in the p-dimensional unit sphere.

p	$P(\|y\|^2 \leq 1)$	N
2	0.785	1
4	0.308	3
8	0.016	63
16	3.59×10^{-6}	278,484
32	1.00×10^{-15}	9.98×10^{14}
64	1.67×10^{-39}	5.99×10^{38}

This inconvenient phenomenon has been termed the 'Curse of Dimensionality'. In order to describe the central tendency of jointly distributed variables in high-dimensional space, we need huge amounts of data as the number of dimensions increases. However, probabilistic dimension reduction methods help to lessen this burden by imposing approximate probability distributions on the observations.

2.3 Literature Review

Early attempts at reducing the dimensionality of complex datasets emerged out of the need to analyze educational and psychological tests [39, 23, 77]. Principal Components Analysis (PCA) was first proposed by Hotelling in [39] as a method for reducing the number of dimensions in a given dataset using a system of linear equations. The main idea driving this method was to find a 'component', or basis vector, that would explain as much of the sample variance as possible. Successive components which explained as much of the remaining variance as possible were then iteratively found. Eckart and Young [23] viewed the problem of dimension reduction from the perspective of optimally approximating a matrix by another matrix of lesser rank. This work is notable in that it used the singular value decomposition of the matrix to be approximated, a method which would later become the standard. Tipping and Bishop [75] introduced probabilistic

PCA (pPCA) by assigning probability distributions to the dataset and to the lower-dimensional points. This method treated the lower-dimensional mapped data points as latent variables [9, 27], upon which the dataset was conditioned. Bayesian inference was then used to determine the model parameters. Using a probability model in this way enlarged PCA into a 'generative' method - new data points could be generated either in the original data space or in the lower-dimensional latent space simply by sampling from the appropriate distribution. The method of pPCA was shown to be equivalent to PCA when the model variance tended to zero. Unlike classical PCA, the principal axes produced by pPCA were subject to an arbitrary orthogonal rotation. Probabilistic PCA was recently extended to the natural exponential family (NEF) of distributions in [16, 17]. Since the method proposed by [16] is explicitly defined for the Poisson distribution with log-Normal canonical parameter, it has been called Poisson log-Normal PCA (PLNPCA).

Neural networks (NNs) are general nonlinear functions which can approximate any continuous function [20, 38, 37]. These nonlinear functions are constructed from successive applications of affine transformations which are separated by intermediate nonlinear 'activation' functions [6]. Though NNs are highly nonlinear through the application of their activation functions, they have been shown to be locally piecewise linear [58]. An autoencoder is a combination of two NNs that uses one NN (the encoder) to reduce the dimensionality of a dataset and another NN (the decoder) to optimally reconstruct the data from its lower-dimensional representation [32]. Kingma and Welling [46] introduced the Variational Autoencoder (VAE), a method for dimension reduction which followed the probabilistic method of [75] with the encoding section of an autoencoder acting as the dimension reducing mapping.

Another related method for dimension reduction, called multi-dimensional scaling (MDS), was introduced by Torgerson in [77]. Instead of using the dataset itself, MDS uses the distances between data points in its formulation. Though this method differed from PCA, it later inspired the so-called 'kernel' methods of dimension reduction, including kernel PCA. Modern formal treatments of classical PCA, pPCA, and MDS can be found in [52], [43], and [9]. Kernel PCA was

introduced in [68]. This method was presented as a nonlinear dimension reduction technique that used a bivariate nonlinear 'kernel' function to implicitly compute the mapping to the lower dimensional space. The method involved diagonalizing the 'kernel matrix' of pairwise distances between data points, where the pairwise distances were computed using the kernel function. While kernel PCA was nonlinear and used PCA in the diagonalization procedure, it differed from classical PCA in that the diagonalization occurred in a possibly infinite dimensional Hilbert space [67]. Further nonlinear dimension reduction techniques that diagonalized the matrix of pairwise distances between data points were ISOMAP [74], Laplacian Eigenmaps (LE) [8], and Diffusion Maps (DM) [19]. A novel nonlinear method for dimension reduction that aimed to preserve the distance between the mapped coordinates of the data, called Locally Linear Embedding (LLE) was proposed in [64]. Though these nonlinear methods differed in their approach to the dimension reduction problem, they were all shown to be variants of kernel PCA with a specially chosen kernel function in [33]. Diffusion Maps were shown to be both equivalent to kernel PCA and to Markov chains in the original work [19]. All of these methods used the 'manifold hypothesis', the assumption that data points lie on low-dimensional manifolds [78] embedded in high-dimensional real space [32].

A comprehensive comparison of the use of classical PCA, kernel methods, and autoencoders on synthetic data is found in [79]. More recent surveys of dimensionality reduction which include variational autoencoders can be found in [5, 28]. Classical PCA and nonlinear kernel methods are implemented using the Python programming language in [59].

Table 2.2: Summary table of the relative advantages and disadvantages of the main dimension reduction models considered in this chapter.

Technique	**Linear**	**Nonlinear**	**Generative**	**Interpretable**
PCA	Yes	No	No	Yes
pPCA	Yes	No	Yes	Yes
kPCA	No	Yes	No	Implicit mapping
PLNPCA	Yes	No	Yes	Yes
VAE	No	Yes	Yes	No

2.4 Principal Components Analysis

Let $\{Y_i\}_{i=1}^N \subset \mathbb{R}^p$ be a collection of random variables. The goal of Principal Components Analysis (PCA) is to find the best possible linear mapping from $\mathbb{R}^p$ to a lower dimensional space, say $\mathbb{R}^q$, where $q < p$. Of course, we need to specify exactly what we mean by 'best possible' mapping. The two main mathematical formulations of PCA address two different criteria for selecting the optimal mapping between spaces. However, both formulations lead to the exact same technique. In both cases, the optimal linear mapping from p to q dimensions is given by a projection of the data onto the eigenvectors corresponding to the q largest eigenvectors of the data covariance matrix. These eigenvectors are often called the 'principal' eigenvectors. The details for both formulations of PCA are from [9].

2.4.1 Minimum Error Formulation of PCA

Suppose $q < p \in \mathbb{R}$, and let $U := \{u_j\}_{j=1}^p \subset \mathbb{R}^p$ be an orthonormal basis of $\mathbb{R}^p$. Since these vectors form an orthonormal basis, we can express any point in our dataset in a basis expansion:

$$y_i = \sum_{j=1}^{p} (u_j^T y_i) u_j \tag{2.2}$$

PCA can be described as the process by which we search for q basis vectors that optimally preserve the squared distances between data points and their 'reconstructions'. WLOG, suppose that we choose the first q vectors from U as our mapping to $\mathbb{R}^q$. We then approximate points in our dataset as:

$$\hat{y}_i = \sum_{j=1}^{q} \alpha_{ij} u_j + \sum_{j=q+1}^{p} \beta_j u_j = \hat{y}_i + \tilde{y}_i. \tag{2.3}$$

The first term in (2.3) gives the approximation of y_i in $\mathbb{R}^q$, while the second term serves as a measure of the approximation error. Note that the constants in the second term are the same for every data point. We want the basis vectors $\{u_j\}_{j=1}^q$ to work for every point in the dataset, so we will minimize the squared deviation between points and their approximations averaged over the dataset. Together with the orthonormality conditions on the basis vectors u_j, we now have the

constrained optimization problem:

$$\min_{\alpha_{ij},\beta_j,u_j} \frac{1}{N}\sum_{i=1}^{N} ||y_i - \hat{y}_i||_2^2 \tag{2.4}$$

s.t.

$$u_j^T u_k = \delta_{jk}. \tag{2.5}$$

The associated Lagrangian for this problem is given by:

$$\begin{aligned} L(\boldsymbol{\alpha}_i, \boldsymbol{\beta}, u_j, \lambda_j) &= \frac{1}{N}\sum_{i=1}^{N} ||y_i - \hat{y}_i||_2^2 - \sum_{j=1}^{p} \lambda_j (u_j^T u_j - 1) \\ &= \frac{1}{N}\sum_{i=1}^{N} \left\| y_i - \sum_{j=1}^{q} \alpha_{ij} u_j - \sum_{j=q+1}^{p} \beta_j u_j \right\|_2^2 - \sum_{j=1}^{p} \lambda_j (u_j^T u_j - 1), \end{aligned} \tag{2.6}$$

where $\boldsymbol{\alpha}_i = [\alpha_{i,1}, \ldots, \alpha_{i,q}]^T$ and $\boldsymbol{\beta} = [\beta_{q+1}, \ldots, \beta_p]^T$ are the respective coefficients from the approximation in (2.3).

First, taking the derivative of (2.6) with respect to α_{ij} and utilizing the orthogonality of the u_j gives:

$$\frac{\partial L}{\partial \alpha_{ij}} = \frac{2}{N}(-y_i^T u_j + \alpha_{ij}). \tag{2.7}$$

Setting this to zero gives the optimal approximation coefficients:

$$\alpha_{ij} = y_i^T u_j. \tag{2.8}$$

Next, we minimize L with respect to β_j:

$$\frac{\partial L}{\partial \beta_j} = \frac{1}{N}\sum_{i=1}^{N} \left[-2y_i^T u_j + 2\beta_j\right]. \tag{2.9}$$

Then, the optimal coefficients β_j are:

$$\begin{aligned} \beta_j &= \frac{1}{N}\sum_{i=1}^{N} y_i^T u_j \\ &= \bar{y}^T u_j, \end{aligned} \tag{2.10}$$

where $\bar{y}$ is the sample mean.

We are finally ready to minimize (2.6) with respect to the vectors u_j. We will first substitute the optimal values of α_{ij} and β_j to obtain the modified Lagrangian:

$$\ell(u_j,\lambda_j) = \frac{1}{N}\sum_{i=1}^{N}\left\| y_i - \sum_{j=1}^{q}\left(y_i^T u_j\right)u_j - \sum_{j=q+1}^{p}\left(\bar{y}^T u_j\right)u_j \right\|_2^2 - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1). \tag{2.11}$$

Substituting the basis expansion (2.2) gives:

$$\begin{aligned}
\ell(u_j,\lambda_j) &= \frac{1}{N}\sum_{i=1}^{N}\left\| \sum_{j=q+1}^{p}\left(y_i^T u_j\right)u_j - \left(\bar{y}^T u_j\right)u_j \right\|_2^2 - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left\| \sum_{j=q+1}^{p}\left[(y_i-\bar{y})^T u_j\right]u_j \right\|_2^2 - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j,k}\left(\left[(y_i-\bar{y})^T u_j\right]u_j\right)^T\left[(y_i-\bar{y})^T u_k\right]u_k\right] - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j,k}u_j^T\left[u_j^T(y_i-\bar{y})(y_i-\bar{y})^T u_k\right]u_k\right] - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j,k}\left[u_j^T(y_i-\bar{y})(y_i-\bar{y})^T u_k\right]u_j^T u_k\right] - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j,k}\left[u_j^T(y_i-\bar{y})(y_i-\bar{y})^T u_k\right]\delta_{jk}\right] - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1) \\
&= \sum_{j=q+1}^{p}u_j^T S u_j - \sum_{j=1}^{p}\lambda_j(u_j^T u_j - 1),
\end{aligned} \tag{2.12}$$

where S is the sample covariance matrix. Now we take the derivative with respect to u_j:

$$\frac{\partial \ell}{\partial u_j} = 2Su_j - 2\lambda_j u_j. \tag{2.13}$$

Setting this to zero gives the solution:

$$Su_j = \lambda_j u_j. \tag{2.14}$$

In other words, choosing eigenvectors of the sample covariance matrix as our basis vectors will minimize the approximation error in (2.4). Consider again the approximation error:

$$\tilde{y}_i = \sum_{j=q+1}^{p}\left(\bar{y}^T u_j\right)u_j. \tag{2.15}$$

This quantity is minimized when the u_j are chosen to be the eigenvectors that correspond to the $p-q$ smallest eigenvalues of S. Thus, the vectors u_j used in the approximation expansion

$$\hat{y}_i = \sum_{j=1}^{q} \left(y_i^T u_j\right) u_j \tag{2.16}$$

should be the eigenvectors corresponding to the q largest eigenvalues of S.

2.4.2 Maximum Variance Formulation of PCA

We can also non-trivially reformulate PCA as a maximum variance problem. We again begin with a sample dataset of random variables $\{Y_i\}_{i=1}^N \subset \mathbb{R}^p$ and an unknown linear mapping $U : \mathbb{R}^p \to \mathbb{R}^q$, $q < p$. During the process of mapping the data to q-dimensional space, we want to maximize the amount of sample variance accounted for by each of the q independent dimensions. It turns out that the mapping U is a $p \times q$ matrix with columns formed by the eigenvectors corresponding to the first q largest eigenvalues of the sample covariance matrix.

To put this formally, we will first map the sample data to a one-dimensional space using the vector u_1. That is, the i^{th} sample point y_i is mapped to $u_1^T y_i$. We will also require that $u_1^T u_1 = 1$. To evaluate the variance in the lower-dimensional space using Definition 0.1.4, we will need to calculate the lower-dimensional mean. This mean is given by:

$$E(u_1^T Y) = \frac{1}{N}\sum_{i=1}^{N} u_1^T y_i = u_1^T \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) = u_1^T \bar{y}, \tag{2.17}$$

where $\bar{y}$ is the empirical mean. The variance is then:

$$\begin{aligned}
\text{Var}\left(u_1^T Y\right) &= E\left[\left(u_1^T Y - E(u_1^T Y)\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[u_1^T y_i - u_1^T \bar{y}\right]^2 \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i^T u_1 u_1^T y_i - 2y_i^T u_1 u_1^T \bar{y} + \bar{y}^T u_1 u_1^T \bar{y} \\
&= \frac{1}{N}\sum_{i=1}^{N} u_1^T y_i y_i^T u_1 - 2u_1^T \bar{y} y_i^T u_1 + u_1^T \bar{y}\bar{y}^T u_1 \\
&= u_1^T \left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(y_i - \bar{y})^T\right] u_1 \\
&= u_1^T S u_1, \qquad (2.18)
\end{aligned}$$

where S is the empirical covariance matrix. To choose the best unit vector u_1 that maximizes this variance, we use Lagrange's method with the Lagrangian:

$$u_1^T S u_1 - \lambda_1(u_1^T u_1 - 1). \qquad (2.19)$$

Taking the derivative of the Lagrangian with respect to u_1 and setting it to zero gives the solution:

$$S u_1 = \lambda_1 u_1. \qquad (2.20)$$

That is, the vector u_1 that maximizes the variance in 1-dimensional space is given by an eigenvector of S. To see exactly which eigenvector this is, multiply (2.20) by u_1^T to get:

$$u_1^T S u_1 = \lambda_1. \qquad (2.21)$$

Therefore, the mapped variance is maximized when u_1 is the eigenvector corresponding to the largest eigenvalue of S. To show how to map Y to a q-dimensional space, we will use induction to prove that projections onto further dimensions of $\mathbb{R}^q$ are given by the eigenvectors corresponding to the $q-1$ next largest eigenvalues of S. Note that this is posed as Exercise 12.1 in [9].

Taking the above as the base case, we assume that the optimal mapping of the dataset onto $\mathbb{R}^{j-1}$ is given by the first $j-1$ eigenvectors of S (where the eigenvectors are ordered according to

their eigenvalues). Call these vectors u_k, $1 \leq k \leq j-1 < q$. Next, select a new vector u_j that is orthogonal to all previously chosen vectors and has norm 1. That is $u_j^T u_k = \delta_{jk}$, where δ_{jk} is the Dirac delta. We again find the optimal value of u_j using the Lagrangian

$$u_j^T u_j - \lambda_j (u_j^T u_j - 1) \tag{2.22}$$

with the familiar solution

$$S u_j = \lambda_j u_j. \tag{2.23}$$

Again applying the orthonormality of u_j tells us that the optimal mapping onto the j^{th} dimension is given by the j^{th} eigenvector of S.

A useful consequence of this formulation of PCA is the correspondence between the eigenvectors and the sample variance that they explain. Mapping the data to a one-dimensional space is achieved by projecting the data onto the first eigenvector of the sample covariance matrix. We can then say that this one-dimensional representation 'explains' a portion of the sample variance. The amount of variance explained by the one-dimensional mapping is simply given by the largest eigenvalue of S. If we want to know the proportion of total variance explained, then we can do a full spectral decomposition of S and divide the largest eigenvalue by the sum of all eigenvalues.

2.5 Probabilistic PCA

Probabilistic PCA (pPCA) was introduced by Tipping and Bishop in [75]. Similarly to classical PCA, pPCA is also used for dimension reduction with linear mappings. However, a major difference is the assumption of probability density functions on the observed data and on unobserved 'latent' variables. These latent variables represent the positions of the mapped data in the lower dimensional space. Unlike classical PCA, pPCA is generative, meaning that new data points in either the original space or the lower-dimensional latent space may be generated by simply sampling from the appropriate distribution. This provides the additional advantage that missing data may be imputed through sampling.

Suppose that we are given a set of observations of random variables $\{Y_i\}_{i=1}^N \subset \mathbb{R}^p$. We assume that there exist latent variables $\{W_i\}_{i=1}^N \subset \mathbb{R}^q$, $q < p$, and $W_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$. We want to use these latent variables to explain affine structure in the observations. To this end, we seek a linear mapping B such that

$$Y_i = \mu + BW_i + \varepsilon_i, \tag{2.24}$$

where $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is random observation noise. This induces a distribution on the variables Y_i, and we can then say that $Y_i \sim \mathcal{N}(\mu, BB^T + \sigma^2 \boldsymbol{I})$. The authors of [16] state this in an alternate form by introducing intermediate variables $Z_i = \mu + BW_i$. Then $Z_i \sim \mathcal{N}(\mu, BB^T)$, and we can rewrite (2.24) as

$$\begin{aligned} \{Z_i\}_{i=1}^N \text{i.i.d.}, \quad & Z_i \sim \mathcal{N}(\mu, \Sigma), \\ \{Y_i | Z_i\} \text{ independent}, \quad & Y_{ij} | Z_{ij} \sim \mathcal{N}(Z_{ij}, \sigma^2), \end{aligned} \tag{2.25}$$

where $\Sigma = BB^T$. For inference of the model parameters using maximum likelihood estimation, we first compute the marginal probability of the observed variables:

$$\begin{aligned} p(y_i|\theta) &= \int p(y_i|w_i, \theta) dw_i \\ &= (2\pi)^{-\frac{p}{2}} |C|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (y_i - \mu)^T C^{-1} (y_i - \mu) \right), \end{aligned} \tag{2.26}$$

where the model parameters have been collected by $\theta = \{\sigma^2, \mu, B\}$, and we have denoted the model covariance $C = \sigma^2 \boldsymbol{I} + BB^T$. Later, when we introduce the piecewise probabilistic PCA model in Chapter 3, we will elucidate some of the details in the computation of (2.26).

Finding the optimal mapping to $\mathbb{R}^q$ given the model (2.25) is done by maximizing the log-likelihood of the data, which is given by

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^N \log p(y_i|\theta) \\ &= -\frac{Np}{2} \log 2\pi - \frac{N}{2} \left[\log |\Sigma| + \operatorname{Tr}\left(\Sigma^{-1} S \right) \right], \end{aligned} \tag{2.27}$$

where S is the covariance matrix of the observations:

$$S := \frac{1}{N} \sum_{i=1}^N (y_i - \mu)(y_i - \mu)^T. \tag{2.28}$$

By taking the derivative of (2.27) with respect to B and setting it to zero, Tipping and Bishop [75] showed that the corresponding non-zero critical points are given by

$$\widehat{B} = U(\Lambda - \sigma^2 \boldsymbol{I})^{\frac{1}{2}} R, \tag{2.29}$$

where Λ is a $q \times q$ diagonal matrix with the q largest eigenvalues of S on its diagonal, U is a $p \times q$ matrix with the corresponding q eigenvectors as its columns, and R is an arbitrary orthogonal matrix. Since R is orthogonal, it is generally thought of as a rotation in the latent space. Thus, the projection down to q dimensions is unique up to arbitrary rotations.

Remark. When $\sigma^2 \to 0$, the model covariance matrix $\sigma^2 \boldsymbol{I} + BB^T \to BB^T$, yielding the classical PCA obtained through maximizing the model variance in § 2.4.2.

It was further shown in [75] that when $\widehat{B}$ is substituted back into the log-likelihood the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{p-q} \sum_{j=q+1}^{p} \lambda_j, \tag{2.30}$$

where $\{\lambda_j\}_{j=q+1}^{p}$ are the $p-q$ smallest eigenvalues of S. Thus, $\widehat{\sigma}^2$ is a measure of the variance in the observations which is not accounted for by the model.

The variance explained by each dimension of the PCA model is captured by the scaled eigenvalues of the sample covariance matrix S (see (2.28)). Suppose that we order the eigenvectors of S by the descending magnitude of their corresponding eigenvalues. The proportion of the variance in the sample which is explained by projecting the data onto the j^{th} eigenvector of S is then given by

$$v_j = \frac{\lambda_j}{\sum_{i \neq j} \lambda_i}. \tag{2.31}$$

2.5.1 Poisson PCA

Chiquet, Mariadassou, and Robin generalized pPCA by assuming prior NEF distributions in [16]. For instance, by assuming a Poisson distribution, the general model in [16] can be used for dimension reduction with count variables. In this section, we will present the model introduced in [16].

As with pPCA, we begin with the set of observations $\{Y_i\}_{i=1}^N$ with an NEF distribution conditioned on the latent variables $\{W_i\}_{i=1}^N$. We assume the same linear model for dimension reduction as (2.24), with one notable extension. Suppose that $X_i \in \mathbb{R}^d$ is a vector of observed covariates, and we suspect a linear relationship between these covariates and the NEF-distributed observations Y_i. The linear effects of these covariates are included in the model through the matrix $\Theta \in \mathbb{R}^{p\times d}$, and intercepts are included through the offset parameter O_i. Thus, the model which accounts for the linear effect of known covariates is given by

$$\begin{aligned} \{W_i\}_{i=1}^N \text{i.i.d.}, \quad W_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \\ Z_i &= O_i + \Theta X_i + BW_i, \\ \{Y_{ij}|Z_{ij}\} \text{ independent}, \quad p(Y_{ij}|Z_{ij}) &= h(Y_{ij})\exp\left(Y_{ij}Z_{ij} - \psi(Z_{ij})\right). \end{aligned} \tag{2.32}$$

The direct analogue to the pPCA model can be obtained by setting $O_i \equiv \mu$ for all i, and setting $\Theta \equiv \mathbf{0}$.

The marginal data distribution leads to difficulties in the maximum likelihood estimation of the model parameters. Therefore, we proceed with maximizing the variational lower bound (VLB) as presented in § 1.3.1. As an approximation to the posterior distribution $p(w|y,\theta)$, let $\tilde{p} = \otimes_{i=1}^N \tilde{p}_i$ be a product of Gaussian distributions, each with mean vector m_i and diagonal covariance matrix Σ_i. We also let the elements on the diagonal of Σ_i be strictly positive for all i. Thus, our variational distribution is given by

$$\begin{aligned} \tilde{p}(w) &= \prod_{i=1}^N \tilde{p}_i(w_i) \\ &= \prod_{i=1}^N (2\pi)^{-\frac{q}{2}} |\Sigma_i|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(w_i - m_i)^T \Sigma_i^{-1}(w_i - m_i)\right) \\ &= (2\pi)^{-\frac{Nq}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^N (w_i - m_i)^T \Sigma_i^{-1}(w_i - m_i)\right) \prod_{i=1}^N \left[\prod_{j=1}^q s_{ij}\right]^{-\frac{1}{2}}, \end{aligned} \tag{2.33}$$

where s_{ij} is the j^{th} element on the diagonal of the i^{th} variational covariance matrix. The VLB for

Poisson PCA is given by

$$\mathcal{L}(\tilde{p},\theta) = \mathbf{1}_N^T \left[\boldsymbol{Y} \odot \left(\boldsymbol{O} + \boldsymbol{X}\Theta^T + \boldsymbol{M}B^T \right) - \exp\left(\boldsymbol{O} + \boldsymbol{X}\Theta^T + \boldsymbol{M}B^T + \frac{1}{2}(\boldsymbol{S}\odot\boldsymbol{S})(\boldsymbol{B}\odot\boldsymbol{B})^T \right) \right] \mathbf{1}_p$$
$$- \frac{1}{2}\mathbf{1}_N^T \left[\boldsymbol{M}\odot\boldsymbol{M} + \boldsymbol{S}\odot\boldsymbol{S} - 2\log\boldsymbol{S} - \mathbf{1}_{n,q} \right] \mathbf{1}_q - \mathbf{1}_N^T a(\boldsymbol{Y}) \mathbf{1}_p, \tag{2.34}$$

where,

$$\boldsymbol{Y} = \begin{bmatrix} Y_1^T \\ \vdots \\ Y_N^T \end{bmatrix} \in \mathbb{R}^{N\times p}, \quad \boldsymbol{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_N^T \end{bmatrix} \in \mathbb{R}^{N\times d}, \quad \boldsymbol{M} = \begin{bmatrix} m_1^T \\ \vdots \\ m_N^T \end{bmatrix} \in \mathbb{R}^{N\times q}, \tag{2.35}$$

$$\boldsymbol{O} = \begin{bmatrix} O_1^T \\ \vdots \\ O_N^T \end{bmatrix} \in \mathbb{R}^{N\times d}, \quad \boldsymbol{S} = \begin{bmatrix} \mathrm{diag}(\Sigma_1) \\ \vdots \\ \mathrm{diag}(\Sigma_N) \end{bmatrix} \in \mathbb{R}^{N\times q}, \tag{2.36}$$

$\odot$ is the Hadamard product, $\mathbf{1}_k = [1,\ldots,1]^T \in \mathbb{R}^k$, and

$$a(\boldsymbol{Y}) = \begin{bmatrix} \log(Y_{11}!) & \cdots & \log(Y_{1p}!) \\ \vdots & \ddots & \vdots \\ \log(Y_{N1}!) & \cdots & \log(Y_{Np}!) \end{bmatrix}. \tag{2.37}$$

Rather than estimate the model parameters with maximum likelihood estimation or an EM algorithm, the authors of [16] used the method of moving asymptotes (MMA) algorithm defined in [73] to maximize the VLB. They further showed that the VLB is maximized by the estimated covariance matrix

$$\widehat{\Sigma} = \frac{1}{N}\widehat{B}\left(\widetilde{\boldsymbol{M}}^T\widetilde{\boldsymbol{M}} + \mathrm{diag}\left[\mathbf{1}_N^T(\widetilde{\boldsymbol{S}}\odot\widetilde{\boldsymbol{S}}) \right] \right)\widehat{B}^T, \tag{2.38}$$

where $\widehat{B}$ are the optimal model parameters, $\widetilde{\boldsymbol{M}}$ is the matrix of optimal variational means, and $\widetilde{\boldsymbol{S}}$ is a matrix containing the diagonals of the optimal variational covariance matrices. Once these optimal

model and variational parameters are found, we can express estimates of the mapped data in the latent space under the linear model as

$$\widetilde{Z} = \boldsymbol{O} + \boldsymbol{X}\widehat{\Theta}^T + \widetilde{\boldsymbol{M}}\widehat{B}^T. \tag{2.39}$$

Evaluating the Model

The authors of [16] developed a log-likelihood ratio for evaluating their model. Define ℓ_m as the log-likelihood for the model with its associated parameters, $\ell_{\max}$ as the log-likelihood of the model with one parameter per observation, and $\ell_{\min}$ as the log-likelihood of the model with only a single intercept parameter. The pseudo R^2 value is defined as

$$R^2 := \frac{\ell_m - \ell_{\min}}{\ell_{\max} - \ell_{\min}}. \tag{2.40}$$

This metric is intended to be a measure of the proportion of variance accounted for by the Poisson PCA model. When performing visualizations of the latent space, the authors orthogonalized $\widehat{B}^T$ and sorted the columns of $\widetilde{M}$ in order of decreasing variation by completing a classical PCA on the mapped points $\widetilde{M}\widehat{B}^T$. This addition of classical PCA allowed them to decompose the variance explained by the principal components as $\lambda_i R^2$, where λ_i is the i^{th} eigenvalue of $\widetilde{M}\widehat{B}^T$.

In Poisson PCA, the number of latent dimensions is not known in advance but is a parameter that must be estimated. Two information criteria were established to guide the selection of the lower dimension. The Bayesian Information Criterion (BIC) and the Integrated Completed Likelihood (ICL) were defined as

$$BIC(q) := \mathcal{L} - \frac{1}{2}p(d+q)\log N, \tag{2.41}$$

and

$$ICL(q) := \mathcal{L} - \frac{1}{2}p(d+q)\log N - \frac{nq}{2}\log 2\pi e - \mathbf{1}_N^T \log(\boldsymbol{S})\mathbf{1}_q, \tag{2.42}$$

where $p(d+q)$ is the number of parameters in the model, N is the number of observations, and $\mathcal{L}$ is the variational lower bound for the log-likelihood.

In practice, the latent dimensionality is started at $q = 1$, the model parameters are estimated, and then the BIC and ICL are computed. This process continues, increasing q by one unit at each

step, until a maximum in either the BIC or the ICL is reached. The user then chooses the latent dimensionality corresponding to the optimal criterion.

2.6 Kernel PCA

Kernel PCA (kPCA) is a technique for nonlinear dimension reduction which inherits some features of classical PCA. However, in order to accomplish dimension reduction, kPCA first projects the data into an infinite dimensional Hilbert space using a 'kernel' function. This function is used to represent an implicitly defined nonlinear dimension reduction mapping Φ.

Recalling Definition 0.4.1, a kernel is a real-valued function over a product space. For example, consider the product space $\mathbb{R} \times \mathbb{R}$. A standard example of a kernel over this space is the 2-dimensional Gaussian function defined by:

$$f(x,y) = \exp\left(-(x^2+y^2)\right). \tag{2.43}$$

This particular kernel is symmetric, but that need not be the case in general. Non-symmetric polynomial kernels of odd degree are also possible, such as:

$$f(x,y) = a_1x^3 + a_2y^3 + a_3x^2y + a_4xy^2 + a_5xy + a_6x + a_7y + a_8, \quad a_i \in \mathbb{R}. \tag{2.44}$$

A kernel is often used to encode prior assumptions or knowledge about the dataset. The main theoretical result that shows the effectiveness of kernels is known as the 'kernel trick' or Representer Theorem (see Theorem 6). This result proves that every positive definite symmetric (PDS) kernel defines an implicit mapping from a general metric space into a reproducing kernel Hilbert space (RKHS) [56, 67]. The idea is that once the data is mapped to a Hilbert space, we can take advantage of the Euclidean geometry there to form a covariance matrix which may then be diagonalized. In this section, we present the construction of the covariance matrix under the nonlinear mapping Φ and its relationship to the kernel function. The details that follow fill in the ideas presented in [33].

As mentioned above, the mapping Φ is implicitly defined. To achieve dimension reduction, we diagonalize the covariance matrix of the implicitly mapped data points. Let $\{y_i\}_{i=1}^N \in \Omega$ and $k : \Omega \times \Omega \to \mathbb{R}$ a PDS kernel. The covariance matrix in the RKHS is given in [68] by

$$C := \frac{1}{N} \sum_{i=1}^{N} \Phi(y_i)\Phi(y_i)^T. \tag{2.45}$$

Our goal is to diagonalize the matrix C. That is, we are looking for solutions to the eigenvalue problem

$$C\boldsymbol{v} = \lambda \boldsymbol{v}. \tag{2.46}$$

This means that the eigenvector solutions $\boldsymbol{v}$ to (2.46) are linear combinations of the images of the data points under Φ. In other words,

$$\boldsymbol{v} = \sum_{i=1}^{N} \alpha_i \Phi(y_i). \tag{2.47}$$

Plugging (2.47) into (2.46), Theorem 6 allows us to express the eigenvalue problem in terms of the kernel k.

$$\begin{aligned} C\boldsymbol{v} = \lambda \boldsymbol{v} \iff \left(\frac{1}{N} \sum_{i=1}^{N} \Phi(y_i)\Phi(y_i)^T \right) \left(\sum_{j=1}^{N} \alpha_j \Phi(y_j) \right) &= \lambda \sum_{i=1}^{N} \alpha_i \Phi(y_i) \\ \frac{1}{N} \sum_{i=1}^{N} \Phi(y_i) \left(\sum_{j=1}^{N} \alpha_j \Phi(y_i)^T \Phi(y_j) \right) &= \lambda \sum_{i=1}^{N} \alpha_i \Phi(y_i) \\ \frac{1}{N} \sum_{i=1}^{N} \Phi(y_i) \left(\sum_{j=1}^{N} \alpha_j k(y_i, y_j) \right) &= \lambda \sum_{i=1}^{N} \alpha_i \Phi(y_i). \end{aligned} \tag{2.48}$$

Since (2.48) is true for each $1 \leq i \leq N$, we have

$$K\boldsymbol{\alpha} = N\lambda\boldsymbol{\alpha} = \tilde{\lambda}\boldsymbol{\alpha}, \tag{2.49}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$ is the vector of coefficients from (2.48), and $[K]_{ij} = k(y_i, y_j)$ is the kernel matrix. Moving forward, we will absorb the scaling by N into the eigenvalues and refer to the resulting scaled eigenvalues as simply λ.

Typically, a point y would be mapped to the implicitly defined RKHS by projecting its image under Φ onto the eigenvectors $\boldsymbol{v}$. However, the Representer Theorem again enables the computation of such projections through the kernel function. The projection is calculated as

$$\begin{aligned}\langle \boldsymbol{v}, \Phi(y) \rangle &= \frac{1}{\sqrt{\lambda}} \left\langle \sum_{i=1}^{N} \alpha_i \Phi(y_i), \Phi(y) \right\rangle \\ &= \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{N} \alpha_i \langle \Phi(y_i), \Phi(y) \rangle \\ &= \frac{1}{\sqrt{\lambda}} \sum_{i=1}^{N} \alpha_i k(y_i, y), \end{aligned} \tag{2.50}$$

where the factor of $\lambda^{-1/2}$ is included to ensure that $\langle \boldsymbol{v}, \boldsymbol{v} \rangle = 1$ [33].

In summary, kPCA reduces the dimensionality of a dataset with nonlinear structure through the diagonalization of the kernel matrix. By the Representer Theorem, the kernel matrix provides the same information as the covariance matrix of the data under the implicitly defined nonlinear mapping Φ. We project the data onto the first q eigenvectors of the kernel matrix and scale by the corresponding eigenvalues.

Kernel PCA is an extremely useful tool for dimensionality reduction, but it does bring with it some disadvantages. First, the kernel function must be defined at the outset, and designing kernels requires expert knowledge about the particular dataset and its underlying nonlinear structure. In contrast to classical PCA where a linear mapping is given by a matrix multiplication, kPCA has no explicitly defined parameters besides those used in the kernel function. Therefore, we lose the interpretability of the kPCA model. There are additional computational considerations when using kPCA: the kernel function must be computed for all pairs of data points which adds computational time.

2.7 Autoencoders

Another method for reducing dimensions is the construction of explicit nonlinear maps between linear spaces of different dimensionality. As we will see later, linear and piecewise linear maps

are relatively simple to optimize. However, when highly nonlinear maps such as neural networks are used for solving the probabilistic dimension reduction problem posed by (2.25), we arrive at the autoencoder. Since autoencoders share so much in common with neural networks, we will first introduce the structure and construction of neural networks.

2.7.1 Feedforward Neural Networks

A neural network (NN) is a nonlinear function formed by successive applications of linear maps followed by a continuous, possibly nonlinear 'activation' function. Let $Y \in \mathbb{R}^p$ be a random variable, and suppose that we want to optimally map Y to the lower-dimensional space $\mathbb{R}^q$. A 'single hidden layer' neural network can be constructed for this task by introducing a linear map $A^{[0]} : \mathbb{R}^p \to \mathbb{R}^q$. If the computation were to stop here, then we would have a simple linear mapping between two linear spaces of different dimension. However, NNs typically introduce additional parameters to the model. After Y has been transformed by the linear mapping $A^{[0]}$, sometimes called the 'input layer', we apply the 'activation' function element-wise, i.e. $\sigma : \mathbb{R}^q \to \mathbb{R}^q$. The most common activation function is the 'rectified linear unit', or `ReLU` function, which is defined by:

$$\texttt{ReLU}(Y)_j := \max_{1 \le i \le q} \{0, y_i\}. \tag{2.51}$$

After the activation function is applied, a second linear map $A^{[1]} : \mathbb{R}^q \to \mathbb{R}^q$ gives the final output. A 'layer' is therefore composed of a linear map followed by the application of the activation function. To recap, the entire structure of the single hidden layer NN, from input to output, is given by:

$$Y = A^{[1]} \sigma \left(A^{[0]} Y \right). \tag{2.52}$$

The NN given by (2.52) has pq^3 parameters to be optimized. Note that we may make either of the linear maps $A^{[0]}, A^{[1]}$ into affine transformations by appending a p-dimensional column vector of ones to the dataset. We will discuss the most frequently employed method of optimization in the next section, but we first provide a general definition of 'feedforward' NNs and a discussion of their properties.

Definition 2.7.1. *Given a non-empty set $S = Y \cup \{\boldsymbol{e}\} \subset \mathbb{R}^p$, $\boldsymbol{e} = [1,\dots,1]^T$, a feedforward NN is a continuous map $\mathcal{N} : S \to \mathbb{R}^q$ defined by*

$$\mathcal{N}(x) := \sigma^{[L]}\left(A^{[L]}\sigma^{[L-1]}\left(\cdots A^{[1]}\sigma^{[0]}\left(A^{[0]}y\right)\right)\right), \tag{2.53}$$

where $\{A^{[\ell]}\}_{\ell=0}^{L}$ are affine transformations and $\{\sigma^{[\ell]}\}_{\ell=0}^{L}$ are continuous 'activation functions'.

NNs of the form (2.53) are, in fact, piecewise linear when the activation functions $\sigma^{[\ell]}$ are piecewise linear [58], such as in the common case when using the ReLU function (2.51). For the sake of illustration, we will show the forward pass of a single data vector through the NN. Suppose we are given a data point $X \in \mathbb{R}^p$. As in Definition 2.7.1, we will append a single unit as the first element of our data vector. The presence of this unit ($\boldsymbol{e}$ in Definition 2.7.1) simply allows us to fit an intercept (sometimes called a bias) to the model. In other words, instead of applying strictly linear transformations, we are able to apply affine transformations. With this in mind, our data vector now has dimension $p+1$ and can be written as:

$$X := [1, X_1, \dots, X_p]^T. \tag{2.54}$$

The first 'layer' of the NN first multiplies X with a $k \times p$ matrix $A^{[0]}$:

$$\begin{aligned} z^{[0]} &:= A^{[0]}X \\ &= \begin{bmatrix} \langle A_1^{[0]}, X\rangle \\ \langle A_2^{[0]}, X\rangle \\ \vdots \\ \langle A_k^{[0]}, X\rangle \end{bmatrix}, \end{aligned} \tag{2.55}$$

where $A_i^{[0]}$ represents the i^{th} row of $A^{[0]}$ and $\langle\cdot,\cdot\rangle$ is the standard inner product on $\mathbb{R}^p$. Note that the output from this affine transformation is k-dimensional. We are free to choose the value of

k depending on the application [32]. For dimension reduction, we typically choose $k < p$. The computations continue by applying the activation function to obtain the output $a^{[0]} = \sigma^{[0]}\left(z^{[0]}\right)$.

Geometrically, we may think of the input layer as a collection of k hyperplanes. Passing the data vector through the transformation $A^{[0]}$ then amounts to computing the inner product of each data vector with the normal vectors from these proposed hyperplanes. That is, the inner products of the rows of $A^{[i]}$ with X represent a comparison of the data with the hyperplanes. The activation function σ checks which side of these affine hyperplanes the data X is on. If the data vector lies above the planes, then the information in the corresponding output is passed forward. On the other hand, if the data vector lies below any of the hyperplanes, then the corresponding output is set to zero and no information is passed forward.

The computations above represent a pass of the data through a single 'layer' of the network. A 'shallow' NN further computes a second affine transformation $A^{[1]}$ and activation function $\sigma^{[1]}$, which takes the vector $a^{[0]}$ to the output dimension. In the case of regression, $\sigma^{[1]}$ is usually the identity. For binary classification, $\sigma^{[1]}$ is the sigmoid function, and for multiple classes $\sigma^{[1]}$ is the softmax function [32]. These last two functions allow the network output to be interpreted as a probability distribution, and so the class label is assigned as the maximum argument of the respective function output. In other words, the index corresponding to the element of $a^{[L]}$ with the highest value gets assigned as the class label. 'Deep' NNs are simply made up of several layers of activated affine transformations.

2.7.2 Training Neural Networks

At the outset, the affine transformations $\{A^{[\ell]}\}$ are randomly initialized. Of course, this means that the parameters contained in these matrices must be updated based on the data. In many cases, NNs are used to optimally approximate some known output quantity. In such applications, the output of the NN can be compared to the expected output. The empirical mean squared error is typically used to compare the outputs of a NN with the expected values over an entire dataset. To update the

NN parameters, we then solve the unconstrained optimization problem:

$$F = \underset{f}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} ||y_i - f(y_i)||_2^2, \tag{2.56}$$

where F is the desired NN. This optimization is typically done using variants of gradient descent, where the gradients of the objective function in (2.56) are found using the chain rule [66].

2.7.3 Autoencoder Architecture

An autoencoder is constructed by placing two feedforward NNs together in two sections, called the 'encoder' and the 'decoder' [32]. First, the encoder is made up of a NN that maps the p-dimensional input variables down to q dimensions, where $q < p$. The decoder section then maps the q-dimensional output from the encoder back to the original p dimensions. Figure 2.1 contains a graphical representation of the resulting autoencoder.

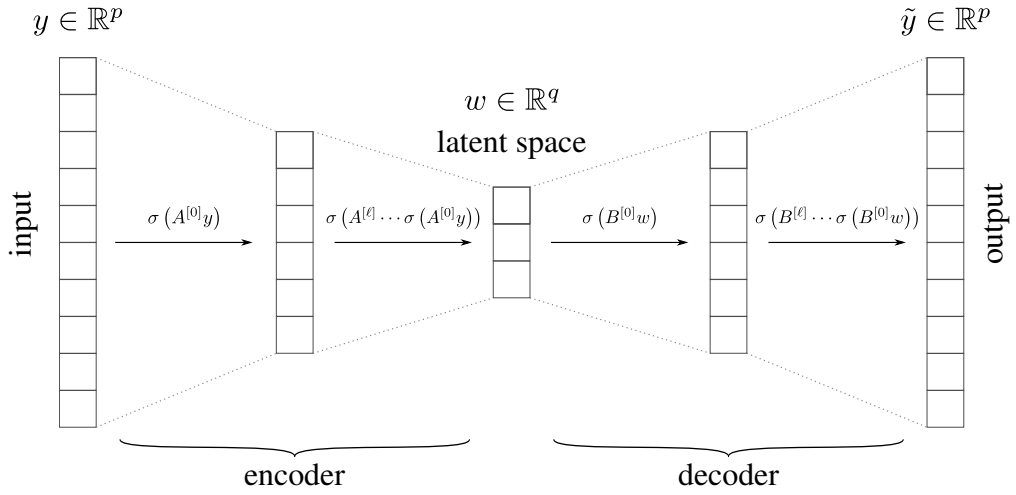


Figure 2.1: Structure of an autoencoder. The encoder section reduces the data dimension from p to q with a feedforward neural network, and the decoder section uses a second feedforward neural network to restore the dimension back to p. Note that the autoencoder in this figure is symmetric, i.e. the same number of affine transformations are used in both the encoder and decoder sections.

Beginning with random variables $\{Y_i\}_{i=1}^N \subset \mathbb{R}^p$, the encoder $h : \mathbb{R}^p \to \mathbb{R}^q$ is a feedforward NN of the form

$$h(y) = A^{[L]}\sigma(A^{[L-1]}\sigma(\cdots A^{[0]}y)), \tag{2.57}$$

where, typically, $L \geq 2$. The decoder $g : \mathbb{R}^q \to \mathbb{R}^p$ is also a feedforward NN of the form

$$g(z) = B^{[K]}\sigma(B^{[K-1]}\sigma(\cdots B^{[0]}z)), \tag{2.58}$$

where K is not necessarily equal to L.

The optimal parameters for the set of affine maps $\boldsymbol{A} := \{A^{[\ell]}\}_{\ell=0}^{L} \cup \{B^{[k]}\}_{k=0}^{K}$ is then found by minimizing the 'reconstruction error', which is measured by the Euclidean distance between the original data and the reconstructed data. The error is minimized by solving the unconstrained optimization problem:

$$\boldsymbol{A} = \underset{\boldsymbol{a}}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} ||y_i - f_{\boldsymbol{a}}(y_i)||_2^2, \tag{2.59}$$

where $f_{\boldsymbol{a}} := g \circ h$ is the autoencoder formed by the composition of the encoder h and the decoder g with combined parameters $\boldsymbol{a}$.

Remark. Other measurements of error may be used instead of the Euclidean reconstruction error in (2.59). When training autoencoders later in this work in Chapter 4, the Euclidean distance will be used.

2.7.4 Variational Autoencoder

The variational autoencoder (VAE), defined in [46], is a nonlinear model for dimension reduction which can be seen as a heavily overparametrized version of pPCA. As in the case of pPCA above, we are given random variables $\{Y_i\}_{i=1}^{N} \subset \mathbb{R}^p$ and latent variables $\{W_i\}_{i=1}^{N} \subset \mathbb{R}^q$, with $q < p$ and $W_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Instead of assuming a linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^q$, we instead let the mapping be defined by a feedforward neural network. We also assume a multivariate Gaussian distribution on the observations, and we approximate the posterior distribution with a product of Gaussians as in (2.33). The model is summarized as

$$\{W_i\}_{i=1}^N \text{i.i.d.}, \quad W_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}),$$

$$Z_i = f(W_i, \theta) =: f_i,$$

$$\{Y_i | Z_i\} \text{ independent}, \quad p(Y_i|Z_i) = \mathcal{N}(f_i, \sigma^2 \boldsymbol{I}), \tag{2.60}$$

where

$$f(W_i, \theta) = \sigma^{[L]}\left(A^{[L]}\sigma^{[L-1]}\left(\cdots A^{[1]}\sigma^{[0]}\left(A^{[0]}W_i\right)\right)\right) \tag{2.61}$$

is a feedforward NN with L layers and parameters given by $\theta := \{A^{[\ell]}\}_{\ell=0}^L$.

Instead of using an EM algorithm to find the optimal parameters for the model, Kingma and Welling proposed using stochastic gradient descent [46]. This optimization technique is a variation of gradient descent (see Definition 0.2.1) where subsets of observations, called 'batches', are used to optimize the parameters, with the resulting log-likelihood being averaged over the batches.

Chapter 3

Piecewise Affine Probabilistic PCA

3.1 Introduction

Suppose that we are presented with a dataset sampled from the surface of a low-dimensional manifold embedded in a high-dimensional real space. Our objective is to map the data points from this high-dimensional space to one of lower dimensions, preferably closer to the dimensionality of the manifold itself. The data points are normally distributed across the surface of the manifold, but the manifold is sufficiently curved such that a simple linear mapping will produce distortions when the points are mapped to the lower dimensional space. Our work provides a probabilistic piecewise linear model that optimally maps the high-dimensional, nonlinear data to a lower-dimensional space. This model is generative, as new points can be generated from either the prior or posterior distribution.

Consider the following simplified example which motivates our model. Suppose that we are given a can of spray paint that emits a Gaussian stream of paint. That is, when the spray paint is applied to a surface from a fixed position, the resulting pattern of paint spots on the surface is approximately Gaussian distributed. Now, imagine using this can of spray paint to paint a flat piece of paper and then folding the paper 90 degrees along a single edge. The resulting blotch of paint will still be roughly Gaussian, but is now spread over a nonlinear surface. Similarly, suppose we align our paper with the xy-plane and apply the Gaussian spray paint. We then split the paper into two halves and give the separate halves different elevations along the vertical z-axis. The blotch of paint will still be approximately Gaussian, but will now be 'broken' across the two surfaces. In both of these examples, the intention behind our model is to optimally reconstruct the two-dimensional surface across which the Gaussian blotches of paint are found.

Our model is a bridge between the linear probabilistic PCA (pPCA) and the highly nonlinear

variational autoencoder. The piecewise model has the advantage over classical and probabilistic PCA in that unimodal distributions may be approximated across the surface of nonlinear manifolds without needing to resort to mixture models.

3.2 Piecewise PCA with Gaussian Distributed Observations

We begin with a finite set of iid data points, $\{Y_i\}_{i=1}^N \subset \mathbb{R}^p$. Corresponding to each point is an unobserved latent variable $W_i \in \mathbb{R}^q$, $q < p$. For convenience, we will choose a prior density function on the latent variable space that is easy to work with later on. In our case, we will assume that $W_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. We want to use maximum likelihood to determine the probability of observing our dataset given the latent variables. Towards this goal, we further assume a functional relationship between the random variables Y and W. That is,

$$Y_i = f(W_i, \theta) + \varepsilon_i, \tag{3.1}$$

where the exact form of the function f with parameters θ will be specified later, and ε_i is the observational noise. We then write the joint probability distribution as

$$p(Y, W|\theta) = \prod_{i=1}^{N} p(Y_i, W_i|\theta) = \prod_{i=1}^{N} p(Y_i|W_i, \theta)p(W_i). \tag{3.2}$$

We also assume a normal distribution with isotropic covariance matrix for the observational noise, $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$.

As an extension of the probabilistic model for PCA in [75], suppose that the latent space is divided into two half-spaces by an affine hyperplane. We will describe this affine hyperplane by its normal vector ν and the intercept parameter β. The latent space will be separated into two subsets depending on whether the points are above or below this hyperplane. That is, we will divide $\mathbb{R}^q$ into the two subsets defined by

$$\begin{aligned} \Omega_+ &:= \left\{ w \in \mathbb{R}^q : \nu^T w + \beta \geq 0 \right\}, \\ \Omega_- &:= \left\{ w \in \mathbb{R}^q : \nu^T w + \beta < 0 \right\}. \end{aligned} \tag{3.3}$$

We are free to put further conditions on the latent space that will make future calculations much easier. At this point, the only simplifying conditions we will impose are that $\beta = 0$ and that ν corresponds to one of the standard basis vectors in $\mathbb{R}^q$. Without loss of generality, we let ν be the q^{th} standard basis vector of $\mathbb{R}^q$. That is, we assume

$$\nu = \boldsymbol{e}_q := [0, \ldots, 0, 1]^T. \tag{3.4}$$

As in [75], we will assume a prior normal distribution on the latent space and a normal distribution on the observed data, conditioned on the latent variables and the model parameters. We will assume that this conditional normal distribution has an isotropic covariance matrix given by $\sigma^2 \boldsymbol{I}$. The piecewise affine model that relates latent variables to the observation space is then given by:

$$f(w, \theta) := \begin{cases} B_1 w + \mu_1, \text{ if } w \in \Omega_+, \\ B_2 w + \mu_2, \text{ if } w \in \Omega_-, \end{cases} \tag{3.5}$$

where θ is the collection of the model parameters. The entire model may be summarized as:

$$\begin{aligned} w \in \mathbb{R}^q, \quad & w \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \\ y \in \mathbb{R}^p, \quad & y = f(w, \theta) + \varepsilon, \\ p(y|w, \theta) = (2\pi\sigma^2)^{-p/2} & \exp\left(-\frac{1}{2\sigma^2} (y - f(w, \theta))^T (y - f(w, \theta)) \right). \end{aligned} \tag{3.6}$$

3.2.1 Relationship to Other Models

The piecewise model in (3.6) is an extension of the probabilistic PCA model defined by Tipping and Bishop ([75]). It can also been seen as a simplification of the autoencoders in § 2.7. As shown in [58], for a feedforward neural network (NN) with L layers, p inputs, hidden layers with dimension p_ℓ at the ℓ^{th} layer, and the `ReLU` activation function at each layer, the resulting constructed function is piecewise linear with at least

$$\left(\prod_{\ell=1}^{L-1} \left\lfloor \frac{p_\ell}{p} \right\rfloor^{p_\ell} \right) \sum_{j=0}^{p} \binom{p_L}{j} \tag{3.7}$$

linear regions. As noted in [32], feedforward NNs need large quantities of data to train accurately. Furthermore, the interpretability of the parameters in the model are lost. Our model does not compute piecewise PCA over as many linear regions as the exponential lower bound in (3.7), but instead offers a relatively simple method for 'straightening out' data lying near a manifold which may be approximated by two or more linear planes. As we will see in § 3.4, we may maximize the log-likelihood of our piecewise model through a variational lower bound with closed-form expressions for the derivatives of the model parameters. This is in contrast to autoencoders, which must be trained solely with numerical approximations such as gradient descent.

The guiding principle in the design of the piecewise PCA model has been to borrow as much as possible from probabilistic PCA while extending the parametrization towards the autoencoder structure. While the above section extends pPCA to a dimension reduction scheme with two linear pieces, we further extend the model to $K \geq 2$ linear regions in § 3.5 below. However, we first investigate the optimization of the model parameters.

3.2.2 Model Inference

We have two options for finding the optimal parameters for this model based on the data. First, we could calculate the marginal distribution of Y dependent on θ, denoted by $p(y|\theta)$, and proceed to maximum likelihood estimation. This marginal data distribution is found by integrating the joint distribution $p(y,w|\theta) = p(y|w,\theta)p(w)$ with respect to the latent variable W and is computed in § 3.9. Alternatively, we could proceed directly to an Expectation-Maximization (EM) algorithm by using a variational approximation in the lower bound of the log-likelihood function. As it is shown in § 3.9, computing the marginal data distribution leads to difficult derivatives which hinder the maximization of the likelihood function. A variational EM algorithm that bypasses these difficulties is established in § 3.4. For completeness, the computations of the marginal data distribution and its likelihood function and the posterior distribution of the latent variables for the Gaussian model are included in § 3.9 and § 3.10, respectively.

Remark. In case where the integration is not amenable to variational approximations, numerical

simulations such as Markov chain Monte Carlo (MCMC) methods may be used to fit the model parameters. However, such methods require the specification of additional probability distributions on the piecewise affine mapping parameters.

Though the variational approximation leads to tractable expressions for the variational lower bound, it still involves integration of Gaussian functions over half spaces. We therefore proceed with a guide to these integrals.

3.3 Integration of Gaussian Functions

Consider the problem of computing the integral

$$Q_+ := \int_{\Omega_+} \exp\left(-\frac{1}{2}(w-m)^T C^{-1}(w-m)\right) dw, \tag{3.8}$$

where

$$\Omega_+ := \{w \in \mathbb{R}^q : v^T w + \beta \geq 0\}, \tag{3.9}$$

C^{-1} is symmetric positive definite (see Definition 0.4.2), and v is the normal vector of a hyperplane in $\mathbb{R}^q$. In order to complete this integration, we first need to transform the integral into something easier to deal with. In its simpler form, we will be integrating Gaussian functions on $\mathbb{R}^{q-1}$ and computing an error function on a subset of $\mathbb{R}$ (see Definition 0.3.2 for definitions of the error function and complementary error function).

First, a translation. Let $z = w - m$, $dz = dw$, and denote

$$\Omega_z := \left\{z \in \mathbb{R}^q : v^T z + (v^T m + \beta) \geq 0\right\}. \tag{3.10}$$

Then

$$Q_+ = \int_{\Omega_z} \exp\left(-\frac{1}{2} z^T C^{-1} z\right) dz. \tag{3.11}$$

Next, rotation and scaling. Since C^{-1} is symmetric positive definite, it admits a decomposition of the form $C^{-1} = U^T D^{-1} U$, where U is an orthogonal matrix and D is a diagonal matrix with strictly

positive elements on its diagonal. Now perform the following substitutions:

$$u = \sqrt{D^{-1}}Uz, \tag{3.12}$$

$$\nu_u = \sqrt{D^{-1}}U\nu, \tag{3.13}$$

$$du = |\sqrt{D}|^{-1}|U|dz = |\sqrt{D}|^{-1}dz, \tag{3.14}$$

and denote

$$\Omega_u := \left\{u \in \mathbb{R}^q : \nu_u^T u + (\nu^T m + \beta) \geq 0\right\} \tag{3.15}$$

to obtain

$$Q_+ = |\sqrt{D}| \int_{\Omega_u} \exp\left(-\frac{1}{2}u^T u\right) du. \tag{3.16}$$

Take $\boldsymbol{e}_q$, the q^{th} standard basis vector for $\mathbb{R}^q$, and find an orthogonal matrix V that represents the projection of ν_u onto $\boldsymbol{e}_q$. That is, find V such that

$$V\nu_u = ||\nu_u||\boldsymbol{e}_q, \tag{3.17}$$

and let $x = Vu$, $dx = |V|du = du$. Denote

$$\begin{aligned}\Omega_x &:= \left\{x \in \mathbb{R}^q : x^T V\nu_u + (\nu^T m + \beta) \geq 0\right\} \\ &= \left\{x \in \mathbb{R}^q : x^T(||\nu_u||\boldsymbol{e}_q) + (\nu^T m + \beta) \geq 0\right\} \\ &= \left\{x \in \mathbb{R}^q : ||\nu_u||x_q + (\nu^T m + \beta) \geq 0\right\},\end{aligned} \tag{3.18}$$

where x_q is the q^{th} element of x. Then:

$$Q_+ = |\sqrt{D}| \int_{\Omega_x} \exp\left(-\frac{1}{2}x^T x\right) dx. \tag{3.19}$$

We now let δ be the point along $\boldsymbol{e}_q$ where $||\nu_u||x_q + (\nu^T m + \beta) \geq 0$, that is,

$$\delta = -\frac{(\nu^T m + \beta)}{||\nu_u||}. \tag{3.20}$$

Now, $\Omega_x = \mathbb{R}^{q-1} \times [\delta, \infty)$. With these transformations in place, Q_+ is now a Gaussian integral over $\mathbb{R}^{q-1}$ multiplied by the complementary error function evaluated over a subset of the real numbers.

In other words,

$$Q_+ = |\sqrt{D}| \int_\delta^\infty \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \exp\left(-\frac{(x_1^2 + \cdots + x_{q-1}^2 + x_q^2)}{2}\right) dx_1 \cdots dx_{q-1} dx_q$$
$$= |\sqrt{D}|(2\pi)^{(1-q)/2} \sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\frac{\delta}{\sqrt{2}}\right). \tag{3.21}$$

We want to avoid calculating ν_u, so we will back-substitute the rotation and scaling to put δ in terms of ν. Since U is an orthogonal matrix,

$$\begin{aligned} ||\nu_u|| &= \left\| \sqrt{D^{-1}} U \nu \right\| \\ &= \left(\left\langle \sqrt{D^{-1}} U \nu, \sqrt{D^{-1}} U \nu \right\rangle \right)^{1/2} \\ &= \left(\left\langle \nu, U^T D^{-1} U \nu \right\rangle \right)^{1/2} \\ &= \left(\left\langle \nu, C^{-1} \nu \right\rangle \right)^{1/2} \\ &= \left(\nu^T C^{-1} \nu \right)^{1/2}. \end{aligned} \tag{3.22}$$

Now,

$$\delta = -\frac{(\nu^T m + \beta)}{\sqrt{\nu^T C^{-1} \nu}}. \tag{3.23}$$

Note that integration over the complement of Ω_+, given by

$$\Omega_- := \{w \in \mathbb{R}^q : \nu^T w + \beta < 0\}, \tag{3.24}$$

would follow in almost the exact same manner, but with the error function in place of the complementary error function:

$$Q_- := |\sqrt{D}| \int_{-\infty}^\delta \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \exp\left(-\frac{(x_1^2 + \cdots + x_{q-1}^2 + x_q^2)}{2}\right) dx_1 \cdots dx_{q-1} dx_q$$
$$= |\sqrt{D}|(2\pi)^{(1-q)/2} \sqrt{\frac{\pi}{2}} \left[\operatorname{erf}\left(\frac{\delta}{\sqrt{2}}\right) + 1\right]. \tag{3.25}$$

We can use this result to integrate other functions over the half spaces Ω_+ and Ω_-. For example, we may compute the integral

$$\int_{\Omega_+} w \exp\left(-\frac{1}{2}(w-m)^T C^{-1} (w-m)\right) dw. \tag{3.26}$$

Note that after performing the first translation as above, the integral becomes

$$\int_{\Omega_z} (z+m) \exp\left(-\frac{1}{2} z^T C^{-1} z\right) dz. \tag{3.27}$$

When we eventually reduce the integrand to the product of univariate Gaussians over $\mathbb{R}^{q-1}$, the integrand $z \exp\left(-\frac{1}{2} z^T C^{-1} z\right)$ is an odd function on symmetric intervals and therefore integrates to zero. Thus, the result is

$$\int_{\Omega_+} w \exp\left(-\frac{1}{2}(w-m)^T C^{-1}(w-m)\right) dw = m|\sqrt{D}|(2\pi)^{(1-q)/2}\sqrt{\frac{\pi}{2}}\operatorname{erfc}\left(\frac{\delta}{\sqrt{2}}\right), \tag{3.28}$$

with an analogous result over Ω_- involving the error function.

We postpone the details of the integration of Gaussian quadratic forms

$$w^T w \exp\left(-\frac{1}{2}(w-m)^T C^{-1}(w-m)\right) \tag{3.29}$$

to Appendix A.3.

3.4 An EM Algorithm for the Variational Lower Bound

As outlined later in § 3.9, directly maximizing the log-likelihood of the marginal data distribution leads to computational difficulties. Therefore, we will maximize the variational lower bound (VLB), given by $\mathcal{L}(q,\theta)$ in Theorem 3, restated here:

$$\mathcal{L}(q,\theta) = \int q(w) \log\left(\frac{p(y|w,\theta)p(w)}{q(w)}\right) dw, \tag{3.30}$$

where we have written $p(y,w|\theta) = p(y|w,\theta)p(w)$ for convenience.

Because the decomposition in (3.30) holds for any distribution over W, we will approximate the posterior conditional distribution $p(W|Y)$ with a product distribution that easily factors, as outlined in § 1.3.1. Our choice of approximating distribution is a product of Gaussians, the same approximation that was used in the Poisson PCA model of [16].

Suppose that $\tilde{p}_i(w_i) = \mathcal{N}(m_i, \Sigma_i)$ is a Gaussian distribution with mean vector $m_i \in \mathbb{R}^q$ and covariance matrix Σ_i. For ease of later calculations, we will say that Σ_i is a diagonal matrix with

strictly positive elements on its diagonal. That is, $\Sigma_i = \text{diag}(s_i \odot s_i)$, with $s_i \in \mathbb{R}_+^q$. We then approximate the posterior distribution $p(W|Y)$ with the product distribution $\tilde{p} = \otimes_{i=1}^N \tilde{p}_i$. Using this factored product distribution allows us to rewrite the VLB as

$$\begin{aligned}
\mathcal{L}(\tilde{p}, \theta) &= \int \tilde{p}(w) \log\left(\frac{p(y|w,\theta)p(w)}{\tilde{p}(w)}\right) dw \\
&= \int \prod_{i=1}^{N} \tilde{p}_i(w_i) \log\left(\frac{p(y_i|w_i,\theta)p(w_i)}{\tilde{p}_i(w_i)}\right) dw_i \\
&= \sum_{i=1}^{N} \int \tilde{p}_i(w_i) \left[\log p(w_i) + \log p(y_i|w_i,\theta) - \log \tilde{p}_i(w_i)\right] dw_i. \qquad (3.31)
\end{aligned}$$

Our choice of product distribution simplifies the VLB, but it also creates two sets of parameters to estimate: the model parameters given by σ^2, B_1, B_2, μ_1, and μ_2, and the variational parameters m_i and Σ_i, $1 \leq i \leq N$.

Applying the variational approximation of the posterior distribution to our product of Gaussian distributions leads to three integrals:

$$A := \int \tilde{p}_i(w_i) \log p(w_i) dw_i, \qquad (3.32)$$

$$B := \int \tilde{p}_i(w_i) \log p(y_i|w_i,\theta) dw_i, \qquad (3.33)$$

$$C := \int \tilde{p}_i(w_i) \log \tilde{p}_i(w_i) dw_i. \qquad (3.34)$$

The integrals defined by (3.32) and (3.34) are relatively simple to compute.

$$\begin{aligned}
A &= \int \log\left((2\pi)^{-\frac{q}{2}} \exp\left(-\frac{1}{2} w_i^T w_i\right)\right) \tilde{p}_i(w_i) dw_i \\
&= \int \left(-\frac{q}{2}\log(2\pi) - \frac{1}{2} w_i^T w_i\right) \tilde{p}_i(w_i) dw_i \\
&= -\frac{q}{2}\log(2\pi) - E_{\tilde{p}_i}\left(w_i^T w_i\right) \\
&= -\frac{1}{2}\left(q\log(2\pi) + \text{Tr}\left(\Sigma_i\right) + m_i^T m_i\right). \qquad (3.35)
\end{aligned}$$

$$
\begin{aligned}
C &= \int \log\left((2\pi)^{-\frac{q}{2}}|\Sigma_i|\exp\left(-\frac{1}{2}(w_i-m_i)^T\Sigma_i^{-1}(w_i-m_i)\right)\right)\tilde{p}_i(w_i)dw_i \\
&= \int\left(-\frac{q}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma_i|-\frac{1}{2}(w_i-m_i)^T\Sigma_i^{-1}(w_i-m_i)\right)\tilde{p}_i(w_i)dw_i \\
&= -\frac{q}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma_i|-\frac{1}{2}E_{\tilde{p}_i}\left((W_i-m_i)^T\Sigma_i^{-1}(W_i-m_i)\right) \\
&= -\frac{q}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma_i|-\frac{1}{2}\mathrm{Tr}\left(\Sigma_i^{-1}\Sigma_i\right) \\
&= -\frac{q}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma_i|-\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{I}_q\right) \\
&= -\frac{1}{2}\left(q\log(2\pi)+\log|\Sigma_i|+q\right).
\end{aligned} \tag{3.36}
$$

We next calculate the integral (3.33).

$$
\begin{aligned}
B &= \int \log\left((2\pi\sigma^2)^{-\frac{p}{2}}\exp\left(-\frac{1}{2\sigma^2}||y_i-f(w_i,\theta)||^2\right)\right)\tilde{p}_i(w_i)dw_i \\
&= \int\left(-\frac{p}{2}\log\left(2\pi\sigma^2\right)+\log\left(\exp\left(-\frac{1}{2\sigma^2}||y_i-f(w_i,\theta)||^2\right)\right)\right)\tilde{p}_i(w_i)dw_i \\
&= -\frac{p}{2}\log\left(2\pi\sigma^2\right)+ \\
&\int_{\Omega_+}\log\left(\exp\left(-\frac{1}{2\sigma^2}||y_i-\mu_1-B_1w_i||^2\right)\right)\tilde{p}_i(w_i)dw_i+ \\
&\quad\int_{\Omega_-}\log\left(\exp\left(-\frac{1}{2\sigma^2}||y_i-\mu_2-B_2w_i||^2\right)\right)\tilde{p}_i(w_i)dw_i \\
&= -\frac{p}{2}\log\left(2\pi\sigma^2\right)- \\
&\frac{1}{2\sigma^2}\int_{\Omega_+}\left((y_i-\mu_1)^T(y_i-\mu_1)-2(y_i-\mu_1)^TB_1w_i+w_i^TB_1^TB_1w_i\right)\tilde{p}_i(w_i)dw_i- \\
&\quad\frac{1}{2\sigma^2}\int_{\Omega_-}\left((y_i-\mu_2)^T(y_i-\mu_2)-2(y_i-\mu_2)^TB_2w_i+w_i^TB_2^TB_2w_i\tilde{p}_i(w_i)\right)dw_i.
\end{aligned} \tag{3.37}
$$

Using a variation of the computations in § 3.3 and the integration of Gaussian quadratic forms from § A.1, we obtain:

$$
\begin{aligned}
B = -\frac{p}{2}\log\left(2\pi\sigma^2\right) &- \frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2}\left[\delta_i\exp\left(-\frac{\delta_i^2}{2}\right)\left(\left[B_1^TB_1\Sigma_i\right]_{qq}-\left[B_2^TB_2\Sigma_i\right]_{qq}\right)+\right. \\
&\sqrt{\frac{\pi}{2}}\mathrm{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right)\left(\mathrm{Tr}\left(B_1^TB_1\Sigma_i\right)+m_i^TB_1^TB_1m_i+(y_i-\mu_1)^T\left[(y_i-\mu_1)-2B_1m_i\right]\right)+ \\
&\left.\sqrt{\frac{\pi}{2}}\left(\mathrm{erf}\left(\frac{\delta_i}{\sqrt{2}}\right)+1\right)\left(\mathrm{Tr}\left(B_2^TB_2\Sigma_i\right)+m_i^TB_2^TB_2m_i+(y_i-\mu_2)^T\left[(y_i-\mu_2)-2B_2m_i\right]\right)\right],
\end{aligned} \tag{3.38}
$$

where $\delta_i = -\frac{[m_i]_q}{\sqrt{[\Sigma_i]_q}}$, and $[\Sigma_i]_q$ is the q^{th} element on the main diagonal of Σ_i. Note that the individual terms in (3.38) are not calculated over the entire data set. To make this distinction clearer in the VLB, we define two index sets:

$$I_1 := \{1 \leq i \leq N : \delta_i \geq 0\}, \tag{3.39}$$

$$I_2 := \{1 \leq i \leq N : \delta_i < 0\}. \tag{3.40}$$

With the integrals (3.32), (3.33), and (3.34) computed, the VLB for the log-likelihood of our piecewise affine model is given by:

$$\begin{aligned}
\mathcal{L}(\tilde{p}, \theta) &= \sum_{i=1}^{N} A + B - C \\
&= \sum_{i=1}^{N} -\frac{1}{2}\left(q\log(2\pi) + \operatorname{Tr}(\Sigma_i) + m_i^T m_i + p\log\left(2\pi\sigma^2\right)\right) - \\
&\qquad \frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2} \sum_{i\in I_1} \left[\delta_i \exp\left(-\frac{\delta_i^2}{2}\right)\left[B_1^T B_1 \Sigma_i\right]_{qq} + \right. \\
&\qquad \left. \sqrt{\frac{\pi}{2}}\operatorname{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right)\left(\operatorname{Tr}\left(B_1^T B_1 \Sigma_i\right) + m_i^T B_1^T B_1 m_i + (y_i - \mu_1)^T\left[(y_i - \mu_1) - 2B_1 m_i\right]\right)\right] - \\
&\qquad \frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2} \sum_{i\in I_2} \left[-\delta_i \exp\left(-\frac{\delta_i^2}{2}\right)\left[B_2^T B_2 \Sigma_i\right]_{qq} + \right. \\
&\qquad \left. \sqrt{\frac{\pi}{2}}\left(\operatorname{erf}\left(\frac{\delta_i}{\sqrt{2}}\right) + 1\right)\left(\operatorname{Tr}\left(B_2^T B_2 \Sigma_i\right) + m_i^T B_2^T B_2 m_i + (y_i - \mu_2)^T\left[(y_i - \mu_2) - 2B_2 m_i\right]\right)\right] + \\
&\qquad \frac{1}{2}\sum_{i=1}^{N}\left(q\log(2\pi) + \log|\Sigma_i| + q\right).
\end{aligned}$$

Gathering and factoring terms gives the VLB in its final form:

$$\mathcal{L}(\tilde{p},\theta) = \frac{1}{2}\Bigg[N\left(q - p\log(2\pi\sigma^2)\right) + \sum_{i=1}^{N}\left[\log|\Sigma_i| - \mathrm{Tr}\left(\Sigma_i\right) - m_i^T m_i\right] - \frac{(2\pi)^{-\frac{1}{2}}}{\sigma^2}\Bigg[\sum_{i\in I_1}\delta_i \exp\left(-\frac{\delta_i^2}{2}\right)\left[B_1^T B_1 \Sigma_i\right]_{qq} + \sqrt{\frac{\pi}{2}}\mathrm{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right)\left(\mathrm{Tr}\left(B_1^T B_1 \Sigma_i\right) + m_i^T B_1^T B_1 m_i + (y_i-\mu_1)^T\left[(y_i-\mu_1) - 2B_1 m_i\right]\right) + \sum_{i\in I_2} -\delta_i \exp\left(-\frac{\delta_i^2}{2}\right)\left[B_2^T B_2 \Sigma_i\right]_{qq} + \sqrt{\frac{\pi}{2}}\left(\mathrm{erf}\left(\frac{\delta_i}{\sqrt{2}}\right)+1\right)\left(\mathrm{Tr}\left(B_2^T B_2 \Sigma_i\right) + m_i^T B_2^T B_2 m_i + (y_i-\mu_2)^T\left[(y_i-\mu_2) - 2B_2 m_i\right]\right)\Bigg]\Bigg]. \tag{3.41}$$

Theorem 8. *The VLB given by* (3.41) *is convex in the model parameters* $\{\sigma^2, \mu_1, \mu_2, B_1, B_2\}$ *with the variational parameters* $\{m_i, \Sigma_i\}$ *fixed, and vice-versa.*

Proof.

For convenience, we restate the VLB from (1.11):

$$\begin{aligned}\mathcal{L}(\tilde{p},\theta) &= \sum_{i=1}^{N}\int \tilde{p}_i(w_i)\left(\log p(y_i|w_i,\theta) + \log p(w_i) - \log\tilde{p}_i(w_i)\right)dw_i \\ &= \sum_{i=1}^{N}\int \tilde{p}_i(w_i)\log p(y_i|w_i,\theta)dw_i + \int \tilde{p}_i(w_i)\log p(w_i)dw_i - \int \tilde{p}_i(w_i)\log\tilde{p}_i(w_i)dw_i\end{aligned} \tag{3.42}$$

It suffices to prove convexity for each of the individual integrands in (3.42). The three integrands are:

$$\log p(y_i|w_i,\theta) = -\frac{1}{2}\left(p\log(2\pi\sigma^2) + \sigma^{-2}||y_i - f(w_i,\theta)||^2\right) \tag{3.43}$$

$$\log p(w_i) = -\frac{1}{2}\left(q\log(2\pi) + w_i^T w_i\right) \tag{3.44}$$

$$\log\tilde{p}_i(w_i) = -\frac{1}{2}\left(q\log(2\pi) + \log|\Sigma_i| + (w_i - m_i)^T\Sigma_i^{-1}(w_i - m_i)\right). \tag{3.45}$$

First, fix the model parameters $\{\sigma^2, \mu_1, \mu_2, B_1, B_2\}$. We will show that (3.43)-(3.45) are convex in the variational parameters $\{m_i, \Sigma_i\}$. Note that the functions (3.43) and (3.44) are constant with respect to the variational parameters, hence convex. Since (3.45) is a quadratic form, then by Lemma A.13 it is convex in both m_i and Σ_i.

Next, fix $\{m_i, \Sigma_i\}$. Since (3.44) and (3.45) are constant with respect to the variational parameters, they are both convex. The remaining term, (3.44), is quadratic in the model parameters and thus convex. □

We maximize the VLB using maximum likelihood estimation. For this we need to calculate derivatives of the VLB with respect to both the model parameters and the variational parameters and find their respective critical points. Fortunately, differentiating with respect to the model parameters leads to closed form solutions for the corresponding critical points.

3.4.1 Optimizing the Model Parameters - Gaussian Observations

Taking the derivative of the VLB with respect to the isotropic observational covariance gives:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{Np}{2\sigma^2} + \frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^4}\left[\sum_{i\in I_1}(\cdots) + \sum_{i\in I_2}(\cdots)\right], \tag{3.46}$$

where the sums over I_1 and I_2 are computed as in (3.41). The critical point is found at:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = 0 \iff \sigma^2 = \frac{(2\pi)^{-\frac{1}{2}}}{Np}\left[\sum_{i\in I_1}(\cdots) + \sum_{i\in I_2}(\cdots)\right]. \tag{3.47}$$

For the derivatives with respect to the model means μ_1 and μ_2, we use identities (69) and (86) from [60].

$$\frac{\partial \mathcal{L}}{\partial \mu_1} = -\frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2}\left[\sum_{i\in I_1}\sqrt{\frac{\pi}{2}}\operatorname{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right)[-2(y_i-\mu_1)+2B_1 m_i]\right], \tag{3.48}$$

and, similarly,

$$\frac{\partial \mathcal{L}}{\partial \mu_2} = -\frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2}\left[\sum_{i\in I_2}\sqrt{\frac{\pi}{2}}\left(\operatorname{erf}\left(\frac{\delta_i}{\sqrt{2}}\right)+1\right)[-2(y_i-\mu_2)+2B_2 m_i]\right]. \tag{3.49}$$

Setting these two derivatives to zero and solving for μ_1, μ_2 gives their respective critical points:

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \mu_1} = 0 \\
&\qquad \Longleftrightarrow \\
&\mu_1 = \left[\sum_{i \in I_1} \operatorname{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right)\right]^{-1} \sum_{i \in I_1} \operatorname{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right)(y_i - B_1 m_i),
\end{aligned} \tag{3.50}$$

and

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \mu_2} = 0 \\
&\qquad \Longleftrightarrow \\
&\mu_2 = \left[\sum_{i \in I_2} \left(\operatorname{erf}\left(\frac{\delta_i}{\sqrt{2}}\right) + 1\right)\right]^{-1} \sum_{i \in I_2} \left(\operatorname{erf}\left(\frac{\delta_i}{\sqrt{2}}\right) + 1\right)(y_i - B_2 m_i),
\end{aligned} \tag{3.51}$$

where the division by the error function is possible as long as $\delta_i = -[m_i]_q / \sqrt{[\Sigma_i]_q}$ does not tend to $\pm\infty$. Therefore, it is necessary that the variational means are bounded and that the diagonal elements of Σ_i are strictly positive. We will return to this necessary condition when we calculate the derivatives of the VLB with respect to the variational parameters.

Remark. The mean parameters μ_1 and μ_2 are only fitted when the dataset is not centered or when the observations come from two disjoint subsets.

We now calculate the derivatives of the VLB with respect to the linear transformations B_1 and B_2.

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial B_1} = -\frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2} \Bigg(&\sum_{i \in I_1} \delta_i \exp\left(-\frac{\delta_i^2}{2}\right) \frac{\partial \left[B_1^T B_1 \Sigma_i\right]_{qq}}{\partial B_1} + \\
&\sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right) \left[2B_1\Sigma_i + 2B_1 m_i m_i^T - 2(y_i - \mu_1) m_i^T\right] \Bigg).
\end{aligned} \tag{3.52}$$

In order to compute $\frac{\partial \left[B_1^T B_1 \Sigma_i\right]_{qq}}{\partial B_1}$, we define $q \times q$ real matrices $\{M_n\}_{n=1}^q$, with

$$\begin{cases}
[M_n]_{jk} = 1, & j = k = n, \\
[M_n]_{jk} = 0, & \text{otherwise.}
\end{cases} \tag{3.53}$$

Then, we decompose Σ_i into a sum of 'basis' matrices:

$$\Sigma_i = \sum_{n=1}^{q} [\Sigma_i]_n M_n, \tag{3.54}$$

where $[\Sigma_i]_n$ is the n^{th} term along the diagonal. In this notation,

$$[B_1^T B_1 \Sigma_i]_q q = \text{Tr}\left(B_1^T B_1 [\Sigma_i]_q M_q\right) = [\Sigma_i]_q \text{Tr}\left(B_1^T B_1 M_q\right). \tag{3.55}$$

Using the form (3.55) and identity (113) from [60], we compute the derivative:

$$\begin{aligned} \frac{\partial \left[B_1^T B_1 \Sigma_i\right]_{qq}}{\partial B_1} &= [\Sigma_i]_q \frac{\partial \text{Tr}\left(B_1^T B_1 M_q\right)}{\partial B_1} \\ &= 2[\Sigma_i]_q B_1 M_q. \end{aligned} \tag{3.56}$$

Therefore,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial B_1} = -\frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2} \Bigg(&\sum_{i \in I_1} 2\delta_i \exp\left(-\frac{\delta_i^2}{2}\right) [\Sigma_i]_q B_1 M_q + \\ &\sqrt{\frac{\pi}{2}} \text{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right) \left[2B_1\Sigma_i + 2B_1 m_i m_i^T - 2(y_i - \mu_1) m_i^T\right] \Bigg). \end{aligned} \tag{3.57}$$

Similarly,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial B_2} = -\frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2} \Bigg(&\sum_{i \in I_2} -2\delta_i \exp\left(-\frac{\delta_i^2}{2}\right) [\Sigma_i]_q B_2 M_q + \\ &\sqrt{\frac{\pi}{2}} \left(\text{erf}\left(\frac{\delta_i}{\sqrt{2}}\right) + 1\right) \left[2B_2\Sigma_i + 2B_2 m_i m_i^T - 2(y_i - \mu_2) m_i^T\right] \Bigg). \end{aligned} \tag{3.58}$$

The critical points are found at:

$$\frac{\partial \mathcal{L}}{\partial B_1} = 0$$

$$\iff$$

$$B_1 = \left[\sum_{i \in I_1} (y_i - \mu_1) m_i^T\right] \left[\sum_{i \in I_1}^{N} \sqrt{\frac{\pi}{2}} \text{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right) (\Sigma_i + m_i m_i^T) + \delta_i \cdot e^{-\frac{\delta_i^2}{2}} [\Sigma_i]_q M_q\right]^{-1}, \tag{3.59}$$

and

$$\frac{\partial \mathcal{L}}{\partial B_2} = 0$$

$$\iff$$

$$B_2 = \left[\sum_{i\in I_2}(y_i-\mu_2)m_i^T\right]\left[\sum_{i\in I_2}^{N}\sqrt{\frac{\pi}{2}}\left(\operatorname{erf}\left(\frac{\delta_i}{\sqrt{2}}\right)+1\right)(\Sigma_i+m_i m_i^T)-\delta_i\cdot e^{-\frac{\delta_i^2}{2}}[\Sigma_i]_q M_q\right]^{-1}. \tag{3.60}$$

3.4.2 Optimizing the Variational Parameters - Gaussian Observations

We now compute the derivatives of the VLB with respect to the variational parameters (m_i, Σ_i) over the index sets I_1 and I_2.

$$\left.\frac{\partial \mathcal{L}}{\partial m_i}\right|_{I_1} = -m_i + \frac{1}{2\sigma^2}\left((2\pi)^{-\frac{1}{2}}[\Sigma_i]_q^{-\frac{1}{2}}\exp\left(-\frac{\delta_i^2}{2}\right)\left[[B_1^T B_1\Sigma_i]_{qq}(1-\delta_i^2)-\operatorname{Tr}\left(B_1^T B_1\Sigma_i\right)-\right.\right.$$
$$m_i^T B_1^T B_1 m_i-(y_i-\mu_1)^T\left[(y_i-\mu_1)-2B_1 m_i\right]\Big]\boldsymbol{e}_q+$$
$$\left.\left.\operatorname{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right)\left[B_1^T B_1 m_i - B_1^T(y_i-\mu_1)\right]\right), \tag{3.61}$$

and

$$\left.\frac{\partial \mathcal{L}}{\partial m_i}\right|_{I_2} = -m_i - \frac{1}{2\sigma^2}\left((2\pi)^{-\frac{1}{2}}[\Sigma_i]_q^{-\frac{1}{2}}\exp\left(-\frac{\delta_i^2}{2}\right)\left[[B_2^T B_2\Sigma_i]_{qq}(1-\delta_i^2)-\operatorname{Tr}\left(B_2^T B_2\Sigma_i\right)-\right.\right.$$
$$m_i^T B_2^T B_2 m_i-(y_i-\mu_2)^T\left[(y_i-\mu_2)-2B_2 m_i\right]\Big]\boldsymbol{e}_q+$$
$$\left.\left.\left(\operatorname{erf}\left(\frac{\delta_i}{\sqrt{2}}\right)+1\right)\left[B_2^T B_2 m_i - B_2^T(y_i-\mu_2)\right]\right), \tag{3.62}$$

where $\mathbf{e}_q$ is the q^{th} standard basis vector of $\mathbb{R}^q$. From these derivatives, we can write the more convenient expression:

$$\left.\frac{\partial \mathcal{L}}{\partial m_i}\right|_{I_j} = -m_i + \frac{(-1)^{j-1}}{2\sigma^2}\left[c_j\boldsymbol{e}_q + f_j(\delta_i)\left(B_j^T B_j m_i - B_j^T(y_i-\mu_j)\right)\right], \tag{3.63}$$

where

$$f_j(x) := \begin{cases}\operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right), & j=1,\\ \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)+1, & j=2,\end{cases} \tag{3.64}$$

and

$$c_j := (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{\delta_i^2}{2}\right) [\Sigma_i]_q^{-\frac{1}{2}} \left([B_j^T B_j \Sigma_i]_{qq}(1-\delta_i^2) - \mathrm{Tr}\left(B_j^T B_j \Sigma_i\right) - \right. \tag{3.65}$$
$$\left. m_i^T B_j^T B_j m_i - (y_i - \mu_j)^T \left[(y_i - \mu_j) - 2B_j m_i\right]\right).$$

We finally compute the derivatives of the VLB with respect to the latent covariance matrices.

$$\left.\frac{\partial \mathcal{L}}{\partial \Sigma_i}\right|_{I_1} = \frac{1}{2}\left[\Sigma_i^{-1} - I - \frac{(2\pi)^{-\frac{1}{2}}}{\sigma^2}\left(\exp\left(-\frac{\delta_i^2}{2}\right)\left[\delta_i B_1^T B_1 \boldsymbol{e}_q \boldsymbol{e}_q^T + \right.\right.\right. \tag{3.66}$$
$$\left.\left.\left.\frac{\partial \delta_i}{\partial \Sigma_i}\left([B_1^T B_1 \Sigma_i]_{qq}(1-\delta_i^2) - \mathrm{Tr}\left(B_1^T B_1 \Sigma_i\right)\right)\right] + \sqrt{\frac{\pi}{2}}\mathrm{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right) B_1^T B_1\right)\right],$$

and

$$\left.\frac{\partial \mathcal{L}}{\partial \Sigma_i}\right|_{I_2} = \frac{1}{2}\left[\Sigma_i^{-1} - I - \frac{(2\pi)^{-\frac{1}{2}}}{\sigma^2}\left(-\exp\left(-\frac{\delta_i^2}{2}\right)\left[\delta_i B_2^T B_2 \boldsymbol{e}_q \boldsymbol{e}_q^T + \right.\right.\right. \tag{3.67}$$
$$\left.\left.\left.\frac{\partial \delta_i}{\partial \Sigma_i}\left([B_2^T B_2 \Sigma_i]_{qq}(-1+\delta_i^2) + \mathrm{Tr}\left(B_2^T B_2 \Sigma_i\right)\right)\right] + \sqrt{\frac{\pi}{2}}\left(\mathrm{erf}\left(\frac{\delta_i}{\sqrt{2}}\right)+1\right) B_2^T B_2\right)\right],$$

with

$$\frac{\partial \delta_i}{\partial \Sigma_i} = -\frac{1}{2}\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & [m_i]_q [s_i]_q^{-1} \end{bmatrix}. \tag{3.68}$$

From (3.66) and (3.67), we can write the general form for the derivative of the VLB with respect to the variational parameter Σ_i:

$$\left.\frac{\partial \mathcal{L}}{\partial \Sigma_i}\right|_{I_j} = \frac{1}{2}\left[\Sigma_i^{-1} - I - \frac{(2\pi)^{-\frac{1}{2}}}{\sigma^2}\left((-1)^{j-1}\exp\left(-\frac{\delta_i^2}{2}\right)\left[D_j + \frac{\partial \delta_i}{\partial \Sigma_i} d_j\right] + g_j(\delta_i)\right)\right], \tag{3.69}$$

where

$$D_j := \delta_i B_j^T B_j \boldsymbol{e}_q \boldsymbol{e}_q^T, \tag{3.70}$$

$$d_j := [B_j^T B_j \Sigma_i]_{qq}(1-\delta_i^2) - \mathrm{Tr}\left(B_j^T B_j \Sigma_i\right), \tag{3.71}$$

and

$$g_j(x) := \sqrt{\frac{\pi}{2}} B_j^T B_j \begin{cases} \mathrm{erfc}\left(\frac{\delta_i}{\sqrt{2}}\right), & j = 1, \\ \left(\mathrm{erf}\left(\frac{\delta_i}{\sqrt{2}}\right)+1\right), & j = 2. \end{cases} \tag{3.72}$$

Due to the presence of both the exponential and error functions in (3.63) and (3.69), these derivatives are highly nonlinear and we cannot find a closed form for their respective critical points. Therefore, in practical implementations of piecewise PCA, we will resort to iterative solutions such as gradient descent (see Definition 0.2.1) to maximize the VLB with respect to m_i and Σ_i over the index sets I_1 and I_2. In the implementation, we include the condition that the diagonals of Σ_i only contain strictly positive elements. We have experimented with both bounded and unbounded optimization of the variational means, and have found that choosing starting values based on Multivariate Adaptive Regression Splines (MARS) models ([25]) leads to successful unbounded optimization.

3.5 Multiple Partitions in the Latent Space

We may generalize the VLB from § 3.4 to account for more partitions in the latent space. Suppose that we partition the latent space using two parallel hyperplanes given in intercept form by

$$\begin{aligned} w_q &= 0, \\ w_q &= 1, \end{aligned} \tag{3.73}$$

where w_q is the q^{th} element of the q-dimensional vector w. In other words, the latent space is partitioned into the three sets defined by:

$$\begin{aligned} \Omega_1 &:= \{w \in \mathbb{R}^q : w_q > 1\}, \\ \Omega_2 &:= \{w \in \mathbb{R}^q : 0 < w_q \leq 1\}, \\ \Omega_3 &:= \{w \in \mathbb{R}^q : w_q \leq 0\}. \end{aligned} \tag{3.74}$$

Define the piecewise function relating data points Y_i to latent points W_i:

$$y_i = f(w_i, \theta) := \begin{cases} \mu_1 + B_1 w_i, & w_i \in \Omega_1, \\ \mu_2 + B_2 w_i, & w_i \in \Omega_2, \\ \mu_3 + B_3 w_i, & w_i \in \Omega_3. \end{cases} \tag{3.75}$$

As in § 3.2, we assume a Gaussian prior distribution $W_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and a Gaussian distribution with isotropic covariance matrix $\sigma^2 \boldsymbol{I}$ on the observations. We also use the same variational approximation as in § 3.4. Recall that in § 3.4, the VLB was composed of three integrals, A, B, and C. Of these three integrals, A and C depend only on the prior distribution in the latent space and the variational approximation, but not on the number of partitions in the latent space. Thus, we need only concern ourselves with computing the integral B for the latent space with three partitions. Recall that B was defined as

$$B := \int_{\mathbb{R}^q} \tilde{p}_i(w_i) \log p(y_i | w_i, \theta) dw_i. \tag{3.76}$$

Using the calculations from § 3.3, the above integral for the Gaussian model with three partitions in latent space is given by

$$\begin{aligned} B = &- \frac{p}{2} \log 2\pi\sigma^2 - \\ &\frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2} \left[\sqrt{\frac{\pi}{2}} \left(\operatorname{erfc}\left(\frac{\delta_i}{\sqrt{2}} \right) G_1 + \operatorname{erf}\left(\frac{\delta_i}{\sqrt{2}} \right) (G_2 - G_3) + \operatorname{erf}\left(\frac{\gamma_i}{\sqrt{2}} \right) G_2 + G_3 \right) + \right. \\ &\gamma_i \exp\left(-\frac{\gamma_i^2}{2} \right) \left[[B_1^T B_1 \Sigma_i]_{qq} - [B_2^T B_2 \Sigma_i]_{qq} \right] + \\ &\left. \delta_i \exp\left(-\frac{\delta_i^2}{2} \right) \left[[B_2^T B_2 \Sigma_i]_{qq} - [B_3^T B_3 \Sigma_i]_{qq} \right] \right], \end{aligned} \tag{3.77}$$

where

$$\begin{aligned} \gamma_i &:= -\frac{([m_i]_q + 1)}{\sqrt{[\Sigma_i]_q}}, \\ \delta_i &:= -\frac{[m_i]_q}{\sqrt{[\Sigma_i]_q}}, \end{aligned} \tag{3.78}$$

and for $1 \leq k \leq 3$,

$$G_k := \operatorname{Tr}\left(B_k^T B_k \Sigma_i \right) + m_i^T B_k^T B_k m_i + (y_i - \mu_k)^T (y_i - \mu_k) - 2(y_i - \mu_k)^T B_k m_i. \tag{3.79}$$

Suppose that we now divide the latent space into K partitions. Define the K partitions of the

latent space as

$$\begin{aligned}
\Omega_1 &:= \{w \in \mathbb{R}^q : w_q > K-1\},\\
&\vdots\\
\Omega_k &:= \{w \in \mathbb{R}^q : k-2 < w_q \leq k-1\},\\
&\vdots\\
\Omega_K &:= \{w \in \mathbb{R}^q : w_q \leq 0\},
\end{aligned} \tag{3.80}$$

and define the piecewise affine function

$$y_i = f(w_i,\theta) := \begin{cases}
\mu_1 + B_1 w_i, & w_i \in \Omega_1,\\
\vdots & \vdots\\
\mu_k + B_k w_i, & w_i \in \Omega_k,\\
\vdots & \vdots\\
\mu_K + B_K w_i, & w_i \in \Omega_K.
\end{cases} \tag{3.81}$$

Also, define the $K-1$ points of separation between the latent space partitions as

$$\delta_i^{[k]} := -\frac{[m_i]_q + (k-1)}{\sqrt{[\Sigma_i]_{qq}}}, \quad 1 \leq k \leq K-1. \tag{3.82}$$

The integral defined by B in the VLB for the Gaussian model with K partitions in latent space is then given by:

$$\begin{aligned}
B := &-\frac{p}{2}\log 2\pi\sigma^2 -\\
&\frac{(2\pi)^{-\frac{1}{2}}}{2\sigma^2}\left[\sum_{k=1}^{K-1}\delta_i^{[k]}\exp\left(-\frac{\left(\delta_i^{[k]}\right)^2}{2}\right)\left([B_k^T B_k \Sigma_i]_{qq} - [B_{k+1}^T B_{k+1}\Sigma_i]_{qq}\right) +\right.\\
&\left.\sqrt{\frac{\pi}{2}}\left[\operatorname{erfc}\left(\frac{\delta_i^{[1]}}{\sqrt{2}}\right) + \left(\operatorname{erf}\left(\frac{\delta_i^{[K-1]}}{\sqrt{2}}\right)+1\right)G_K + \sum_{k=2}^{K-1} G_k\left(\operatorname{erf}\left(\frac{\delta_i^{[k]}}{\sqrt{2}}\right) - \operatorname{erf}\left(\frac{\delta_i^{[k-1]}}{\sqrt{2}}\right)\right)\right]\right],
\end{aligned} \tag{3.83}$$

where, for $1 \leq k \leq K$,

$$G_k := \operatorname{Tr}\left(B_k^T B_k \Sigma_i\right) + m_i^T B_k^T B_k m_i + (y_i-\mu_k)^T(y_i-\mu_k) - 2(y_i-\mu_k)^T B_k m_i. \tag{3.84}$$

3.6 Piecewise PCA with Poisson Distributed Observations

We may also adapt piecewise PCA to observations with alternate NEF distributions. In this section we establish a model for performing PCA in a partitioned latent space while using a Poisson distribution on the observations. For this model, we again assume that the latent variables follow a Gaussian distribution with unit covariance. We then segment the latent space into two half-spaces as in § 3.2, where Ω_+ and Ω_- were defined.

Given the latent variables $\{W_i\}_{i=1}^N$, suppose that we assume independent Poisson distributions on the observations $\{Y_i\}_{i=1}^N$ with canonical parameter given by

$$f(w_i,\theta) := \begin{cases} \mu_1 + B_1 w_i, & w_i \in \Omega_+, \\ \mu_2 + B_2 w_i, & w_i \in \Omega_-, \end{cases} \tag{3.85}$$

where Ω_+ and Ω_- were defined in (3.3).

The Poisson piecewise PCA model can be summarized as:

$$\begin{aligned} \{W_i\}_{i=1}^N \subset \mathbb{R}^q, &\quad W_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q), \\ \{Y_i\}_{i=1}^N \subset \mathbb{R}^p, &\quad Y_i = f(W_i,\theta) + \varepsilon_i, \\ \{Y_i|W_i\}_{1\le i\le N} &\quad \text{independent}, \\ p(Y_i|W_i,\theta) &= h(Y_i)\exp\left(y_i^T f(W_i,\theta) - \psi(f(W_i,\theta))\right), \end{aligned} \tag{3.86}$$

where, for $\boldsymbol{x} \in \mathbb{R}^p$,

$$h(\boldsymbol{x}) = \exp\left(-\sum_{j=1}^{p}\sum_{k=2}^{x_j}\log k\right), \text{ and} \tag{3.87}$$

$$\psi(\boldsymbol{x}) = \sum_{j=1}^{p}\exp\left(x_j\right). \tag{3.88}$$

Remark. Since the prior distribution is Gaussian and the canonical parameter can be written as $\lambda = \log f(W_i,\theta)$, Poisson models like (3.86) and the Poisson PCA model in § 2.5.1 are often referred to as Poisson log-normal models. However, since our current model is not based on a diagonalization of the posterior covariance matrix, we do not use the covariance structure of the Poisson log-normal distribution as described in [3].

There are difficulties with computing the marginal distribution of the data, $p(Y_i|\theta)$, that will lead to an intractable log-likelihood function (see § 3.4 and § 3.9.1). Therefore, instead of integrating out the latent variables W_i under the joint distribution $p(Y,W)$, we proceed directly with a variational approximation of the log-likelihood function. As above when we assumed a Gaussian distribution on the observations, we approximate the posterior with a product of Gaussians $\tilde{r} = \otimes_{i=1}^{N} \tilde{r}_i$, where $\tilde{r}_i(w_i) = \mathcal{N}(n_i, \Phi_i)$, and Φ_i is diagonal with strictly positive entries. The VLB for the Poisson prior is then

$$\begin{aligned}
\mathcal{J}(\tilde{r},\theta) &= \sum_{i=1}^{N} \int \tilde{r}_i(w_i) \left[\log p(w_i) + \log p(y_i|w_i,\theta) - \log \tilde{r}_i(w_i)\right] dw_i \\
&= \sum_{i=1}^{N} \left[\int \tilde{r}_i(w_i) \log p(w_i) dw_i + \int \tilde{r}_i(w_i) \log p(y_i|w_i,\theta) dw_i - \int \tilde{r}_i(w_i) \log \tilde{r}_i(w_i) dw_i \right] \\
&= \sum_{i=1}^{N} A' + B' - C'. \qquad (3.89)
\end{aligned}$$

The computations for the integrals defined by A' and C' are exactly the same as those done for A and C in (3.35) and (3.36), respectively. That is,

$$A' = -\frac{1}{2}\left(q\log(2\pi) + \mathrm{Tr}\left(\Phi_i\right) + m_i^T m_i\right), \qquad (3.90)$$

and

$$C' = -\frac{1}{2}\left(q\log(2\pi) + \log|\Phi_i| + q\right). \qquad (3.91)$$

Therefore, we focus on the integral given by B'. We will use the results from § 3.3 to complete the

integration. We have

$$\begin{aligned}
B' &:= \int_{\mathbb{R}^q} \tilde{r}_i(w_i) \log p(y_i|w_i,\theta) dw_i \\
&= \left[\int_{\Omega_+} \tilde{r}_i(w_i) \log \left[h(y_i) \exp\left(y_i^T (\mu_1 + B_1 w_i) - \psi(\mu_1 + B_1 w_i) \right) \right] dw_i + \right. \\
&\qquad \left. \int_{\Omega_-} \tilde{r}_i(w_i) \log \left[h(y_i) \exp\left(y_i^T (\mu_2 + B_2 w_i) - \psi(\mu_2 + B_2 w_i) \right) \right] dw_i \right] \\
&= c_i \left[\log h(y_i) \int_{\mathbb{R}^q} \tilde{r}_i(w_i) dw_i + y_i^T \mu_1 \int_{\Omega_+} \tilde{r}_i(w_i) dw_i + y_i^T \mu_2 \int_{\Omega_-} \tilde{r}_i(w_i) dw_i + \right. \\
&\qquad \int_{\Omega_+} \tilde{r}_i(w_i) \left[y_i^T B_1 w_i - \psi(\mu_1 + B_1 w_i) \right] dw_i + \\
&\qquad \left. \int_{\Omega_-} \tilde{r}_i(w_i) \left[y_i^T B_2 w_i - \psi(\mu_1 + B_2 w_i) \right] dw_i \right]
\end{aligned} \tag{3.92}$$

where $c_i = (2\pi)^{-\frac{q}{2}} |\Phi_i|^{-\frac{1}{2}}$ and $\psi(\boldsymbol{x}) = \sum_{j=1}^{p} \exp\left(x_j\right)$, $\boldsymbol{x} \in \mathbb{R}^p$.

Since $\tilde{r}_i$ is Gaussian, the first term in (3.92) is easily computed:

$$\log h(y_i) \int_{\mathbb{R}^q} \tilde{r}_i(w_i) dw_i = (2\pi)^{\frac{q}{2}} |\Phi_i|^{\frac{1}{2}} \log h(y_i). \tag{3.93}$$

Next, the terms involving $y_i^T \mu_j$, $j = 1,2$ are integrated to get

$$y_i^T \mu_1 \int_{\Omega_+} \tilde{r}_i(w_i) dw_i = (2\pi)^{\frac{q-1}{2}} |\Phi_i|^{\frac{1}{2}} \sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\frac{\tau_i}{\sqrt{2}} \right) \tag{3.94}$$

and

$$y_i^T \mu_2 \int_{\Omega_-} \tilde{r}_i(w_i) dw_i = (2\pi)^{\frac{q-1}{2}} |\Phi_i|^{\frac{1}{2}} \sqrt{\frac{\pi}{2}} \left(\operatorname{erf}\left(\frac{\tau_i}{\sqrt{2}} \right) + 1 \right), \tag{3.95}$$

where $\tau_i = -\frac{[n_i]_q}{\sqrt{\phi_{iq}}}$, and ϕ_{iq} is the q^{th} element on the diagonal of Φ_i.

The rest of the terms in (3.92) involve two very similar integrals over the spaces Ω_+ and Ω_-, where these spaces were defined in (3.3). We will use the same partitioning hyperplane as in (3.4), where the normal of the hyperplane was given by $\boldsymbol{e}_q = [0,\ldots,0,1]^T$. To keep the length of this section somewhat reasonable, we will only show the details for the integration over Ω_+. We have

$$\begin{aligned}
\int_{\Omega_+} \tilde{r}_i(w_i) \left[y_i^T B_1 w_i - \psi(\mu_1 + B_1 w_i) \right] dw_i &= y_i^T B_1 \int_{\Omega_+} w_i \tilde{r}_i(w_i) dw_i - \\
&\int_{\Omega_+} \sum_{j=1}^{p} \exp\left([\mu_1 + B_1 w_i]_j \right) \tilde{r}_i(w_i) dw_i.
\end{aligned} \tag{3.96}$$

The first term in (3.96) was calculated in § 3.3 and is

$$y_i^T B_1 \int_{\Omega_+} w_i \tilde{r}_i(w_i) dw_i = (2\pi)^{\frac{q-1}{2}} |\Phi_i|^{\frac{1}{2}} y_i^T B_1 m_i \sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\frac{\tau_i}{\sqrt{2}}\right). \tag{3.97}$$

Similarly,

$$y_i^T B_2 \int_{\Omega_-} w_i \tilde{r}_i(w_i) dw_i = (2\pi)^{\frac{q-1}{2}} |\Phi_i|^{\frac{1}{2}} y_i^T B_2 m_i \sqrt{\frac{\pi}{2}} \left(\operatorname{erf}\left(\frac{\tau_i}{\sqrt{2}}\right) + 1\right). \tag{3.98}$$

For the second term in (3.96), note that $[\mu_1 + B_1 w_i]_j = \boldsymbol{e}_j^T (\mu_1 + B_1 w_i)$, with $\boldsymbol{e}_j$ the j^{th} standard basis vector in $\mathbb{R}^p$. We also use the linearity of the integral to get

$$\int_{\Omega_+} \sum_{j=1}^{p} \exp\left([\mu_1 + B_1 w_i]_j\right) \tilde{r}_i(w_i) dw_i = \sum_{j=1}^{p} \int_{\Omega_+} \exp\left(\boldsymbol{e}_j^T (\mu_1 + B_1 w_i)\right) \tilde{r}_i(w_i) dw_i. \tag{3.99}$$

We then integrate:

$$\int_{\Omega_+} \exp\left(\boldsymbol{e}_j^T (\mu_1 + B_1 w_i)\right) \tilde{r}_i(w_i) dw_i = \int_{\Omega_+} \exp\left(\boldsymbol{e}_j^T (\mu_1 + B_1 w_i) - \frac{1}{2}(w_i - n_i)^T \Phi_i^{-1} (w_i - n_i)\right) dw_i. \tag{3.100}$$

We will use the same techniques to integrate (3.100) as those used in § 3.3, but with one additional transformation. Since the integrand differs from the ones integrated above, we will show the details here. First, apply the translation $z = w_i - n_i$, $dz = dw_i$ and define the new domain of integration $\Omega_\oplus := \{z \in \mathbb{R}^q : [z]_q + [n_i]_q \geq 0\}$. We then have

$$\exp\left(\boldsymbol{e}_j^T (\mu_1 + B_1 n_i)\right) \int_{\Omega_\oplus} \exp\left(\boldsymbol{e}_j^T B_1 z - \frac{1}{2} z^T \Phi_i^{-1} z\right) dz. \tag{3.101}$$

Next, we scale the variables of integration by letting $u = \Phi_i^{-\frac{1}{2}} z$, $du = |\Phi_i|^{-\frac{1}{2}} dz$, and redefine the domain of integration as $\Omega_u := \{u \in \mathbb{R}^q : [\Phi_i^{\frac{1}{2}} u]_q + [n_i]_q \geq 0\}$ to get

$$\exp\left(\boldsymbol{e}_j^T (\mu_1 + B_1 n_i)\right) |\Phi_i|^{\frac{1}{2}} \int_{\Omega_u} \exp\left(\boldsymbol{e}_j^T B_1 \Phi_i^{\frac{1}{2}} z - \frac{1}{2} u^T u\right) dz. \tag{3.102}$$

Next, we add and subtract $\frac{1}{2} \boldsymbol{e}_j^T B_1 \Phi_i B_1^T \boldsymbol{e}_j$ inside the exponential in the integrand to get

$$\begin{aligned} &|\Phi_i|^{\frac{1}{2}} \alpha_{ij} \int_{\Omega_u} \exp\left(-\frac{1}{2}\left(u^T u - 2\boldsymbol{e}_j^T B_1 \Phi_i^{\frac{1}{2}} z + \boldsymbol{e}_j^T B_1 \Phi_i B_1^T \boldsymbol{e}_j\right)\right) dz \\ &\quad = |\Phi_i|^{\frac{1}{2}} \alpha_{ij} \int_{\Omega_u} \exp\left(-\frac{1}{2}(u - \Phi_i^{\frac{1}{2}} B_1^T \boldsymbol{e}_j)^T (u - \Phi_i^{\frac{1}{2}} B_1^T \boldsymbol{e}_j)\right) dz, \end{aligned} \tag{3.103}$$

where we have defined

$$\alpha_{ij} = \exp\left(\boldsymbol{e}_j^T(\mu_1 + B_1 n_i) + \frac{1}{2}\boldsymbol{e}_j^T B_1 \Phi_i B_1^T \boldsymbol{e}_j\right). \tag{3.104}$$

We perform one more translation $v = u - \Phi_i^{\frac{1}{2}} B_1^T \boldsymbol{e}_j$, $dv = du$, and redefine the domain of integration as

$$\begin{aligned}\Omega_v &:= \left\{ v \in \mathbb{R}^q : \left[\Phi_i^{\frac{1}{2}} v\right]_q + \left[\Phi_i^{\frac{1}{2}} B_1^T \boldsymbol{e}_j\right]_q + [n_i]_q \geq 0 \right\} \\ &= \left\{ v \in \mathbb{R}^q : \phi_{iq}^{\frac{1}{2}}[v]_q + \left[\Phi_i^{\frac{1}{2}} B_1^T \boldsymbol{e}_j\right]_q + [n_i]_q \geq 0 \right\}. \end{aligned} \tag{3.105}$$

We then obtain the familiar form

$$\begin{aligned}|\Phi_i|^{\frac{1}{2}} \alpha_{ij} \int_{\Omega_v} \exp\left(-\frac{1}{2} v^T v\right) dv &= \alpha_{ij} \int_{\xi_{ij}}^{\infty} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \exp\left(-\frac{t^2}{2}\right) dt \\ &= (2\pi)^{\frac{q-1}{2}} |\Phi_i|^{\frac{1}{2}} \alpha_{ij} \sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\frac{\xi_{ij}}{\sqrt{2}}\right), \end{aligned} \tag{3.106}$$

with

$$\xi_{ij} := -\frac{\left(\left[\Phi_i^{\frac{1}{2}} B_1^T \boldsymbol{e}_j\right]_q + [n_i]_q\right)}{\sqrt{\phi_{iq}}}. \tag{3.107}$$

Similarly,

$$\int_{\Omega_-} \exp\left(\boldsymbol{e}_j^T(\mu_2 + B_2 w_i)\right) \tilde{r}_i(w_i) dw_i = (2\pi)^{\frac{q-1}{2}} |\Phi_i|^{\frac{1}{2}} \beta_{ij} \sqrt{\frac{\pi}{2}} \left(\operatorname{erf}\left(\frac{\zeta_{ij}}{\sqrt{2}}\right) + 1\right), \tag{3.108}$$

where

$$\beta_{ij} = \exp\left(\boldsymbol{e}_j^T(\mu_2 + B_2 n_i) + \frac{1}{2}\boldsymbol{e}_j^T B_2 \Phi_i B_2^T \boldsymbol{e}_j\right), \tag{3.109}$$

and

$$\zeta_{ij} := -\frac{\left(\left[\Phi_i^{\frac{1}{2}} B_2^T \boldsymbol{e}_j\right]_q + [n_i]_q\right)}{\sqrt{\phi_{iq}}}. \tag{3.110}$$

Gathering the terms from (3.93), (3.94), (3.95), (3.97), (3.98), (3.106), and (3.108) and factor-

ing out $(2\pi)^{\frac{q}{2}}|\Phi_i|^{\frac{1}{2}}$ gives:

$$B' = c_i(2\pi)^{\frac{q}{2}}|\Phi_i|^{\frac{1}{2}}\Bigg[\log h(y_i) + (2\pi)^{-\frac{1}{2}}\Bigg[\sqrt{\frac{\pi}{2}}\text{erfc}\left(\frac{\tau_i}{\sqrt{2}}\right) + \sqrt{\frac{\pi}{2}}\left(\text{erf}\left(\frac{\tau_i}{\sqrt{2}}\right)+1\right) + y_i^T B_1 m_i\sqrt{\frac{\pi}{2}}\text{erfc}\left(\frac{\tau_i}{\sqrt{2}}\right) + y_i^T B_2 m_i\sqrt{\frac{\pi}{2}}\left(\text{erf}\left(\frac{\tau_i}{\sqrt{2}}\right)+1\right) - \sum_{j=1}^{p}\alpha_{ij}\sqrt{\frac{\pi}{2}}\text{erfc}\left(\frac{\xi_{ij}}{\sqrt{2}}\right) - \sum_{j=1}^{p}\beta_{ij}\sqrt{\frac{\pi}{2}}\left(\text{erf}\left(\frac{\zeta_{ij}}{\sqrt{2}}\right)+1\right)\Bigg]\Bigg]. \tag{3.111}$$

Now, $c_i = (2\pi)^{-\frac{q}{2}}|\Phi_i|^{-\frac{1}{2}}$, so the global factor of $(2\pi)^{\frac{q}{2}}|\Phi_i|^{\frac{1}{2}}$ vanishes. Where possible, we also gather $\text{erf}(\cdot)$ and $\text{erfc}(\cdot)$ terms to obtain

$$B' = \log h(y_i) + \frac{1}{2}\Bigg[\text{erfc}\left(\frac{\tau_i}{\sqrt{2}}\right)\left[1 + y_i^T B_1 m_i\right] + \left(\text{erf}\left(\frac{\tau_i}{\sqrt{2}}\right)+1\right)\left[1 + y_i^T B_2 m_i\right] - \sum_{j=1}^{p}\alpha_{ij}\text{erfc}\left(\frac{\xi_{ij}}{\sqrt{2}}\right) - \sum_{j=1}^{p}\beta_{ij}\left(\text{erf}\left(\frac{\zeta_{ij}}{\sqrt{2}}\right)+1\right)\Bigg]. \tag{3.112}$$

We may now combine the terms A', B', and C' to give the VLB for the Poisson model. We also note that the individual terms involving the error function and complementary error function are only computed over subsets of $1 \leq i \leq N$. Therefore, we define the two index sets $\mathfrak{I}_1$ and $\mathfrak{I}_2$ by

$$\mathfrak{I}_1 := \{1 \leq i \leq N : \tau_i \geq 0\} \tag{3.113}$$

$$\mathfrak{I}_2 := \{1 \leq i \leq N : \tau_i < 0\}. \tag{3.114}$$

Then, with the order of summing C' and B' conveniently switched, we have the VLB for the Poisson

model:

$$
\begin{aligned}
\mathcal{J}(\tilde{r},\theta) &= \sum_{i=1}^{N} A' - C' + B' \\
&= \sum_{i=1}^{N} -\frac{1}{2}\left(q\log(2\pi) + \mathrm{Tr}\left(\Phi_i\right) + m_i^T m_i\right) - \left(-\frac{1}{2}\left(q\log(2\pi) + \log|\Phi_i| + q\right)\right) + \\
&\qquad \log h(y_i) + \\
&\qquad \frac{1}{2}\left[\sum_{i\in\mathcal{J}_1}\left(\mathrm{erfc}\left(\frac{\tau_i}{\sqrt{2}}\right)\left[1 + y_i^T B_1 m_i\right] - \sum_{j=1}^{p}\alpha_{ij}\mathrm{erfc}\left(\frac{\xi_{ij}}{\sqrt{2}}\right)\right) + \right. \\
&\qquad \left.\sum_{i\in\mathcal{J}_2}\left(\left(\mathrm{erf}\left(\frac{\tau_i}{\sqrt{2}}\right) + 1\right)\left[1 + y_i^T B_2 m_i\right] - \sum_{j=1}^{p}\beta_{ij}\left(\mathrm{erf}\left(\frac{\zeta_{ij}}{\sqrt{2}}\right) + 1\right)\right)\right] \\
&= \frac{1}{2}\sum_{i=1}^{N}\left[2\log h(y_i) - \left(\mathrm{Tr}\left(\Phi_i\right) + m_i^T m_i + \log|\Phi_i| + q\right) + \right. \\
&\qquad \left[\sum_{i\in\mathcal{J}_1}\left(\mathrm{erfc}\left(\frac{\tau_i}{\sqrt{2}}\right)\left[1 + y_i^T B_1 m_i\right] - \sum_{j=1}^{p}\alpha_{ij}\mathrm{erfc}\left(\frac{\xi_{ij}}{\sqrt{2}}\right)\right) + \right. \\
&\qquad \left.\left.\sum_{i\in\mathcal{J}_2}\left(\left(\mathrm{erf}\left(\frac{\tau_i}{\sqrt{2}}\right) + 1\right)\left[1 + y_i^T B_2 m_i\right] - \sum_{j=1}^{p}\beta_{ij}\left(\mathrm{erf}\left(\frac{\zeta_{ij}}{\sqrt{2}}\right) + 1\right)\right)\right]\right],
\end{aligned}
\tag{3.115}
$$

with

$$
\alpha_{ij} = \exp\left(\boldsymbol{e}_j^T\left(\mu_1 + B_1 n_i\right) + \frac{1}{2}\boldsymbol{e}_j^T B_1 \Phi_i B_1^T \boldsymbol{e}_j\right), \tag{3.116}
$$

$$
\beta_{ij} = \exp\left(\boldsymbol{e}_j^T\left(\mu_2 + B_2 n_i\right) + \frac{1}{2}\boldsymbol{e}_j^T B_2 \Phi_i B_2^T \boldsymbol{e}_j\right), \tag{3.117}
$$

$$
\tau_i = -\frac{[n_i]_q}{\sqrt{\phi_{iq}}}, \tag{3.118}
$$

$$
\xi_{ij} = -\frac{\left(\left[\Phi_i^{\frac{1}{2}} B_1^T \boldsymbol{e}_j\right]_q + [n_i]_q\right)}{\sqrt{\phi_{iq}}}, \tag{3.119}
$$

$$
\zeta_{ij} = -\frac{\left(\left[\Phi_i^{\frac{1}{2}} B_2^T \boldsymbol{e}_j\right]_q + [n_i]_q\right)}{\sqrt{\phi_{iq}}}, \tag{3.120}
$$

$$
h(y_i) = \exp\left(\sum_{j=1}^{p}\sum_{k=j}^{y_{ij}}\log k\right). \tag{3.121}
$$

3.6.1 Optimizing the Model Parameters - Poisson Observations

In this section, we compute the derivatives of the VLB from § 3.6 with respect to the model parameters in the case of Poisson distributed observations.

First, let $\boldsymbol{\Delta}_{ij}$ be the $p \times q$ matrix with 1 in its $(i,j)^{\text{th}}$ position and zeros elsewhere, and $\boldsymbol{M}_{jj}$ be the analogous diagonal $p \times p$ matrix. Also, we will relabel the error function and complementary error function for convenience:

$$F(x) := \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right), \tag{3.122}$$

$$G(x) := \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + 1. \tag{3.123}$$

Then,

$$\frac{\partial \mathcal{J}}{\partial \mu_1} = y_i F(\tau_i) + \sum_{j=1}^{p} \boldsymbol{e}_j \alpha_{ij} F(\xi_{ij}), \tag{3.124}$$

$$\frac{\partial \mathcal{J}}{\partial \mu_2} = y_i G(\tau_i) + \sum_{j=1}^{p} \boldsymbol{e}_j \beta_{ij} G(\zeta_{ij}), \tag{3.125}$$

$$\frac{\partial \mathcal{J}}{\partial B_1} = y_i n_i^T F(\tau_i) + \sum_{j=1}^{p} \alpha_{ij} \left[F(\xi_{ij}) \left(e_j n_i^T - \boldsymbol{M}_{jj} B_1 \Phi_i^{\frac{1}{2}} \right) + \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\xi_{ij}^2}{2} \right) \boldsymbol{\Delta}_{ij} \right], \tag{3.126}$$

$$\frac{\partial \mathcal{J}}{\partial B_2} = y_i n_i^T G(\tau_i) + \sum_{j=1}^{p} \beta_{ij} \left[G(\zeta_{ij}) \left(e_j n_i^T - \boldsymbol{M}_{jj} B_2 \Phi_i^{\frac{1}{2}} \right) + \sqrt{\frac{2}{\pi}} \exp\left(-\frac{\zeta_{ij}^2}{2} \right) \boldsymbol{\Delta}_{ij} \right]. \tag{3.127}$$

We do not have closed forms for the critical points of these derivatives, so an iterative procedure such as gradient descent must be used to estimate the corresponding local optima.

3.6.2 Optimizing the Variational Parameters - Poisson Observations

Differentiating with respect to the variational parameters also fails to yield closed forms for the corresponding critical points, and an iterative procedure must be used to find their local optima.

First, the variational means. Noting that $\frac{\partial \xi_{ij}}{\partial n_i} = \frac{\partial \zeta_{ij}}{\partial n_i} = -\boldsymbol{e}_q \phi_{iq}^{-\frac{1}{2}}$, where $\phi_{iq} = [\Phi_i]_q$. Then,

$$\begin{aligned}\left.\frac{\partial \mathcal{J}}{\partial n_i}\right|_{\mathfrak{I}_1} =& \sqrt{\frac{2}{\pi}} \phi_{iq}^{-\frac{1}{2}} \exp\left(-\frac{\tau_i^2}{2}\right) y_i^T \left[\mu_1 + B_1 n_i\right] + B_1^T y_i F(\tau_i) + \\ & \sum_{j=1}^{p} \alpha_{ij} \left[B_1^T \boldsymbol{e}_j F(\xi_{ij}) + \sqrt{\frac{2}{\pi}} \phi_{iq}^{-\frac{1}{2}} \exp\left(-\frac{\xi_{ij}^2}{2}\right) \boldsymbol{e}_q\right], \end{aligned} \tag{3.128}$$

and

$$\begin{aligned}\left.\frac{\partial \mathcal{J}}{\partial n_i}\right|_{\mathfrak{I}_2} =& -\sqrt{\frac{2}{\pi}} \phi_{iq}^{-\frac{1}{2}} \exp\left(-\frac{\tau_i^2}{2}\right) y_i^T \left[\mu_2 + B_2 n_i\right] + B_2^T y_i G(\tau_i) + \\ & \sum_{j=1}^{p} \beta_{ij} \left[B_2^T \boldsymbol{e}_j G(\zeta_{ij}) - \sqrt{\frac{2}{\pi}} \phi_{iq}^{-\frac{1}{2}} \exp\left(-\frac{\zeta_{ij}^2}{2}\right) \boldsymbol{e}_q\right]. \end{aligned} \tag{3.129}$$

Similarly to (3.68), define

$$\frac{\partial \tau_i}{\partial \Phi_i} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & [n_i]_q \phi_{iq}^{-1} \end{bmatrix}, \tag{3.130}$$

$$\frac{\partial \xi_i}{\partial \Phi_i} = \frac{\partial \zeta_i}{\partial \Phi_i} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & [n_i]_q \phi_{iq}^{-3/2} \end{bmatrix}. \tag{3.131}$$

Note that since Φ_i is diagonal, $\frac{d\Phi_i^{\frac{1}{2}}}{d\Phi_i} = \frac{1}{2}\Phi_i^{-\frac{1}{2}}$, where

$$\Phi_i^{-\frac{1}{2}} = \begin{bmatrix} \phi_{11}^{-\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & \phi_{22}^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \phi_{qq}^{-\frac{1}{2}} \end{bmatrix}. \tag{3.132}$$

Then, the derivatives of the VLB with respect to the variational covariance matrices are given by:

$$\left.\frac{\partial \mathcal{J}}{\partial \Phi_i}\right|_{\mathfrak{I}_1} = -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\tau_i^2}{2}\right) y_i^T\left(\mu_1 + B_1 n_i\right)\frac{\partial \tau_i}{\partial \Phi_i} + \sum_{j=1}^{p} \alpha_{ij}\left[\frac{1}{2}F(\xi_{ij})B_1^T \boldsymbol{e}_j \boldsymbol{e}_j^T B_1 \Phi_i^{-\frac{1}{2}} - \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\xi_{ij}^2}{2}\right)\frac{\partial \xi_i}{\partial \Phi_i}\right], \tag{3.133}$$

and

$$\left.\frac{\partial \mathcal{J}}{\partial \Phi_i}\right|_{\mathfrak{I}_2} = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\tau_i^2}{2}\right) y_i^T\left(\mu_2 + B_2 n_i\right)\frac{\partial \tau_i}{\partial \Phi_i} + \sum_{j=1}^{p} \beta_{ij}\left[\frac{1}{2}G(\xi_{ij})B_2^T \boldsymbol{e}_j \boldsymbol{e}_j^T B_2 \Phi_i^{-\frac{1}{2}} + \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\zeta_{ij}^2}{2}\right)\frac{\partial \zeta_i}{\partial \Phi_i}\right]. \tag{3.134}$$

3.7 Estimating the Dimension of the Latent Space

Until now, we have assumed a priori knowledge of the dimension of the latent space, given by q. However, in general, this must be estimated. Therefore, we will use a variation of the Bayesian Information Criterion (BIC) introduced in [69] and defined for Poisson PCA in [16]. For a model fitted on N observations and with k parameters, the BIC is given by

$$\boldsymbol{BIC}(k) := \mathcal{J}(k) - \frac{1}{2}k\log N, \tag{3.135}$$

where $\mathcal{J}$ is the log-likelihood of the model. The piecewise PCA model has $2pq$ parameters when using the two-partition scheme in § 3.2 (Kpq parameters when using a K-partition scheme), so the BIC for the piecewise PCA model is

$$\boldsymbol{BIC}(q) = \mathcal{J}(q) - pq\log N. \tag{3.136}$$

In practice, we will choose a selection of positive integers $1 \leq q < p$ and fit a piecewise PCA model for each. In each model, we will evaluate the log-likelihood and report the corresponding BIC. We may then choose the value of q for which the BIC is maximal.

3.8 Variance Explained by the Piecewise Models

Due to the nonlinear nature of the piecewise affine PCA models (3.6) and (3.86), the diagonalization of the posterior covariance matrix will not yield the proportion of explained variance through its eigenvalues, as was done in § 2.4.2. Instead, we follow the work of [16], where a pseudo-R^2 criterion was developed as an analogue to the explained variance of classical PCA. The resulting criterion is the proportion of the log-likelihood that can be accounted for by the model.

To calculate this criterion, denoted by ρ^2 for convenience, we must first calculate the marginal log-likelihood of our observation data given the fitted parameters from the model. This marginal data log-likelihood can be written using the canonical form for NEF distributions (see Definition 0.5.1):

$$L := \sum_{i=1}^{N} \log p(y_i|\theta) = \sum_{i=1}^{N} \log h(y_i) + T(y_i)^T \cdot \theta - \psi(\theta). \tag{3.137}$$

We next compute the log-likelihood for the null model, i.e. the model with no parameters.

$$\begin{aligned} L_{\min} &:= \sum_{i=1}^{N} \log h(y_i) + T(y_i)^T \cdot \mathbf{0} - \psi(\mathbf{0}) \\ &= \sum_{i=1}^{N} \log h(y_i). \end{aligned} \tag{3.138}$$

Finally, we compute the log-likelihood for the 'saturated' model, i.e. the model with one parameter per observation. We will denote these parameters by η_i and compute:

$$L_{\max} := \sum_{i=1}^{N} \log h(y_i) + T(y_i)^T \cdot \eta_i - \psi(\eta_i).$$

Since the parameters $\{\eta_i\}$ used in $L_{\max}$ are unknown, they must be estimated based on the canonical form of the NEF distribution. For the parameter estimates $\hat{\eta}_i$, we turn to the definition of the canonical link function for NEF distributions. Recall from Definition 0.5.2 that the link function g is defined such that

$$g(\psi'(\eta)) = \eta. \tag{3.139}$$

To obtain an unconstrained estimate $\hat{\eta}_i$ of a single parameter η_i, we take the derivative of $L_{\max}$ with

respect to η_i.

$$\frac{\partial L_{\max}}{\partial \eta_i} = T(y_i) - \psi'(\eta_i) \tag{3.140}$$

Setting this derivative to zero, we obtain the parameter estimate $\hat{\eta}_i$:

$$\hat{\eta}_i = (\psi')^{-1}(T(y_i)) = g(T(y_i)). \tag{3.141}$$

The value of ρ^2 is then defined as

$$\rho^2 := \frac{L - L_{\min}}{L_{\max} - L_{\min}}. \tag{3.142}$$

Now, consider model (3.6), which uses a Gaussian distribution with known isotropic covariance for the prior distribution. For canonical parameter vector $\boldsymbol{\mu}$, and $\psi(\boldsymbol{\mu}) = \frac{1}{\sigma^2}\boldsymbol{\mu}^T\boldsymbol{\mu}$, the canonical link function is given by

$$g(\boldsymbol{\mu}) = \sigma^2\boldsymbol{\mu}. \tag{3.143}$$

Therefore, the parameter estimates for $L_{\max}$ with a Gaussian prior are

$$\hat{\eta}_i = g(T(y_i)) = g(\sigma^{-2}y_i) = y_i. \tag{3.144}$$

Next, consider model (3.86), where a Poisson prior on the observations is used. The canonical parameter vector is now $\boldsymbol{\lambda}$, $\psi(\boldsymbol{\lambda}) = \sum_{j=1}^{p} \exp(\lambda_j)$, and the canonical link function is given by

$$g(\boldsymbol{\lambda}) = \log \boldsymbol{\lambda}. \tag{3.145}$$

Therefore, the parameters estimates for $L_{\max}$ with a Poisson prior are

$$\hat{\eta}_i = g(T(y_i)) = \log T(y_i) = \log y_i. \tag{3.146}$$

In the case of the piecewise model (3.6), the log-likelihood function must be split according to the partition in the latent space. Therefore, we again denote by I_1 and I_2 the index sets corresponding to generative latent points residing in the two disjoint partitions Ω_+ and Ω_-. The marginal data log-likelihood is then partitioned into the two sums:

$$L = \sum_{i \in I_1} \log p_1(y_i|\theta_1) + \sum_{i \in I_2} \log p_2(y_i|\theta_2), \tag{3.147}$$

where $\theta_1 := \{\sigma^2, \mu_1, B_1\}$ and $\theta_2 := \{\sigma^2, \mu_2, B_2\}$, and the computations for the marginal distributions are completed in § 3.9. Borrowing the result before the calculations are completed and assuming that the latent variables are distributed according to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we have

$$\log p_1(y_i|\theta_1) = \log c - \frac{1}{2}(y_i - \mu_1)^T C_{y_1}^{-1}(y_i - \mu_1) + \log\left(\operatorname{erfc}(0)\right), \tag{3.148}$$

$$\log p_2(y_i|\theta_2) = \log c - \frac{1}{2}(y_i - \mu_2)^T C_{y_2}^{-1}(y_i - \mu_2) + \log\left((\operatorname{erf}(0) + 1\right), \tag{3.149}$$

where

$$c = 2^{-q-p/2}\pi^{1-q-p/2}(\sigma^2)^{-p/2}, \tag{3.150}$$

$$C_{y_1}^{-1} = (\sigma^2 \boldsymbol{I} + B_1 B_1^T)^{-1}, \tag{3.151}$$

$$C_{y_2}^{-1} = (\sigma^2 \boldsymbol{I} + B_2 B_2^T)^{-1}, \tag{3.152}$$

and the assumptions of normality on the latent variables mean that $\operatorname{erfc}(0) = \operatorname{erf}(0) + 1 = 1$. In this work, we generate estimates of the saturated model parameters based on the parameters corresponding to the index sets. Over both index sets, the estimated saturated model parameters are $\hat{\eta}_i|_{I_j} = y_i$, $j = 1, 2$. Therefore, for the piecewise PCA model, we have

$$L_{\max} = N \log c. \tag{3.153}$$

For the piecewise PCA model with no parameters, we have

$$L_{\min} = N \log c - \frac{\sigma^2}{2} \sum_{i=1}^{N} y_i^T y_i. \tag{3.154}$$

In order to compute an analogue to the proportion of explained variance provided by classical PCA, we may first find the lower-dimensional latent representation using piecewise PCA and compute its ρ^2 value. Next, we may decorrelate the dimensions in the latent representation with a full-rank classical PCA on the latent space covariance matrix, which provides the proportion of explained variance along each of the decorrelated latent dimensions. The explained variance from the classical PCA is given by the scaled eigenvectors d_j, $1 \leq j \leq q$, so we then define the explained variance of the piecewise PCA model along the j^{th} latent dimension by $d_j \rho^2$, as in [16].

3.9 Computing the Marginal Data Distribution

In this section we compute the marginal data distribution for the piecewise affine dimension reduction model (3.6). Difficulties relating to the direct maximization of the likelihood function using an EM algorithm are detailed in § 3.9.1.

First, note that the partitioning of the latent space into two half spaces will necessarily result in two corresponding partitions in the observation space. We will denote the partitioned observation spaces as $\mathfrak{P}_+$ and $\mathfrak{P}_-$ and compute the marginal likelihood for observations y over the appropriate partitions. We will also make use of the indicator function:

$$I_{\mathfrak{P}}(y) := \begin{cases} 1, & y \in \mathfrak{P}, \\ 0, & y \notin \mathfrak{P}. \end{cases} \tag{3.155}$$

Given the prior distribution $p(w) = (2\pi)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}w^T w\right)$ and the joint distribution of the model, we compute the marginal data distribution as:

$$\begin{aligned} p(y|\theta) &= \int_{\mathbb{R}^q} p(y|w,\theta)p(w)dw \\ &= (2\pi)^{-(p+q)/2}\sigma^{-p}\int_{\mathbb{R}^q}\exp\left(-\frac{1}{2\sigma^2}(y-f(w,\theta))^T(y-f(w,\theta))\right)\exp\left(-\frac{1}{2}w^T w\right)dw \\ &= c_1\left[I_{\mathfrak{P}_+}(y)\int_{\Omega_+}\exp\left(-\frac{1}{2\sigma^2}\left[(y-\mu_1-B_1w)^T(y-\mu_1-B_1w)+\sigma^2 w^T w\right]\right)dw + \right. \\ &\qquad \left. I_{\mathfrak{P}_-}(y)\int_{\Omega_-}\exp\left(-\frac{1}{2\sigma^2}\left[(y-\mu_2-B_2w)^T(y-\mu_2-B_2w)+\sigma^2 w^T w\right]\right)dw\right], \end{aligned} \tag{3.156}$$

where $c_1 := (2\pi)^{-(p+q)/2}\sigma^{-p}$. Since the calculation is very similar for both Ω_+ and Ω_-, we will focus on integrating over Ω_+.

Remark. Given Gaussian distributed latent variables $\{W_i\}$ and observations with a canonical NEF distribution (see Definition 0.5.1), we may also derive more general marginal distributions $p(y_i|\theta)$. In such a case, the marginal distribution is given by

$$p(y_i|\theta) = (2\pi)^{-\frac{q}{2}}h(y_i)\int\exp\left(T(y_i)^T f(w_i,\theta) - \psi(f(w_i,\theta)) - \frac{1}{2}w_i^T w_i\right)dw_i. \tag{3.157}$$

Since the integration of (3.157) requires knowledge of the exact functional form of ψ, we will leave the derivation of the marginal distribution for the general NEF case for the future and instead focus on the Gaussian case. As seen earlier in § 3.4 and § 3.6, we used variational approximations to the log-likelihood function for Gaussian and Poisson prior distributions, respectively.

We have:

$$\begin{aligned}
P &:= c_1 \int_{\Omega_+} \exp\left(-\frac{1}{2\sigma^2}\left[(y-\mu_1-B_1w)^T(y-\mu_1-B_1w)+\sigma^2 w^T w\right]\right) dw \\
&= c_1 \int_{\Omega_+} \exp\left(-\frac{1}{2\sigma^2}\left[y^T y - 2\mu_1^T y + \mu_1^T\mu_1 - 2(y-\mu_1)^T B_1 w + (B_1 w)^T B_1 w + \sigma^2 w^T w\right]\right) \\
&= c_1 \exp\left(-\frac{1}{2\sigma^2}(y-\mu_1)^T(y-\mu_1)\right) \times \\
&\quad \int_{\Omega_+} \exp\left(-\frac{1}{2\sigma^2}\left[w^T(B_1^T B_1+\sigma_2 I)w - 2(y-\mu_1)^T B_1 w\right]\right) dw.
\end{aligned} \tag{3.158}$$

Now, as in [75], let

$$C_{w_1}^{-1} = \sigma^{-2}(B_1^T B_1 + \sigma^2 I) \tag{3.159}$$

and

$$\mu_{w_1}^T C_{w_1}^{-1} = \sigma^{-2}(y-\mu_1)^T B_1. \tag{3.160}$$

Next, add and subtract $\mu_{w_1}^T C_{w_1}^{-1}\mu_{w_1}$ inside the exponential function in the integrand of (3.158) to get:

$$\begin{aligned}
P &= c_1 \exp\left(-\frac{1}{2\sigma^2}(y-\mu_1)^T(y-\mu_1)\right) \times \\
&\quad \int_{\Omega_+} \exp\left(-\frac{1}{2}\left[w^T C_{w_1}^{-1} w - 2\mu_{w_1}^T C_{w_1}^{-1}\right] w\right) dw \\
&= c_1 \exp\left(-\frac{1}{2\sigma^2}(y-\mu_1)^T(y-\mu_1)\right) \times \\
&\quad \int_{\Omega_+} \exp\left(-\frac{1}{2}\left[w^T C_{w_1}^{-1} w - 2\mu_{w_1}^T C_{w_1}^{-1} w + \mu_{w_1}^T C_{w_1}^{-1}\mu_{w_1} - \mu_{w_1}^T C_{w_1}^{-1}\mu_{w_1}\right]\right) dw \\
&= c_1 \exp\left(-\frac{1}{2\sigma^2}\left[(y-\mu_1)^T(y-\mu_1) - \sigma^2\mu_{w_1}^T C_{w_1}^{-1}\mu_{w_1}\right]\right) \times \\
&\qquad \int_{\Omega_+} \exp\left(-\frac{1}{2}(w-\mu_{w_1})^T C_{w_1}^{-1}(w-\mu_{w_1})\right) dw.
\end{aligned} \tag{3.161}$$

Back-substitution of $\mu_{w_1}^T C_{w_1}^{-1}$ into (3.161) gives:

$$P = c_1 \exp\left(-\frac{1}{2}\left[(y-\mu_1)^T(\sigma^{-2}I - \sigma^{-2}B_1(\sigma^{-2}B_1^T B_1 + I)^{-1}B_1^T\sigma^{-2}I)(y-\mu_1)\right]\right) \times \int_{\Omega_+} \exp\left(-\frac{1}{2}(w-\mu_{w_1})^T C_{w_1}^{-1}(w-\mu_{w_1})\right) dw. \tag{3.162}$$

There is now a quadratic form inside the exponential function in front of the integral. However, it is hard to see due to the number of terms. To put this exponential function in a more usable form, we use the Woodbury identity for matrix inverses (see [60], identity (156)), which states that for three invertible matrices $\mathcal{A}, \mathcal{B}, \mathcal{C}$,

$$\mathcal{A}^{-1} - \mathcal{A}^{-1}\mathcal{C}\left(\mathcal{B}^{-1} + \mathcal{C}^T\mathcal{A}\mathcal{C}\right)^{-1}\mathcal{C}^T\mathcal{A}^{-1} = \left(\mathcal{A} + \mathcal{C}\mathcal{B}\mathcal{C}^T\right)^{-1}. \tag{3.163}$$

Applying this identity to (3.162) gives:

$$\begin{aligned} P &= c_1 \exp\left(-\frac{1}{2}\left[(y-\mu_1)^T\left(\sigma^2 I + B_1 B_1^T\right)^{-1}(y-\mu_1)\right]\right) \times \\ &\qquad \int_{\Omega_+} \exp\left(-\frac{1}{2}(w-\mu_{w_1})^T C_{w_1}^{-1}(w-\mu_{w_1})\right) dw \\ &= c_1 \exp\left(-\frac{1}{2}\left[(y-\mu_1)^T C_{y_1}^{-1}(y-\mu_1)\right]\right) \times \\ &\qquad \int_{\Omega_+} \exp\left(-\frac{1}{2}(w-\mu_{w_1})^T C_{w_1}^{-1}(w-\mu_{w_1})\right) dw, \end{aligned} \tag{3.164}$$

where $C_{y_1}^{-1} = \left(\sigma^2 I + B_1 B_1^T\right)^{-1}$. Now, by the computations in § 3.3, the expression (3.164) becomes

$$c_1 \exp\left(-\frac{1}{2}\left[(y-\mu_1)^T\left(\sigma^2 I + B_1 B_1^T\right)^{-1}(y-\mu_1)\right]\right) |\sqrt{D_{w_1}}| (2\pi)^{(1-q)/2}\sqrt{\frac{\pi}{2}}\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right), \tag{3.165}$$

where $s_1 = -\frac{(\nu^T\mu_{w_1}+\beta)}{\sqrt{\nu^T C_{w_1}^{-1}\nu}}$, and the factor of $|\sqrt{D_{w_1}}|$ is the product of the singular values of the inverse latent covariance matrix $C_{w_1}^{-1}$.

A similar result holds over the half-space Ω_-, and the resulting marginal probability density function for y is:

$$\begin{aligned} p(y|\theta) = c\Bigg[&|\sqrt{D_{w_1}}| I_{\mathfrak{P}_+}(y) \exp\left(-\frac{1}{2}(y-\mu_1)^T C_{y_1}^{-1}(y-\mu_1)\right)\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right) + \\ &|\sqrt{D_{w_2}}| I_{\mathfrak{P}_-}(y) \exp\left(-\frac{1}{2}(y-\mu_2)^T C_{y_2}^{-1}(y-\mu_2)\right)\left(\operatorname{erf}\left(\frac{s_2}{\sqrt{2}}\right)+1\right)\Bigg], \end{aligned} \tag{3.166}$$

where

$$c := 2^{-q-p/2}\pi^{1-q-p/2}(\sigma^2)^{-p/2}. \tag{3.167}$$

3.9.1 Marginal Data Likelihood Function

The purpose of this section is to illustrate the computational difficulties faced when attempting to maximize the log-likelihood using the marginal probability density presented above in § 3.9.

We want to maximize

$$p(Y|\theta) = \prod_{i=1}^{N} p(Y_i|\theta) \tag{3.168}$$

over the dataset $(Y_i)_{i=1}^N$, where $\theta = \{\sigma^2, \mu_1, \mu_2, B_1, B_2\}$. To make things slightly simpler, we will take the logarithm of this product to obtain:

$$\begin{aligned}
\log p(Y|\theta) &= \sum_{i=1}^{N} \log p(Y_i|\theta) \\
&= \sum_{i=1}^{N} \log \Bigg(2^{-(p/2+1)}\pi^{-p/2}\times \\
&\qquad \Bigg[|\sqrt{D_{w_1}}|\exp\left(-\frac{1}{2}(y_i-\mu_1)^T C_{y_1}^{-1}(y_i-\mu_1)\right)\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right)+ \\
&\qquad |\sqrt{D_{w_2}}|\exp\left(-\frac{1}{2}(y_i-\mu_2)^T C_{y_2}^{-1}(y_i-\mu_2)\right)\left(\operatorname{erf}\left(\frac{s_2}{\sqrt{2}}\right)+1\right)\Bigg]\Bigg) \\
&= \sum_{i=1}^{N} -\left(\frac{p}{2}+1\right)\log(2) - \frac{p}{2}\log(\pi) + \\
&\qquad \log\Bigg[|\sqrt{D_{w_1}}|\exp\left(-\frac{1}{2}(y_i-\mu_1)^T C_{y_1}^{-1}(y_i-\mu_1)\right)\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right)+ \\
&\qquad |\sqrt{D_{w_2}}|\exp\left(-\frac{1}{2}(y_i-\mu_2)^T C_{y_2}^{-1}(y_i-\mu_2)\right)\left(\operatorname{erf}\left(\frac{s_2}{\sqrt{2}}\right)+1\right)\Bigg]
\end{aligned} \tag{3.169}$$

Because of the presence of summation inside the logarithm function, we will use an Expectation-Maximization (EM) algorithm (see § 1.3) to maximize (3.169). However, as will soon be shown, a naive EM algorithm leads to difficult derivatives involving quotients.

In the E-step, we fix initial values for the model parameters,

$$\theta^{[0]} = \left\{\sigma^{2^{[0]}}, \mu_1^{[0]}, \mu_2^{[0]}, B_1^{[0]}, B_2^{[0]}\right\},$$

and then we calculate the expectation of the joint pdf with respect to the posterior distribution $p(w|y,\theta)$. For the purposes of the calculations below, we will slightly abuse the notation and use θ

to denote the fixed model parameters. Unless otherwise stated, all integration takes place over $\mathbb{R}^q$.

$$\begin{aligned}
\langle L\rangle &= \int \sum_{i=1}^{N} \log\left(p(y_i,w_i|\theta)\right) p(w_i|y_i,\theta)dw_i \\
&= \sum_{i=1}^{N} \int \log\left[p(y_i|w_i,\theta)p(w_i)\right] p(w_i|y_i,\theta)dw_i \\
&= \sum_{i=1}^{N} \int \left[\log p(y_i|w_i,\theta) + \log p(w_i)\right] p(w_i|y_i,\theta)dw_i \\
&= \sum_{i=1}^{N} \int \left[\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}||y_i - f(w_i,\theta)||^2 + \log 2\pi^{-q/2} - \frac{1}{2}w_i^T w_i\right] p(w_i|y_i,\theta)dw_i \\
&= \sum_{i=1}^{N} \left(-\frac{p}{2}\log 2\pi\sigma^2 - \frac{q}{2}\log 2\pi - \frac{1}{2\sigma^2}\int \left[||y_i - f(w_i,\theta)||^2 + \sigma^2 I w_i^T w_i\right] p(w_i|y_i,\theta)dw_i\right) \\
&= -\frac{Np}{2}\log 2\pi\sigma^2 - \frac{Nq}{2}\log 2\pi - \sum_{i=1}^{N}\frac{1}{2\sigma^2}\int \left[||y_i - f(w_i,\theta)||^2 + \sigma^2 I w_i^T w_i\right] p(w_i|y_i,\theta)dw_i.
\end{aligned} \tag{3.170}$$

Since $f(w_i,\theta)$ is defined piecewise, the integral in (3.170) splits into two integrals over the previously defined disjoint sets Ω_+ and Ω_- (see (3.3)). Since these two integrals are very similar, only the details for the integration over Ω_+ will be shown.

$$\begin{aligned}
E_1 &:= -\frac{1}{2\sigma^2}\int_{\Omega_+} \left[||y_i - f(w_i,\theta)||^2 + \sigma^2 I w_i^T w_i\right] p(w_i|y_i,\theta)dw_i \\
&= -\frac{1}{2\sigma^2}\int_{\Omega_+} \left[(y_i - \mu_1 - B_1 w_i)^T (y_i - \mu_1 - B_1 w_i) + \sigma^2 w_i^T w_i\right] p(w_i|y_i,\theta)dw_i \\
&= -\frac{1}{2\sigma^2}\int_{\Omega_+} \left[(y_i-\mu_1)^T(y_i-\mu_1) - 2(y_i-\mu_1)^T B_1 w_i + w_i^T(B_1^T B_1 + \sigma^2 I)w_i\right] p(w_i|y_i,\theta)dw_i \\
&= -\frac{1}{2\sigma^2}\left[(y_i-\mu_1)^T(y_i-\mu_1)\int_{\Omega_+} p(w_i|y_i,\theta)dw_i - \right. \\
&\qquad 2(y_i-\mu_1)^T B_1 \int_{\Omega_+} w_i p(w_i|y_i,\theta)dw_i + \\
&\qquad \left.\int_{\Omega_+} w_i^T(B_1^T B_1 + \sigma^2 I)w_i p(w_i|y_i,\theta)dw_i\right].
\end{aligned} \tag{3.171}$$

We use the results from above and the posterior distribution from § 3.10 to obtain

$$\begin{aligned}
E_1 = -\frac{c_2(2\pi)^{(1-q)/2}}{2\sigma^2(P_1+P_2)}&\left[|\sqrt{D_{w_1}}|\exp\left(-\frac{1}{2}(y_i-\mu_1)^T C_{y_1}^{-1}(y_i-\mu_1)\right)\times\right. \\
&\left.\left(\sqrt{\frac{\pi}{2}}\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right)\left(2 - 2(y_i-\mu_1)^T B_1\mu_{w_1} + \mu_{w_1}^T C_{w_1}^{-1}\mu_{w_1}\right) + s_1\exp\left(-s_1^2/2\right)\right)\right].
\end{aligned} \tag{3.172}$$

Similarly, the expectation over Ω_- is given by:

$$E_2 := -\frac{c_2(2\pi)^{(1-q)/2}}{2\sigma^2(P_1+P_2)}\left[|\sqrt{D_{w_2}}|\exp\left(-\frac{1}{2}(y_i-\mu_2)^T C_{y_2}^{-1}(y_i-\mu_2)\right)\times\right.$$
$$\left.\left(\sqrt{\frac{\pi}{2}}\left(\operatorname{erf}\left(\frac{s_2}{\sqrt{2}}\right)+1\right)\left(2-2(y_i-\mu_2)^T B_2\mu_{w_2}+\mu_{w_2}^T C_{w_2}^{-1}\mu_{w_2}\right)-s_2\exp\left(-s_2^2/2\right)\right)\right]. \tag{3.173}$$

Therefore, we have:

$$\begin{aligned}\langle L\rangle = &-\frac{Np}{2}\log(2\pi\sigma^2)-\frac{Nq}{2}\log(2\pi)-\sum_{i=1}^{N}\frac{c_2(2\pi)^{(1-q)/2}}{2\sigma^2(P_1+P_2)}\times\\
&\left[|\sqrt{D_{w_1}}|\exp\left(-\frac{1}{2}(y_i-\mu_1)^T C_{y_1}^{-1}(y_i-\mu_1)\right)\times\right.\\
&\left(\sqrt{\frac{\pi}{2}}\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right)\left(2-2(y_i-\mu_1)^T B_1\mu_{w_1}+\mu_{w_1}^T C_{w_1}^{-1}\mu_{w_1}\right)+s_1\exp\left(-s_1^2/2\right)\right)+\\
&|\sqrt{D_{w_2}}|\exp\left(-\frac{1}{2}(y_i-\mu_2)^T C_{y_2}^{-1}(y_i-\mu_2)\right)\times\\
&\left.\left(\sqrt{\frac{\pi}{2}}\left(\operatorname{erf}\left(\frac{s_2}{\sqrt{2}}\right)+1\right)\left(2-2(y_i-\mu_2)^T B_2\mu_{w_2}+\mu_{w_2}^T C_{w_2}^{-1}\mu_{w_2}\right)-s_2\exp\left(-s_2^2/2\right)\right)\right].\end{aligned} \tag{3.174}$$

For the M-step, we will need to take derivatives of the above expression with respect to the parameters in θ. However, the presence of $P1+P2$ in the denominator of (3.174) leads to difficult derivatives with singularities.

The two problems of the summation inside the logarithm in (3.169) and the difficult derivatives in (3.174) have combined to lead us to an impasse. While it is possible to proceed with the EM algorithm here due to the differentiability of (3.174), we are faced with the computationally simpler method of maximizing the variational lower bound presented in § 3.4. Therefore, to keep the computations reasonably simpler, we will sidestep these problems through the variational approximation to the marginal log-likelihood. An EM algorithm for this approximation was presented above in § 3.4.

3.10 Posterior Distribution of the Latent Variables

In this section, we compute the posterior distribution to ensure that the model is fully generative. Given the probability density function in (3.166), we may write the posterior probability density

function of the latent variables W using Bayes' Theorem (Theorem 1):

$$\begin{aligned} p(w|y,\theta) &= \frac{p(y,w|\theta)}{p(y|\theta)} \\ &= \frac{c_1 \exp\left(-\frac{1}{2\sigma^2}(y-f(w,\theta))^T(y-f(w,\theta))\right)\exp\left(-\frac{1}{2}w^Tw\right)}{c\left[|\sqrt{D_{w_1}}|\exp\left(-\frac{1}{2}(y-\mu_1)^TC_{y_1}^{-1}(y-\mu_1)\right)\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right)+\right.} \\ &\quad \left.|\sqrt{D_{w_2}}|\exp\left(-\frac{1}{2}(y-\mu_2)^TC_{y_2}^{-1}(y-\mu_2)\right)\left(\operatorname{erf}\left(\frac{s_2}{\sqrt{2}}\right)+1\right)\right] \\ &\quad c_2\left(\exp\left(-\frac{1}{2\sigma^2}((y-\mu_1)-B_1w)^T((y-\mu_1)-B_1w)-\frac{1}{2}w^Tw\right)I_{\Omega_+}(w)+\right. \\ &= \left.\frac{\exp\left(-\frac{1}{2\sigma^2}((y-\mu_2)-B_2w)^T((y-\mu_2)-B_2w)-\frac{1}{2}w^Tw\right)I_{\Omega_-}(w)\right)}{I_{\mathfrak{P}_+}(y)P_1+I_{\mathfrak{P}_-}(y)P_2}, \end{aligned} \tag{3.175}$$

where

$$c_2 = 2^{q/2}\pi^{q/2-1}, \tag{3.176}$$

$$P_1 = |\sqrt{D_{w_1}}|\exp\left(-\frac{1}{2}(y-\mu_1)^TC_{y_1}^{-1}(y-\mu_1)\right)\operatorname{erfc}\left(\frac{s_1}{\sqrt{2}}\right), \tag{3.177}$$

$$P_2 = |\sqrt{D_{w_2}}|\exp\left(-\frac{1}{2}(y-\mu_2)^TC_{y_2}^{-1}(y-\mu_2)\right)\left(\operatorname{erf}\left(\frac{s_2}{\sqrt{2}}\right)+1\right), \tag{3.178}$$

and

$$I_X(x) = \begin{cases} 1, & x \in X, \\ 0, & x \notin X. \end{cases} \tag{3.179}$$

The analytic expression in (3.175) is theoretically useful, but in practical applications we will use MCMC simulations to approximate the posterior distribution.

3.11 Examples

3.11.1 Example 1 - 3D Hinge to 2D

In the first experiment, we begin with a 'hinged' 2-dimensional Gaussian distribution. To create the dataset, we first generated $N = 500$ data points from the 2-dimensional Gaussian distribution

defined by $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. These points represent the latent space. The 'hinge' was then formed by taking the absolute value of the x-coordinates to generate z-coordinates. Gaussian noise was sampled from $\mathcal{N}(0, 0.1)$ and added to every data point. The dataset is visualized in Figure 3.1.

Figure 3.1: Original positions of the hinged Gaussian dataset. Colors are added for visualization purposes.

The 2-partition piecewise PCA model (PWPCA) from § 3.2 was fitted to the observations. A classical PCA model (PCA) was also fitted to the observations for comparison.

Initializing the Parameters

The prior isotropic covariance parameter σ^2 was initialized by sampling a single real number from $\mathcal{U}[0, 1)$. The two model mean parameters, μ_1 and μ_2, were both initialized as zero arrays. The mapping parameters, B_1 and B_2, both had their elements initialized by sampling from the univariate standard Gaussian distribution $\mathcal{N}(0, 1)$.

In order to initialize the variational means, a Multivariate Adaptive Regressive Spline (MARS, [25, 34]) model with a single hinge point was fitted to the observations using Py-earth [65]. This

approach models the observations as a linear combination of piecewise linear basis functions in the same way as a `ReLU` neural network does, but with far fewer parameters. The basis functions were chosen as

$$h_1(x) = \begin{cases} x - t, & x > t, \\ 0, & \text{else}, \end{cases} \tag{3.180}$$

$$h_2(x) = \begin{cases} t - x, & x < t, \\ 0, & \text{else}, \end{cases} \tag{3.181}$$

where t is the hinge point determined by the fitting algorithm. The piecewise linear model used for initialization was then given by

$$f_0(Y) = b_0 + \sum_{k=1}^{2} b_k h_k(Y). \tag{3.182}$$

The variational means given by $\{m_i\}$ then had their elements initialized by the resulting local coordinates of each piece of f_0. The latent covariance matrices given by $\{\Sigma_i\}$ were all initialized as $\boldsymbol{I}_q$.

Optimizing the Parameters

An EM algorithm using the variational lower bound on the log-likelihood from § 3.4 was used to find the optimal model and variational parameters. Since the observations were centered at the origin, the intercept parameters μ_1 and μ_2 were not fitted. As discussed above, the critical points for the derivatives of the variational parameters have no closed form. Therefore, we used gradient descent with a step size of 1e-2 and a squared L_2-error tolerance of 1e-3 in their respective optimizations. The overall EM algorithm was given an error tolerance of 1e-3 for convergence of the relative error in log-likelihood.

Simulations

Once the optimal model parameters were found, the posterior distribution was approximated using MCMC simulations. The MCMC simulations were implemented using Just Another Gibbs Sam-

pler (JAGS) [61] and PyJAGS [54]. Latent points were then sampled from the posterior distribution and mapped to predictions in the observation space using the fitted model parameters.

Example 1 - Results

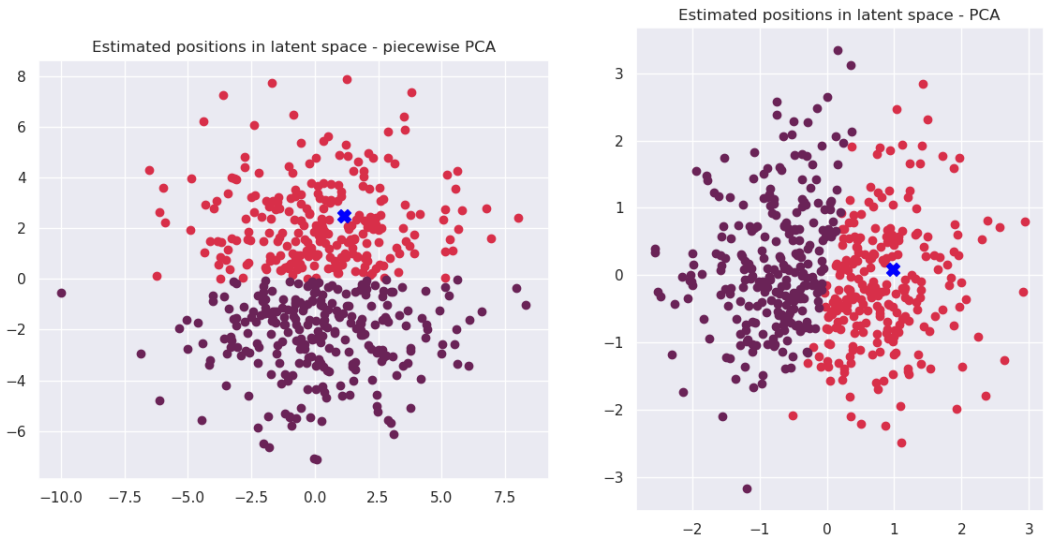


Figure 3.2: Results from applying PWPCA (left) and classical PCA (right) to the hinged Gaussian dataset. Colors are added for visualizations purposes.

The two-dimensional latent representations of the hinged Gaussian dataset using both PWPCA and PCA are shown in Figure 3.2. The resulting latent dimension of the PCA model is displayed with an arbitrary orthogonal rotation, an aspect of PCA previously discussed in § 2.5. In the PCA result, the latent points are more densely clustered along the center of the data than their original density across the surface of the hinge. This is due to the prior assumptions of classical and probabilistic PCA, where the geometry of the latent manifold is not taken into account. However, the PWPCA model generated latent representations of the observation points without spatial distortions. It appears that the scale along the latent axes was increased in the PWPCA model, whereas PCA has more closely preserved the scale due to its orthogonal projection of the three-dimensional points onto a two-dimensional subspace.

After a full-rank classical PCA was performed on the resulting latent dimensions, the total proportion of variance explained by PWPCA was 88.0%: 46.5% of the variance was explained by the

first latent dimension and 41.5% was explained by the second latent dimension. By contrast, the PCA model on the raw observations explained 83.9% of the sample variance. Thus, the PWPCA model explained more of the sample variability than the linear mappings produced by PCA.

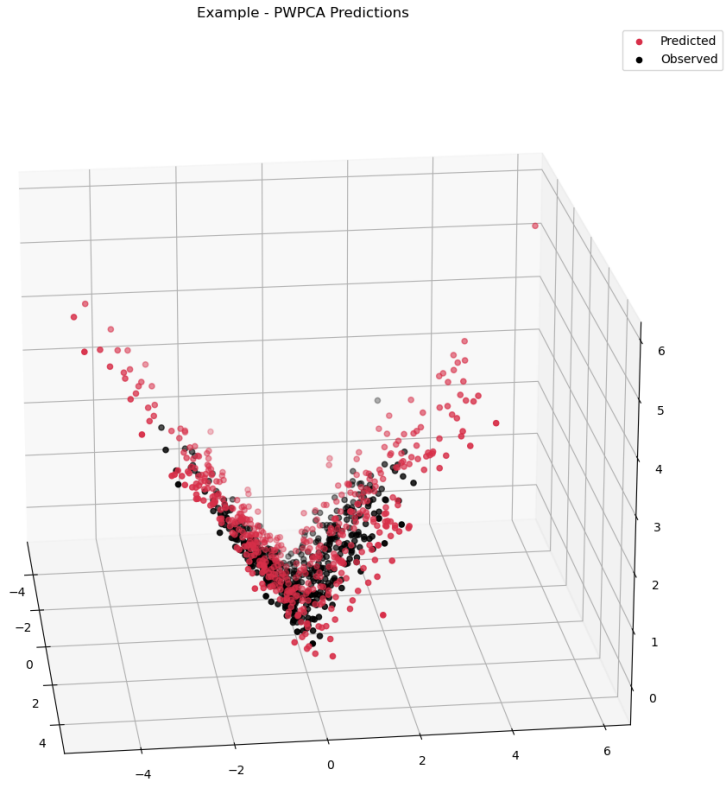


Figure 3.3: PWPCA predictions in the observation space generated from simulated latent points.

After the model parameters were fitted, the posterior distributions of each model were approximated using MCMC. Latent points were generated from the posterior distributions and then mapped to the observation space using the model parameters. A graphical comparison was used to check the predictive ability of each model. In the case of PCA, the simulated observations did not match the dataset well, as the predictions fell on a linear subspace of the observation space. The predictions produced by PWPCA matched the pattern observed in the dataset almost perfectly. The predictions generated by both models can be seen in Figures 3.3 and 3.4.

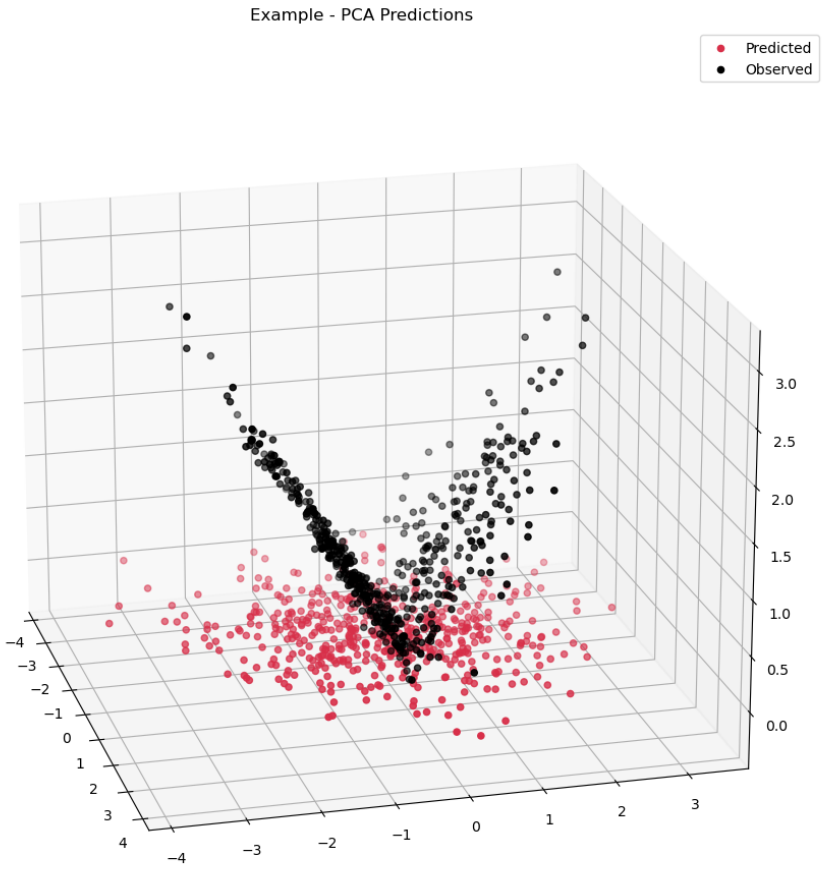


Figure 3.4: PCA predictions in the observation space generated from simulated latent points.

Chapter 4

Bibliometric Study

4.1 Introduction

Children's Health and Wellness (CHW) research is crucial to the Canadian research landscape. Scholars affiliated with CHW research institutes come from diverse backgrounds and work in a broad spectrum of roles, from clinicians to theoreticians. Scholarly output has historically been measured by the number of publications produced, the number of citations generated by a publication, or some function of both variables such as the h-index and its variants. However, basing the evaluation of scholarly productivity solely on document counts or citations is problematic for several reasons. For example, not all citations are made with the same intention [12], and citation counts may be artificially inflated through self-citation practices [31, 2]. Furthermore, document counts are sometimes augmented through the practice of 'salami slicing' - the superfluous partitioning of one piece of work into several smaller papers [76] to boost the overall number of publications. To measure all scholarship on the basis of a single metric runs the risk of obscuring other aspects of a researcher's work, such as collaborative reach.

Thus, in this chapter we introduce a novel application of dimensionality reduction for publication data. The collection of variables that arise from indexing academic publications is known as 'bibliometrics'. The purpose of the present work is to create two-dimensional scholarly profiles for Canadian CHW researchers based on 16 bibliometric variables. To our knowledge, this is the first use of linear dimensionality reduction techniques with bibliometric data.

We acknowledge the range of interpretations that many of these variables take. Since no current method exists for disambiguating intention, we use the variables at face value.

4.2 Literature Review

Several methods have been proposed to quantify scholarly output and research quality. A survey of the early development of bibliometrics is presented in [47]. In the early 20th century, Lotka proposed a frequency distribution for document counts based on abstracts from the natural sciences [50]. The resulting distribution was found by plotting published document counts against the number of authors producing the quantities of documents (both log-transformed). Using this empirical technique, Lotka concluded that the distribution of authorship roughly followed an inverse power law. In [11], Bradford proposed a power law to model the dispersion of documents published in different subject areas. An inverse power law to explain the relative frequencies of common and rare words in a body of text was introduced by Zipf in [85]. The Zipf distribution was later generalized using information theory by Mandelbrot in [51]. The Zipf and Mandelbrot distributions were first applied to bibliometrics by Buckland and Hindle [13]. A survey of empirically determined distributions for bibliometric random variables, including the so-called Bradford-Zipf-Mandelbrot distributions, is found in [24].

It was proposed in [62] that the above bibliometric distributions can be generalized to a distribution proportional to the one-parameter beta function. Lotka's distribution was empirically evaluated in [4], where the authors identified what they called an 'accumulative advantage' in scholarly output and citation trends. The accumulative advantage was a hypothesized effect that occurred when recognition of a scholarly work, whether through social feedback, citations, or funding opportunities, led to further increases in research productivity. In essence, early research recognition was believed to set a productivity expectation which then motivated scholars to continue producing output. The authors found a strong linear relationship between the respective Gini indexes [1] of document counts and citations and the number of years since completion of the PhD degree. More recently, Stewart proposed in [72] that the Poisson distribution may be used to model the probability densities of bibliometric variables.

[1] The Gini index was first defined in [29] as the arithmetic mean of differences between several measurements [30, 49, 15].

Goldstein and Spiegelhalter [31] proposed simple mathematical models for the evaluation of institutional outputs. Their hierarchical models were used to model the variance of test scores and health outcomes, but may also be applied to bibliometrics. Though the authors developed a numeric method for evaluating outputs, they also argued that the information obtained from such metrics should not take too much weight in evaluating the entities from which the metrics came. In 2005, Hirsch proposed the h-index [36]: the number of documents that have at least h citations for a given scholar. Hirsch argued that his metric was more accurate for evaluating research output than document count, citation count, or citations per document. Molinari and Molinari [57] developed a metric for comparing productivity between institutions or groups. This approach assumed that citations act as measures of the quality of a given scholarly work. By assuming that the h-index was governed by a power function, the h_m-index was then defined as the vertical intercept of the associated power function. Waltman and van Eck determined that, as of 2011, there were at least 37 variants of the h-index [80]. In their resulting critique of the h-index, Waltman and van Eck demonstrated that when relative or absolute performance of two scholars changes by the same amount, the corresponding changes in respective h-indexes leads to a completely different ranking of the scholars with respect to each other. The authors suggested replacing the h-index with the 'highly cited publications' indicator, which was defined as the number of documents exceeding a predetermined number of citations. A more recent attempt to aggregate several measures of research output was made in [42], which made use of Scopus - the extensive citation index provided by the publisher Elsevier. The proposed composite metric for a given scholar was computed as the unweighted sum of six bibliometric variables: total citations, Hirsch's h-index [36], a weighted variant of the h-index, and the number of citations as single author, single+first author, or single+first+last author. Each variable was log-transformed and scaled by its maximum, and the composite metric was computed both with and without self-citations included.

With the increased availability and use of bibliometrics to measure scholarly output, there have also been recent calls to acknowledge the limitations of such methods and to avoid relying wholly

on bibliometrics for the evaluation of scholars, research groups, or departments. Garfield [26] discussed the problematic citing of incorrect or obsolete data - a phenomenon which indicated that not all citations are made with the same intention [26]. Further evidence for the complex motives behind citations was presented by Brooks in [12]. In his survey of 437 citations made by 20 scholars, Brooks found that over 70% of the citations had complex motivations. Citations may be made to convince the reader of a fact through past work, borrow conclusions or results, assign positive or negative credit, give further readings to the reader, show how up-to-date the author(s) were, cite friends, and for perceived social consensus. Self-citation practices were verified in [42], where it was found that over 1% of the top cited scholars have over 40% self-citations, and that approximately 8.5% have over 25% self-citations. This indicated the concerning possibility that scholars may have artificially inflated their citation counts by continually and unnecessarily citing their own works. Goldstein and Spiegelhalter [31] further noted that measurement and comparison of a scholar or scholarly group's performance can lead to incentivizing the increase of the measurements. In such ways, scholars could try to 'game the system' to make their own evaluations more positive. In view of the potential for misuse of bibliometrics, Martin [53] advocated for the complementary inclusion of peer review when evaluating the work of scholarly entities. Martin posed that bibliometrics only partially explain the production of scientific knowledge: they do not tell the social, political, or career motivations behind scholarly output [53]. Lee and Bozeman [48] argued that factors such as collaboration, citizenship, gender, funding, and other demographic differences should not be ignored when attempting to measure research productivity. In their study, which measured productivity by document count, Lee and Bozeman found that productivity was positively correlated with collaboration, but not when productivity was measured by divided credit. The authors of [2] argued strongly against using journal impact factors and Hirsch's h-index (see [36]) for comparing any scholarly output. This paper recommended that any statistical information should be informed by correct statistical knowledge and practices. A set of recommendations for the continued use of bibliometrics was established by journal publishers and editors in the San

Francisco Declaration on Research Assessment (DORA) [14]. In particular, the authors of DORA proposed eliminating the journal impact factor, a field-specific metric with an opaque computation process, as a means of evaluating scholars. The authors of the Leiden Manifesto [35] presented a list of 10 guidelines for the continued development and evaluation of research metrics. Their 10 points were formed with the specific goal of avoiding rankings based on arbitrary or inaccurate data. A core thread running through all points is that quantitative metrics should not supersede expert knowledge in research evaluation, usually in the form of objective peer review. The Leiden Manifesto was a call for transparency in the development of research metrics and against the use of black-box evaluation processes. Stephan, Veugelers, and Wang [71] argued that a reliance on bibliometrics for the assessment of scholars led to less risk taking and less ambitious research in favor of consistent publication in journals with high impact factor. All of these documents signal a shift from strictly quantitative metrics to allow for the inclusion of quality-based assessments, and from metrics to content.

4.3 Limitations

Our method makes no assumptions about the quality of the scholarly works analyzed, we only present a method for systematically aggregating easily obtained bibliometric measurements. We therefore acknowledge that many other factors contributing to the production of scholarly output have not been included in our analysis. Furthermore, our model does not distinguish between the various motivations behind citations or the individual efforts or work ethics of the authors. Not everything that is measured counts, and not everything that counts can be measured.

4.4 Methods

4.4.1 Variables

Our dataset was composed of bibliometric measurements for scholars affiliated with eight participating children's hospital research institutes across Canada ($N = 2230$). Current membership rosters were obtained from: Alberta Children's Hospital Research Institute (ACHRI), BC Children's Hospital Research Institute (BCCHR), Children's Hospital Research Institute of Manitoba (CHRIM), CHU Sainte-Justine (CHUSJ), Holland Bloorview Kids Rehabilitation Hospital (HBV), McMaster Children's Hospital, the Hospital for Sick Children (SickKids), and Women and Children's Health Research Institute (WCHRI). The data was downloaded using the Scopus API between December 14, 2021 and Janurary 11, 2022 via

https://dev.elsevier.com/documentation/ScopusSearchAPI.wadl.

All variables considered were counted or computed over the 7-year period 2015-2021 (inclusive). Observations were composed of 16 different variables, with 9 discrete count variables and 7 continuous variables. The 16 variables used in this analysis can be grouped into 4 categories:

- **Document-level metrics**: publication count, published peer-reviewed article count, open access document count, publication count as first author, and publication count as last author.
- **Citation metrics**: citation count (number of times this author's publications have been cited in the text of another publication), cited-by count (number of documents citing this author), h-index, mean citations per publication, mean citations per year, and estimated linear citation growth rate over the 7-year period.
- **Collaborative metrics**: total count of coauthors, mean number of authors in authorship lists, and mean authorship position.

- **Journal metrics**: mean Scientific Journal Ranking (SJR) of all publications and mean journal h-index.

4.4.2 Models

In our approach, we apply the dimension reduction methods from Chapter 2 and Chapter 3 to bibliometric data. The models are: classical PCA using only count variables (PCA-C), classical PCA with all 16 variables (PCA), probabilistic PCA (pPCA) with all 16 variables, piecewise probabilistic PCA (PWPCA) with all 16 variables, Poisson PCA using only count variables (PLNPCA), kernel PCA using a Gaussian kernel with scale parameter $\frac{1}{16}$ and with all 16 variables (KPCA), an autoencoder with 1 hidden layer (AE1), and an autoencoder with 2 hidden layers (AE2). For the autoencoder models, we constructed the encoder and decoder sections symmetrically. The `ReLU` activation function was used in all autoencoder layers, and the dimensions of the between-layer affine transformations for AE1 and AE2 are given in Table 4.2. The number of parameters for each PCA model is given in Table 4.1.

The general model for dimension reduction is written as functional relationship between two sets of variables. First, in all cases except PLNPCA, we assume that our observations $\{Y_i\}_{i=1}^N$ lie in a p-dimensional real space. In the case of PLNPCA, the variables lie in $\mathbb{N}^p$. We then associate to each observation a latent point W_i in a lower-dimensional space $\mathbb{R}^q$, $q < p$. We assume that these latent variables have a Gaussian prior distribution with zero mean and unit covariance, and that the observations follow a Gaussian distribution with isotropic covariance when conditioned on the latent variables. In summary, the general model is

$$\begin{aligned} \{W_i\}_{i=1}^N \subset \mathbb{R}^q, \qquad & W_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \\ \{Y_i\}_{i=1} \in \mathbb{R}^p, \qquad & Y_i = \mu + BW_i + \varepsilon =: f(W_i, \theta) + \varepsilon, \\ p(Y_i|W_i, \theta) = (2\pi\sigma^2)^{-p/2} & \exp\left(-\frac{1}{2\sigma^2}(Y_i - f(W_i,\theta))^T (Y_i - f(W_i,\theta)) \right), \end{aligned} \tag{4.1}$$

where σ^2 is the observation covariance parameter and $\theta = \{\mu, B\}$ is the collection of model parameters. In the PCA-C and PCA models, we assume $\sigma^2 \to 0$, since these models only assume

the existence of a linear relationship between the observations and the latent variables. In pPCA, the full model (4.1) is used. The PWPCA model assumes that the dataset may be partitioned into two subsets, where the separating hyperplane lies along a single dimension as in Chapter 3. The PLNPCA model only changes the form of the distributions on the observations to independent Poisson distributions over each dimension. That is, for the PLNPCA model,

$$p(Y_{ij}|W_{ij},\theta) = \frac{1}{Y_{ij}!}\exp\left(Y_{ij}\cdot\log f(W_{ij},\theta) - f(W_{ij},\theta)\right). \tag{4.2}$$

For the KPCA model, affinities between observations were computed using the Gaussian kernel

$$\exp\left(-\frac{1}{16}||Y_i - Y_j||_2^2\right), \tag{4.3}$$

and then standard PCA was completed on the covariance matrix of the resulting affinities.

The singular value decomposition (SVD) method was used to fit the PCA-C, PCA, and KPCA models (see [43, 59]), and Expectation-Maximization (EM) algorithms were used to fit the pPCA, PWPCA, and PLNPCA models ([75], [16]).

Table 4.1: Number of parameters for the PCA-based models: PCA-C, PCA, PWPCA, PLNPCA, and KPCA.

Model	**Number of parameters**
PCA-C	18
PCA	32
PWPCA	64
PLNPCA	18
KPCA	1

4.4.3 Transforming Variables

PCA-C, PCA, pPCA, and PWPCA all require the input variables to be on similar scales in order to ensure that one variable cannot dominate the resulting principal eigenvectors. Similarly, autoencoders require scaled variables to avoid exploding gradients during training. Since the count

Table 4.2: Architectures for the autoencoder models. The additional parameters at each layer are a result of including intercepts.

Model: AE1		
Layer	**Transformation shape**	**Number of parameters**
Input layer	16	
Encoder hidden	$(16+1)\times 8$	136
Encoder output	$(8+1)\times 2$	18
Decoder input	2	
Decoder hidden	$(2+1)\times 8$	24
Decoder output	$(8+1)\times 16$	144
	Total parameters	322

Model: AE2		
Layer	**Transformation shape**	**Number of parameters**
Input layer	16	
Encoder hidden 1	$(16+1)\times 16$	272
Encoder hidden 2	$(16+1)\times 8$	136
Encoder output	$(8+1)\times 2$	18
Decoder input	2	
Decoder hidden 1	$(2+1)\times 8$	24
Decoder hidden 2	$(8+1)\times 16$	144
Decoder output	$(16+1)\times 16$	272
	Total parameters	866

variables were highly right skewed and varied in scale, we applied the log transform $\log(x+1)$. All continuous variables were similarly right-skewed and were log transformed with $\log(x+1)$. For the PLNPCA model, where the input data are assumed to be Poisson distributed and thus skewness poses no issue, the raw counts were used as inputs. Histograms of all variables before and after the log transformations can be found in Appendix B. For the PCA-C and PCA models, the empirical histograms of the log-transformed variables appeared approximately Gaussian, and so z-score

standardization was applied. That is, all log-transformed variables were scaled by: $z = \frac{y-\bar{y}}{s}$, where $\bar{y}$ is the empirical mean and s is the empirical standard deviation.

A piecewise linear relationship was observed between the log-transformed mean journal h-index and mean SJR (see Figure 4.1), so the log-mean SJR was selected as the response variable in the initializing MARS model, as in § 3.11. The resulting MARS model correctly selected the log-mean journal h-index as the dimension along which the hinge point should be placed. Thus, the variational means for the PWPCA model were initialized by a piecewise affine model with the nonlinearity restricted to the above identified dimensions. The pairwise relationships between all log-transformed variables are displayed in Figure 4.2.

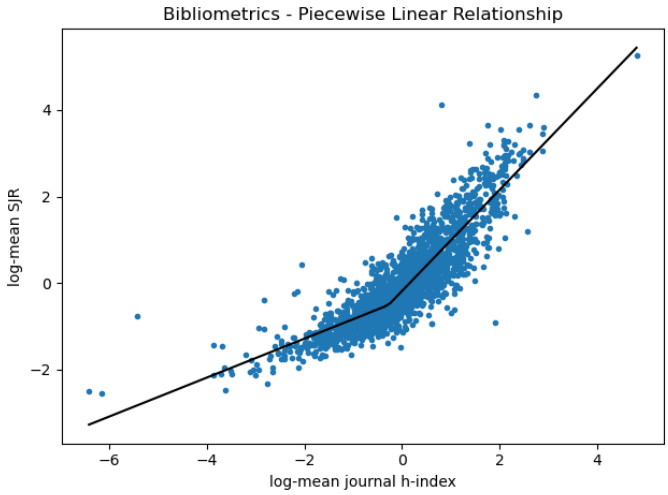
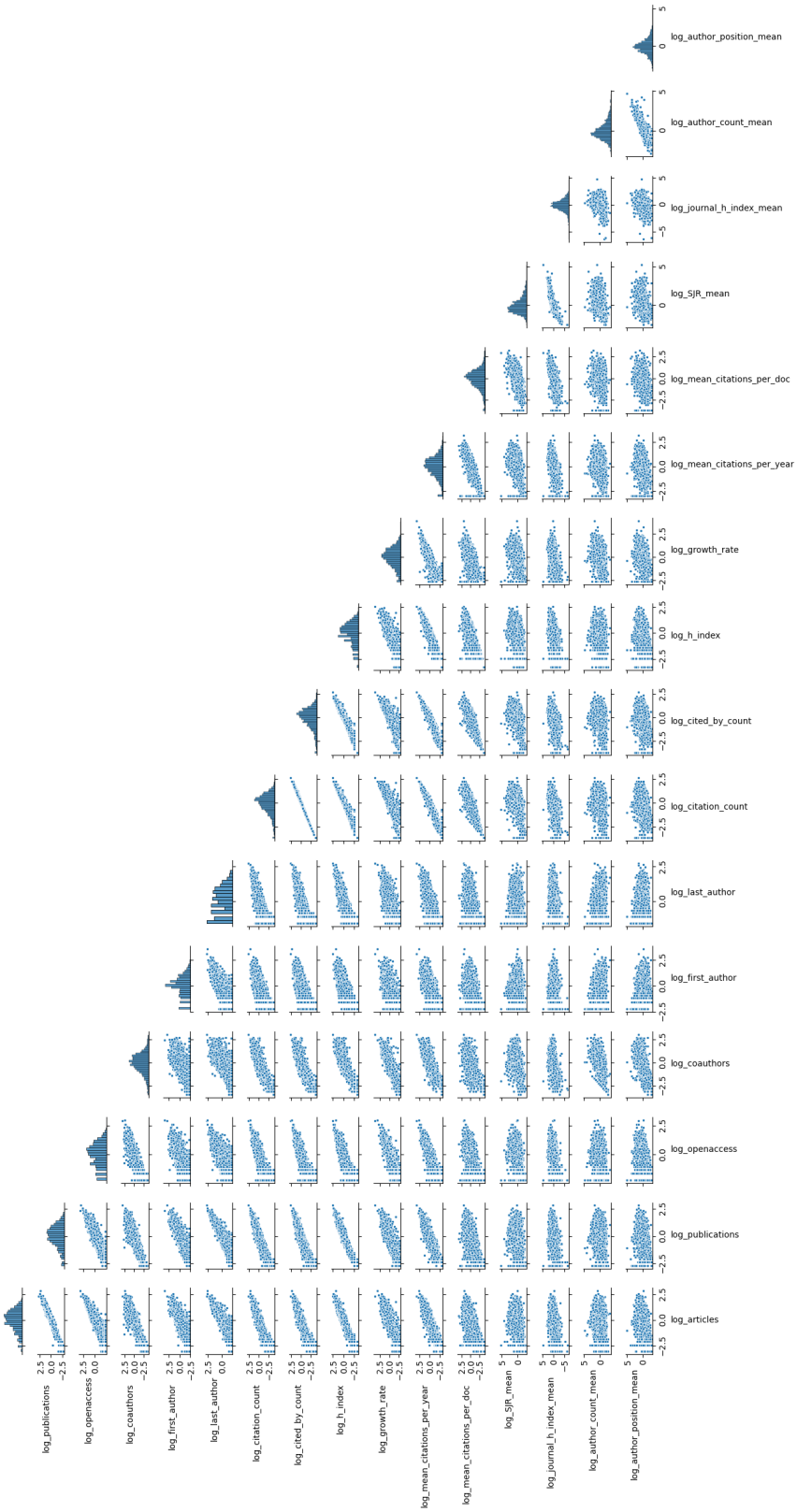


Figure 4.1: Piecewise linear relationship between the log-mean journal h-index and the log-mean SJR.

Figure 4.2: Pairwise scatterplots of all log-transformed bibliometric variables.



4.4.4 Model Comparison

For the models based on the eigenvector decomposition of the empirical covariance matrix (PCA-C, PCA, and pPCA) and the EM-optimized models (PLNPCA and PWPCA), we computed the variance explained by the resulting principal eigenvectors and the cumulative explained variance. For all models except KPCA and the autoencoders, the linear contributions of the input variables to the principal eigenvectors are called the 'loadings'. We created plots for the loadings of the principal eigenvectors of every such model.

In order to develop an easily visualized 2-dimensional scholarly 'profile', the latent dimension was chosen as $q = 2$. However, the PCA-C, PCA, and PLNPCA models provided more principal eigenvectors than were needed for this profile. Therefore, we mapped the observations to a two-dimensional latent space by using the first two principal components of each model. Graphical plots were then made of the mapped observations in the respective latent spaces for visual comparison.

4.4.5 Software

With the help of several software packages, the R programming language [63] was used to fit all linear PCA models. The R package `pcaMethods` was used for fitting classical and probabilistic PCA [70]. We used the R package `PLNmodels` to fit the Poisson PCA model [16]. Further R packages used were `readxl` for loading the data [83], `ggplot2` and `reshape` for plotting [82, 81], and `factoextra` for extracting the eigenvalues from the models [44]. The Python package `sklearn` was used to fit the KPCA model [59]. The Python programming language was also used to fit the autoencoders: the Python package `keras` was used with `tensorflow` as a backend [18, 1]. JAGS [61] and PyJAGS [54] were used to simulate the posterior distributions. All code and data files are available at

`https://github.com/uniformlyMatt/compare-dim-reduction`.

Custom Python software was written to fit the PWPCA model and is available at

`https://github.com/uniformlyMatt/piecewise-pca`.

4.5 Results

The PCA and pPCA models resulted in identical principal components and percentages of explained variance. Therefore, we only present the results from the PCA model unless otherwise specified. For all models, we will refer to the principal components in order of their explained variance. For example, the principal component which explains the highest proportion of sample variance is p_1, and the component with the next highest explained variance is p_2.

The PCA-C model accounted for 92.0% of the sample variance with its first two principal components (86.4% in p_1 and 5.6% in p_2). The explained sample variance was higher for the PLNPCA model, which accounted for 96.1% of the sample variance with the first two components (91.6% in p_1 and 4.5% in p_2). The PCA model, which used all 16 input variables, accounted for 88.3% of the variance in the first two components (81.8% in p_1 and 6.5% in p_2). Finally, the PWPCA model explained 99.5% of the sample variance (81.7% in p_1 and 17.8% in p_2). A table of the variance explained per component and the cumulative explained sample variance for the first five components of all linear or piecewise linear models is given in Table 4.3.

Strong positive correlations were observed among the first principal components of all linear models. In addition, a strong negative correlation was observed between p_1 of the KPCA model and p_1 of each of the linear models. Moderate positive correlations were observed between p_1 of each of the linear models and the first latent dimension of the AE2 model, and a moderate negative correlation was observed between p_1 of the KPCA models and the first latent dimension of AE2. The first latent dimension of AE1 had weak correlations with all other models. The PWPCA model showed weak positive correlations between its associated p_1 and those of the linear models, and a weak negative correlation with p_1 from KPCA. Correlations between the first principal components and latent dimensions of all models is given in Table 4.4.

In all linear models, the effect of each variable on the first principal component was positive.

Table 4.3: Variance explained by the first five principal components of the PCA-C, PCA, and PLNPCA models, with the cumulative variance explained.

	PC1	PC2	PC3	PC4	PC5
PCA-C					
Variance explained	86.4%	5.6%	3.4%	2.2%	1.5%
Cumulative variance explained	86.4%	92.0%	95.4%	97.6%	>99%
PCA					
Variance explained	81.8%	6.5%	4.3%	2.1%	1.6%
Cumulative variance explained	81.8%	88.3%	92.6%	94.7%	96.3%
PLNPCA					
Variance explained	91.6%	4.5%	2.9%	<1%	<1%
Cumulative variance explained	91.6%	96.1%	99.0%	>99%	>99%
PWPCA					
Variance explained	81.7%	17.8%	-	-	-
Cumulative variance explained	81.7%	99.5%	-	-	-

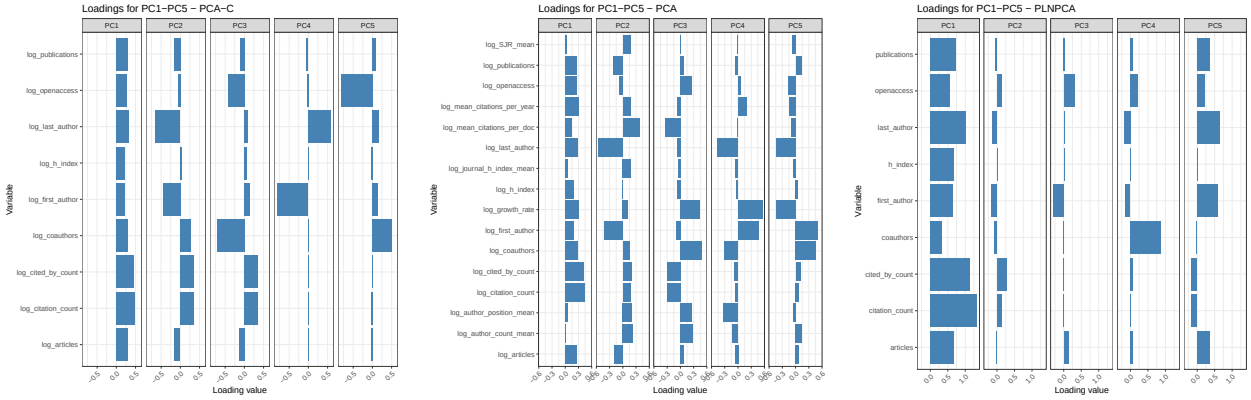


Figure 4.3: Loadings for the three fitted linear models: PCA-C (left), PCA (middle), and PLNPCA (right).

In the PCA-C model, all variable loadings on p_1 were of similar magnitude, with citations and cited-by counts having the largest loadings. This pattern was also seen in the PLNPCA model. There was more variety in the magnitudes of the p_1 loadings in the PCA model, where the highest loadings on p_1 were seen from the log-transformed count variables, the log-transformed mean

citations per year, and the log-transformed growth rate. The loadings for the principal components that explained progressively less variance were also more varied in their sign in all models.

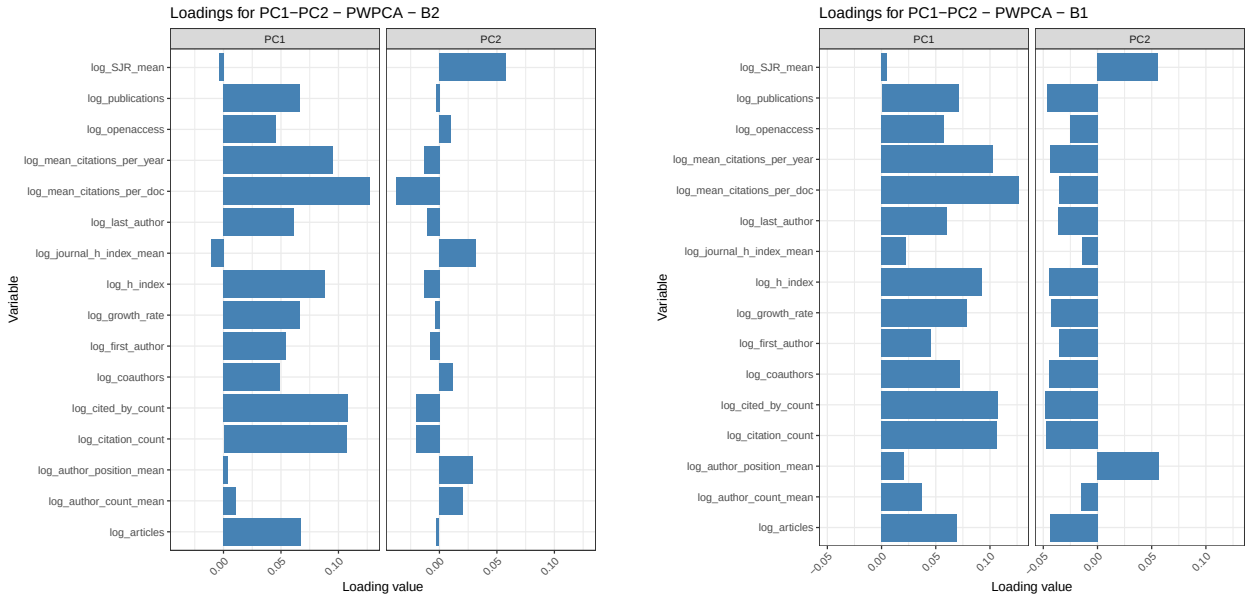


Figure 4.4: Loadings for the two components of the PWPCA model. Loadings for B_2 (left) correspond to points to the 'left' of the one-dimensional hinge point, and loadings for B_1 (right) correspond to points to the 'right' of the one-dimensional hinge point.

For the PWPCA model, the loadings are split between two subsets of the data: those points falling to the right of the hinge point (where the hinge point is placed along the log-mean journal h-index) and those falling to the left. For convenience, we will refer to these two subsets of the dataset as S_r and S_ℓ, respectively. The p_1 loadings for points in S_r were all positive. The same

Table 4.4: Correlations between first latent dimension scores of all models.

	PCA-C	PCA	PLNPCA	PWPCA	KPCA	AE1
PCA	0.996	-	-	-	-	-
PLNPCA	0.998	0.994	-	-	-	-
PWPCA	0.211	0.202	0.213	-	-	-
KPCA	-0.826	-0.828	-0.810	-0.183	-	-
AE1	0.303	0.258	0.299	-0.104	-0.273	-
AE2	0.633	0.669	0.618	-0.004	-0.614	-0.243

pattern was observed in the loadings for p_1 in S_ℓ, with the exception of the negative loadings for log-mean SJR and log-mean journal h-index. Noting this exception, all positive loadings on both data subsets had approximately the same magnitudes.

The PCA-C and PCA models shared a pattern in their p_2 loadings. In both models, all p_2 loadings for the document-level metrics were negative, and all p_2 loadings for collaborative and journal-level metrics were positive. Additionally, the loadings for citation metrics were positive except for the loadings for h-index. Citation-level metrics also had positive loadings on p_2 for the PLNPCA model, but the pattern observed in the PCA-C and PCA loadings for the other categories of input variables was not observed. The PWPCA model showed different p_2 loadings than the linear models. For the PWPCA model over S_r, only log-SJR mean and log-mean author position were positively loaded on p_2, with all other variables loaded negatively. Over S_ℓ, all journal- and collaborative-level variables were positively loaded on p_2, all citation-level variables had negative p_2 loadings, and document-level variables had loadings of varied sign on p_2. Plots of the loadings for the first five principal components of the PCA-C, PCA, and PLNPCA models are found in Figure 4.3. The loadings for the two components of the PWPCA model on both data subsets S_ℓ and S_r are shown in Figure 4.4. Since the parameters from the autoencoder models have no geometric or physical interpretation, we refrain from reporting their variable loadings.

We estimated the positions of the observations in the two-dimensional latent space for all of the models. The resulting latent space representations for the linear models are shown in Figure 4.6, the PWPCA latent space is visualized in Figure 4.5, and the results from the KPCA and autoencoder models are shown in Figure 4.7. The relative densities of the respective latent points generated by all models are displayed along with the scores from PC1 and PC2. The PLNPCA model shows the greatest dispersion along the vertical axis and a left skewed density along the horizontal axis. The latent spaces generated by the linear models appear superficially similar, with only slight differences in density. The latent spaces from all models except KPCA have similar scales along their horizontal and vertical axes.

When the points were mapped to the latent space by the PWPCA model, the spread of latent points appeared to approximate a two-dimensional Gaussian distribution. The partition in the latent space was formed by the horizontal axis. Thus, the upper half of the latent plane corresponds to observations with log-mean journal h-index greater than zero, and the lower half corresponds to the complement.

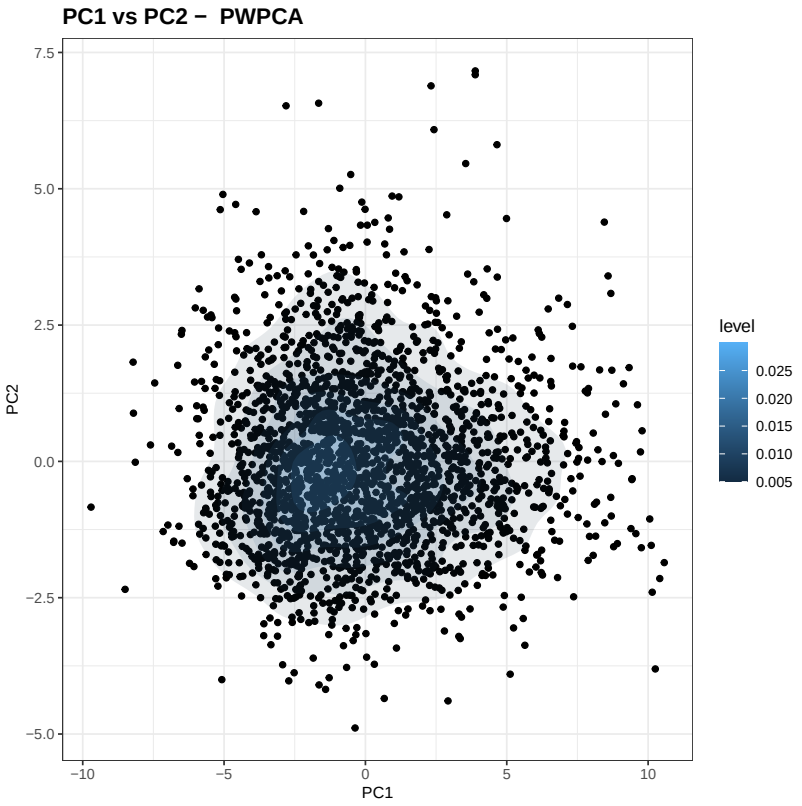


Figure 4.5: Observations mapped to a two-dimensional latent space using PWPCA.

The latent space generated by the KPCA model is markedly heart-shaped, with higher density around the outer boundary of the latent points. The latent spaces generated by the AE1 and AE2 models both appeared to accurately approximate two-dimensional Gaussian distributions, where more rotational symmetry was observed in the latent space generated by the AE2 model.

Predictions were generated for all generative models by mapping points from the simulated latent space to the observation space. Graphical comparisons were used to evaluate the quality of the predictions. Overall, each PCA-based model trained on all variables appeared to generate

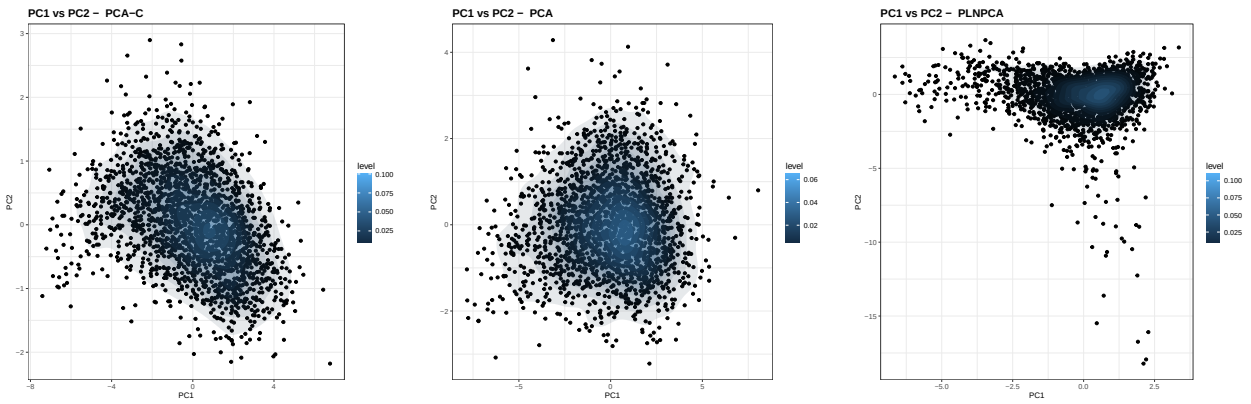


Figure 4.6: Observations mapped to a two-dimensional latent space using (left) PCA-C, (middle) PCA, and (right) PLNPCA.

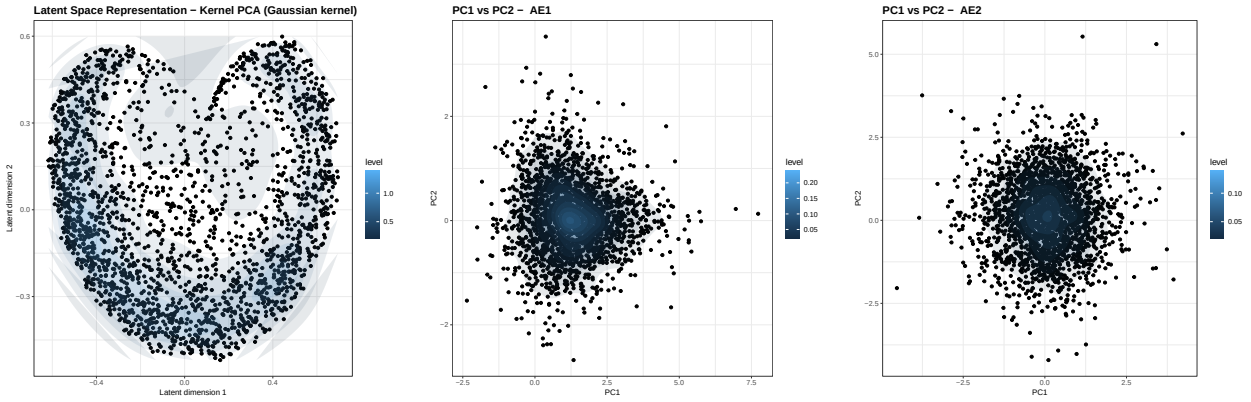


Figure 4.7: Observations mapped to a two-dimensional latent space using (left) KPCA, (middle) AE1, and (right) AE2.

predictions that matched the original observations, both in terms of their distribution and their pairwise relationships. However, only the PWPCA model produced predictions with a piecewise linear relationship between the mean journal *h*-index and the mean SJR (compare Figures B.7 and B.9). The PCA-C model was trained on only the log-transformed count variables, and generated predictions that accurately matched both the distributions and the pairwise relationships of the original observations (see Figures B.3 and B.4). On the other hand, the predictions produced by PLNPCA displayed strong linear pairwise relationships that were not observed between the original observations (see Figures B.5 and B.6). The predictions produced by the two autoencoder models did not accurately match the original observations, and can be seen in Figures B.10 and

B.11.

4.6 Discussion

The purpose of this study was to investigate the use of principal components analysis for generating two-dimensional scholarly 'profiles' from publication data. Since the bulk of the sample variance was explained by the first principal component in each of the three linear models, we'll focus our discussion on the possible interpretation of this component.

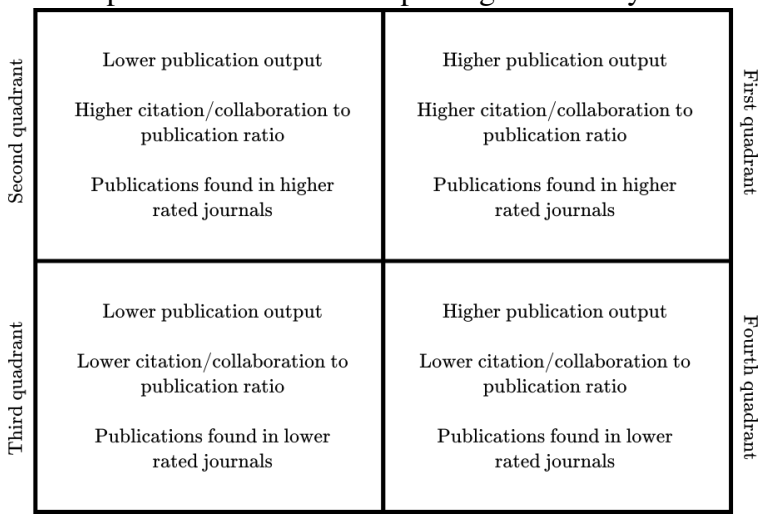
A strong positive correlation was observed between the p_1 scores of all linear models. Therefore, it seems that each model is measuring the same quantity with its first component. Of the three linear models, the model with the highest proportion of explained variance was PLNPCA. This model also possessed the highest drop in explained variance between the first and second principal components (87.1%). However, the PLNPCA model also weighted citations, cited-by counts, and last authorship more heavily in its p_1 loadings, and so using p_1 to express some measure of publication activity will favor authors with high citation counts and who publish in the last authorship position. By contrast, the PCA-C model showed a more balanced set of variable loadings on p_1 and was still able to explain approximately 86% of the sample variance. In other words, if p_1 is to be used as any measure of scholarly publication productivity and we consider all inputs as roughly equal, then the PCA-C model provides the most conservative approach with the fewest assumptions on the data. Not surprisingly, the output from the KPCA and autoencoder models yielded little real information. The latent space representations of both AE1 and AE2 show interesting patterns, but have no interpretable parameters.

Interestingly, the p_1 scores from the PWPCA model were not strongly correlated with those from any other model. However, the p_1 loadings were positive for all variables on S_r and similarly positive on S_ℓ with the exception of the journal-level metrics. Futhermore, the PWPCA model explained the highest proportion of variance among all models (99.5%). This is especially surprising since the PWPCA model included all 16 variables. Additionally, a substantial proportion

of variance (81.7%) was explained by the first component of PWPCA. Thus, it appears that the inclusion of a single nonlinearity in the model has led to a novel measurement of overall scholarly productivity which captures a high degree of variability in the data. It seems that PWPCA provides an advantage in explaining variability in the dataset, but that including a nonlinearity significantly changes the nature of the loadings on the principal components.

For the PCA-C and PCA models, all document-level inputs showed negative loadings on p_2. In contrast, journal-level, citation, and collaborative input loadings were positive on p_2 with the exception of the h-index. Thus, it seems that p_2 acts as a measure of the trade-off between the quantity of publications and the degree to which those publications are known to the larger community (including collaborators). Taking this view, we may interpret the four quadrants of the latent spaces generated by PCA-C and PCA as in Figure 4.8.

Figure 4.8: Interpretation of the latent spaces generated by PCA-C and PCA.



The PWPCA latent space has a slightly different interpretation. The horizontal axis can be interpreted in the same way as in the linear models, but the vertical axis displayed a different pattern than in the linear models. The p_2 loadings for PWPCA indicate that scholars in the upper half of the latent plane have higher SJR and mean author position, but all other variables are weighted negatively. This changes in the lower half of the latent plane, where the p_2 loadings show that authors have higher journal- and collaboration-level metrics, and lower citation-level metrics. Of

the document-level metrics, the lower half of the latent space positively weights open access but negatively weights overall document count, first and last authorship, and article count. Thus, the lower half of the PWPCA latent space appears to show a trade-off between open, collaborative output in highly ranked journals and overall document output.

It is not surprising that adding in more variables resulted in a lower percentage of explained variance between the PCA and PCA-C models. More variables should inherently lead to more variability. However, it is noteworthy that adding 7 continuous variables to the PCA model still accounted for 81.8% of the sample variance in p_1. This is likely due to some of the continuous variables being derived from the count variables, such as the mean citations per document, mean citations per year, and the mean coauthors per document.

The PWPCA model appeared to produce the most accurate predictions when using all transformed variables to fit the model. The predictions produced by PCA were overall accurate to the original observations, but the PCA model failed to capture the known nonlinearity in the data. When no nonlinearity was observed between variables, as was the case when using only the log-transformed count variables, PCA-C appeared to produce the most accurate predictions.

In light of these results, it appears that PWPCA is a preferable choice for the two-dimensional representation of the bibliometrics dataset when compared to linear models. When piecewise linear relationships are identified between input variables, the PWPCA model can effectively differentiate between data lying on the resulting partitions while still explaining a substantial proportion of the sample variance. PCA-C remains an effective choice when the data lie on a linear manifold.

4.6.1 Future Directions

The use of linear dimensionality reduction techniques with bibliometric data offers a convenient, interpretable framework for generating scholarly profiles. This study suggests more directions for further investigation. For example, the PLNPCA model can be extended in the future to include continuous covariates and offset parameters for unequal sampling regimes ([16]). Such extensions are attractive as they allow for both continuous and discrete data to be included in the same

dimension reduction scheme.

The PWPCA shows promise as a flexible method that can explain the variability in nonlinear datasets. The inclusion of more nonlinearities along a single dimension or one-dimensional nonlinearities in multiple variables should similarly lead to a high proportion of explained variance and should be investigated further.

Beyond the continuous variables used in the present study, the inclusion of PlumX metrics is another direction which could prove fruitful. PlumX metrics are complementary to citations in that they measure the number of times that a given scholarly publication is accessed, read, or downloaded online, or when the publication is mentioned in online blog posts, YouTube videos, tweets, and news articles.

4.6.2 Declarations

The classical PCA part of this study was completed for the Alberta Children's Hospital Research Institute (ACHRI) in support of the University of Calgary's 2022 Canada First Research Excellence Fund (CFREF) application, and in collaboration with Benedikt Hallgrimsson, Christina Hirota, and Susa Benseler. Matthew Adams is supported by an Alberta Innovates Graduate scholarship.

Chapter 5

Conclusion

This work has presented a full probabilistic model for computing piecewise affine dimensionality reduction on high-dimensional datasets. The resulting piecewise PCA (PWPCA) model only shows a linear increase in parameters, yet is applicable to a much wider range of datasets than classical PCA. Through the use of PWPCA, dimensionality reduction may be achieved on nonlinear datasets lying near low-dimensional affine subspaces, including datasets containing disjoint regions of such subspaces. In contrast to other nonlinear models such as kernel PCA and autoencoders, the model parameters of PWPCA are fully interpretable as simple affine transformations. The probabilistic framework adopted in this work resulted in a generative model similar to probabilistic PCA and variational autoencoders. We also established a method for estimating the underlying dimension of the lower-dimensional latent space through a Bayesian Information Criterion (BIC). Furthermore, visualization techniques were introduced for the approximate latent points produced by PWPCA. The ρ^2 measure of the log-likelihood accounted for by the piecewise PCA model was also defined and used in concert with classical PCA to quantify the proportion of variance explained by the PWPCA model.

Our model is useful for datasets where piecewise linear relationships are observed between pairs of variables or when the data may be partitioned into piecewise linear regions by parallel hyperplanes. The PWPCA model is also flexible in terms of the probability assumptions, and may be applied to any datasets when the observations are assumed to have a distribution from the natural exponential family, conditioned on the latent variables. This enables PWPCA to be flexible enough for both discrete and continuous data types under a single framework.

A comparison of dimensionality reduction methods for use in bibliometrics was also completed as part of this dissertation. From the results of this work, it appears that PCA models used for the

summarization of scholarly publication output are able to capture most of the variability in the sample. However, applying the PWPCA model to the bibliometric data provided the insight that nonlinearities in the dataset were not captured by linear PCA models. Thus, the nonlinearity of our model is crucial in modeling the underlying nonlinear manifold of the bibliometric data. In addition, PWPCA explained 99.5% of the sample variance while using all 16 bibliometric variables as input, a higher proportion of variance than any of the other models fitted. Therefore, PWPCA and its future variants should be the preferred method for reducing the dimensions of piecewise linear datasets.

5.1 Future Work

This work is the first step in bridging the gap between linear probabilistic dimensionality reduction methods and variational autoencoders, and offers many avenues for future consideration. First, we have so far only allowed nonlinearities along a single pair of dimensions in the observation space. This was largely done for computational convenience, but could be extended to dataset partitioned by nonlinearities in multiple dimensions. The resulting form of the multidimensional partitions would be quite similar to the case calculated here, but would involve more Gaussian error functions in the marginal and variational log-likelihood functions.

In the case of data containing multiple piecewise linear regions, we have assumed that the linear regions are separated by parallel hyperplanes. A possible future direction would be to remove the assumption that these separating hyperplanes are parallel and allow piecewise linear separating boundaries. Similarly, nonlinear boundaries could be approximated by piecewise linear segments and used in the partitioning of the dataset.

Another pressing matter is the estimation of the dimensionality of the lower-dimensional latent space. The present method is to compute the BIC for a range of dimensions $1 \leq j < p$, which requires fitting a separate PWPCA model for each j and then selecting the model with the highest BIC. This is an inefficient method for estimation, especially for large datasets. Therefore, further

investigation into more efficient dimensionality estimation methods must be undertaken in future iterations.

In the present work, we have only considered datasets which follow unimodal distributions from the natural exponential family. Mixture models allow multiple peaks in the data distribution, and may be a fruitful direction for investigation in the continued development of PWPCA. The current framework may also be applied to other forms of the multivariate Poisson distribution that allow the presence of discrete and continuous covariates. The variational approach taken in this work is extremely flexible, and may also lead to developments in versions of PWPCA for distributions outside of the natural exponential family.

Appendix A

General Results

A.1 Integrating a Gaussian Quadratic Form

Suppose that U is a q-dimensional random variable and $U \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_q)$, and let A be a $q \times q$ symmetric matrix. We will integrate

$$\int_{\Omega} u^T A u \exp\left(-\frac{1}{2} u^T u\right) du, \tag{A.1}$$

where $\Omega := \{u \in \mathbb{R}^q : \boldsymbol{e}_q^T u \geq s\}$, for some fixed $s \in \mathbb{R}$, with $\boldsymbol{e}_q$ the q^{th} standard basis vector of $\mathbb{R}^q$. Using Fubini's Theorem, we can compute the integral in (A.1) as an interated integral. That is,

$$\int_{\Omega} u^T A u \exp\left(-\frac{1}{2} u^T u\right) du = \int_s^{\infty} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u^T A u \exp\left(-\frac{1}{2} u^T u\right) du_1 \ldots du_{q-1} du_q. \tag{A.2}$$

Lemma 1. *Over* $\mathbb{R}^{q-1}$*, the integral given by* (A.1) *takes the form*

$$(2\pi)^{\frac{q-1}{2}} \left(\sum_{j=1}^{q-1} [A]_{jj} + [A]_{qq} u_q^2\right) exp\left(-\frac{u_q^2}{2}\right), \tag{A.3}$$

where u_q *is the* q^{th} *element of* u*.*

Proof.

The result follows by induction on j. For the base case, let $j = 1$. Then:

$$\begin{aligned}
I_1 &:= \int_{\mathbb{R}} u^T A u \exp\left(-\frac{1}{2}u^T u\right) du_1 \\
&= \int_{\mathbb{R}} \sum_{j,k\geq 1} [A]_{jk} u_j u_k \exp\left(-\frac{u_1^2}{2}\right) \exp\left(-\frac{1}{2}\sum_{j=2}^{q} u_j^2\right) du_1 \\
&= \exp\left(-\frac{1}{2}\sum_{j=2}^{q} u_j^2\right) \sum_{j,k\geq 1} [A]_{jk} \int_{\mathbb{R}} u_j u_k \exp\left(-\frac{u_1^2}{2}\right) du_1 \\
&= \exp\left(-\frac{1}{2}\sum_{j=2}^{q} u_j^2\right) \left[[A]_{11} \int_{\mathbb{R}} u_1^2 \exp\left(-\frac{u_1^2}{2}\right) du_1 + \right. \\
&\qquad 2\sum_{k=2}^{q} [A]_{1k} u_k \int_{\mathbb{R}} u_1 \exp\left(-\frac{u_1^2}{2}\right) du_1 + \\
&\qquad\qquad \left. \sum_{j,k\geq 2} [A]_{jk} u_j u_k \int_{\mathbb{R}} \exp\left(-\frac{u_1^2}{2}\right) du_1, \right.
\end{aligned} \tag{A.4}$$

where we have used the fact that A is symmetric to get the middle term. Note that the middle term in (A.4) is odd, and therefore integrates to 0. The other terms are standard one-dimensional Gaussian functions and integrate to $\sqrt{2\pi}$. Thus,

$$I_1 = \int_{\mathbb{R}} u^T A u \exp\left(-\frac{1}{2}u^T u\right) du_1 = \sqrt{2\pi} \exp\left(-\frac{1}{2}\sum_{j=2}^{q} u_j^2\right) \left[[A]_{11} + \sum_{j,k\geq 2} u_j u_k \right]. \tag{A.5}$$

For the inductive hypothesis, suppose that (A.5) holds for $j = J-1$. We show that it holds for J.

$$\begin{aligned}
I_J &= \int_{\mathbb{R}} I_{J-1} du_J \\
&= (2\pi)^{\frac{J-1}{2}} \int_{\mathbb{R}} \left(\sum_{j=1}^{J-1} [A]_{jj} + \sum_{j,k\geq J} [A]_{jk} u_j u_k \right) \exp\left(-\frac{1}{2}\sum_{j\geq J} u_j^2\right) du_J \\
&= (2\pi)^{\frac{J-1}{2}} \exp\left(-\frac{1}{2}\sum_{j\geq J+1} u_j^2\right) \left[\int_{\mathbb{R}} \left(\sum_{j=1}^{J-1} [A]_{jj} + [A]_{JJ} u_J^2 \right) \exp\left(-\frac{u_J^2}{2}\right) du_J + \right. \\
&\qquad \left. \sum_{k\geq J+1} [A]_{Jk} u_k \int_{\mathbb{R}} u_J \exp\left(-\frac{u_J^2}{2}\right) du_J + \sum_{j,k\geq J+2} [A]_{jk} u_j u_k \int_{\mathbb{R}} \exp\left(-\frac{u_J^2}{2}\right) du_J \right] \\
&= (2\pi)^{\frac{J-1}{2}} \exp\left(-\frac{1}{2}\sum_{j\geq J+1} u_j^2\right) \sqrt{2\pi} \left[\sum_{i=1}^{J-1} [A]_{jj} + [A]_{JJ} + \sum_{j,k\geq J+2} [A]_{jk} u_j u_k \right] \\
&= (2\pi)^{\frac{J}{2}} \exp\left(-\frac{1}{2}\sum_{j\geq J+1} u_j^2\right) \left[\sum_{i=1}^{J} [A]_{jj} + \sum_{j,k\geq J+2} [A]_{jk} u_j u_k \right].
\end{aligned} \tag{A.6}$$

Applying this result to finite q ends the proof. □

Using Lemma A.3, we can easily calculate (A.2):

$$\begin{aligned}
\int_\Omega u^T Au \exp\left(-\frac{1}{2}u^T u\right) du &= \int_s^\infty \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} u^T Au \exp\left(-\frac{1}{2}u^T u\right) du_1 \ldots du_{q-1} du_q \\
&= \int_s^\infty \left[(2\pi)^{\frac{q-1}{2}} \sum_{j=1}^{q-1} [A]_{jj} + [A]_{qq} u_q^2\right] \exp\left(-\frac{u_q^2}{2}\right) du_q \\
&= (2\pi)^{\frac{q-1}{2}} \int_s^\infty \left[\sum_{j=1}^{q-1} [A]_{jj} + [A]_{qq} u_q^2\right] \exp\left(-\frac{u_q^2}{2}\right) du_q \\
&= (2\pi)^{\frac{q-1}{2}} \left[\sum_{j=1}^{q-1} [A]_{jj} \sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\frac{s}{\sqrt{2}}\right) + \right. \\
&\quad \left. [A]_{qq} \left(\sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(\frac{s}{\sqrt{2}}\right) + s \cdot \exp\left(-\frac{s^2}{2}\right)\right)\right] \\
&= (2\pi)^{\frac{q-1}{2}} \left[\sqrt{\frac{\pi}{2}} \operatorname{Tr}(A) \operatorname{erfc}\left(\frac{s}{\sqrt{2}}\right) + [A]_{qq} s \cdot \exp\left(-\frac{s^2}{2}\right)\right]. \quad \text{(A.7)}
\end{aligned}$$

A.2 Some Probability Results

The next result is stated in another form in [9], with a guide to the proof presented in [55].

Lemma 2. *Let $\{W_i\}_{i=1}^N \subset \mathbb{R}^q$ be a set of* iid *random variables and $\tilde{p} = \otimes_{i=1}^N \tilde{p}_i$ be a product distribution. Suppose that $\{Y_i\}_{i=1}^N$ is a set of random variables with conditional probability distribution $p(Y_i|W_i)$. Then, for a measurable function $f(Y,W)$,*

$$E_{\tilde{p}}[f(Y_i,W_i)] = E_{\tilde{p}_i}[f(Y_i,W_i)]. \quad \text{(A.8)}$$

Proof. We have

$$\begin{aligned}
E_{\tilde{p}}[f(Y_i,W_i)] &= \int_{\mathbb{R}^q} f(y_i,w_i)\tilde{p}(w)dw \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(y_i,w_i) \prod_{i=1}^N \tilde{p}_i(w_i) dw_1 \ldots dw_N, \quad \text{(A.9)}
\end{aligned}$$

where (A.9) is possible due to Theorem 5 (Fubini's Theorem). Again using Fubini's Theorem, this

time together with the factorization of $\tilde{p}$, we get:

$$\begin{aligned} E_{\tilde{p}}\left[f(Y_i,W_i)\right] &= \int_{\mathbb{R}^{q-1}} \prod_{j\neq i} \tilde{p}_j(w_j) \int_{\mathbb{R}} f(y_i,w_i)\tilde{p}_i(w_i)dw_i dw_j \\ &= \int_{\mathbb{R}^{q-1}} \prod_{j\neq i} \tilde{p}_j(w_j) E_{\tilde{p}_i}\left[f(Y_i,W_i)\right] dw_j \\ &= E_{\tilde{p}_i}\left[f(Y_i,W_i)\right] \int_{\mathbb{R}^{q-1}} \prod_{j\neq i} \tilde{p}_j(w_j) dw_j \\ &= E_{\tilde{p}_i}\left[f(Y_i,W_i)\right], \end{aligned} \tag{A.10}$$

by the independence of the W_j and the fact that $1 = \int_{\mathbb{R}} \tilde{p}_k(w_k)dw_k$ for $1 \leq k \leq N$, $k \neq i$. □

A.3 Convexity of Quadratic Forms

Lemma 3. *Let A be a real invertible $q \times q$ matrix. If A is positive definite, then so is A^{-1}.*

Proof. Suppose A is positive definite. Then for any nonzero $x \in \mathbb{R}^q$,

$$x^T A x \geq 0. \tag{A.11}$$

In particular, this is true for $y = Ax$. Then, by the symmetry of A,

$$y^T A^{-1} y = (Ax)^T A^{-1} Ax = x^T A^T A^{-1} Ax = x^T A (A^{-1} A) x = x^T A x \geq 0. \tag{A.12}$$

Hence, A^{-1} is positive definite. □

Lemma 4. *Suppose that $x, m \in \mathbb{R}^q$ and A is a $q \times q$ positive definite matrix. Then, the quadratic form defined by*

$$Q(m,A) := (x-m)^T A^{-1} (x-m) = x^T A^{-1} x - 2x^T A^{-1} m + m^T A^{-1} m \tag{A.13}$$

is convex with respect to both m and A.

Proof. First, fix A. Since $x^TA^{-1}x$ has no dependence on m, we will ignore it and instead prove the convexity of $\tilde{Q}(m) = -2x^TA^{-1}m + m^TA^{-1}m$. Suppose $\lambda \in (0,1)$ and $m_1, m_2 \in \mathbb{R}^q$. Then:

$$\begin{aligned}\tilde{Q}(\lambda m_1 + (1-\lambda)m_2) - \lambda\tilde{Q}(m_1) - (1-\lambda)\tilde{Q}(m_2) = &-2\lambda x^TA^{-1}m_1 - 2(1-\lambda)x^TA^{-1}m_2 + \\ &(1-\lambda)^2 m_2^TA^{-1}m_2 + 2\lambda(1-\lambda)m_1^TA^{-1}m_2 + \\ &\lambda^2 m_1^TA^{-1}m_1 + 2\lambda x^TA^{-1}m_1 - \lambda m_1^TA^{-1}m_1 + \\ &2(1-\lambda)x^TA^{-1}m_2 - (1-\lambda)m_2^TA^{-1}m_2 \\ &= -\lambda(1-\lambda)(m_1-m_2)^TA^{-1}(m_1-m_2).\end{aligned}$$

Now, since A is positive definite, then by Lemma 3 so is A^{-1}. Combining this with the fact that $\lambda \in (0,1)$ gives:

$$-\lambda(1-\lambda)(m_1-m_2)^TA^{-1}(m_1-m_2) \leq 0. \tag{A.14}$$

Hence,

$$\tilde{Q}(\lambda m_1 + (1-\lambda)m_2) \leq \lambda\tilde{Q}(m_1) + (1-\lambda)\tilde{Q}(m_2), \tag{A.15}$$

and $\tilde{Q}$ is convex in m. Alternatively, by [10, p.71], $Q(m)$ is convex since $\nabla_m^2 Q(m) = A^{-1}$, which is positive definite.

Now, fix m. The remainder of this proof is based on the work of [40]. To show convexity of $(x-m)^TA^{-1}(x-m)$ it suffices to show the convexity of the matrix inverse A^{-1}, where A is positive definite. To this end, suppose that A and B are two positive definite matrices, let $\lambda \in (0,1)$, and let

$$f(\lambda) := \lambda A + (1-\lambda)B. \tag{A.16}$$

Note that since f is the sum of two positive definite matrices, it is also positive definite, and so is f^{-1} by Lemma 3. By [60], identity (59), we have that

$$\frac{df^{-1}}{d\lambda} = -f^{-1}\frac{df}{d\lambda}f^{-1}. \tag{A.17}$$

Note also that

$$\frac{d^2f}{d\lambda^2} = \frac{d}{d\lambda}(\lambda A + (1-\lambda)B) = 0. \tag{A.18}$$

Moving forward, we'll denote the derivative of f^{-1} with respect to λ as $\left(f^{-1}\right)'$. Taking the derivative of (A.17) gives

$$\begin{aligned}(f^{-1})'' &= -(f^{-1})'f'f^{-1} - f^{-1}f''f^{-1} - f^{-1}f'(f^{-1})' \\ &= f^{-1}f'f^{-1}f'f^{-1} - 0 + f^{-1}f'f^{-1}f'f^{-1} \\ &= 2f^{-1}f'f^{-1}f'f^{-1}. \end{aligned} \tag{A.19}$$

Choose any nonzero vector $u \in \mathbb{R}^q$ and let $\varphi(\lambda) = u^T f^{-1} u$. Then

$$\varphi''(\lambda) = u^T (f^{-1})'' u = 2u^T f^{-1} f' f^{-1} f' f^{-1} u. \tag{A.20}$$

Let $\upsilon = f' f^{-1} u$ so that $\varphi''(\lambda) = 2\upsilon^T f' \upsilon \geq 0$, since f^{-1} is positive definite. Therefore, φ is a convex function over $[0,1]$. That is, for any $\lambda \in (0,1)$,

$$\lambda\varphi(1) + (1-\lambda)\varphi(0) \geq \varphi(\lambda). \tag{A.21}$$

We also have that

$$\varphi(0) = f^{-1}(0) = B^{-1}, \quad \varphi(1) = f^{-1}(1) = A^{-1}. \tag{A.22}$$

Then, substituting these values into (A.21) gives

$$\begin{gathered}\lambda u^T A^{-1} u + (1-\lambda) u^T B^{-1} u \geq u^T \left(\lambda A + (1-\lambda)B\right)^{-1} u \\ \Longleftrightarrow \\ u^T \left(\lambda A^{-1} + (1-\lambda)B^{-1}\right) u \geq u^T \left(\lambda A + (1-\lambda)B\right)^{-1} u.\end{gathered}$$

Since u was an arbitrary nonzero vector, this means that

$$\lambda A^{-1} + (1-\lambda)B^{-1} \succeq \left(\lambda A + (1-\lambda)B\right)^{-1}, \tag{A.23}$$

and A^{-1} is convex.

By Lemma 3, when A is positive definite then so is A^{-1}. By the proof we have just undertaken, when A^{-1} is positive definite, then A is convex. Hence, $Q(m,A)$ is convex with respect to A. □

Appendix B

Supplementary Material for the Bibliometric Study

B.1 Histograms of Bibliometric Input Variables

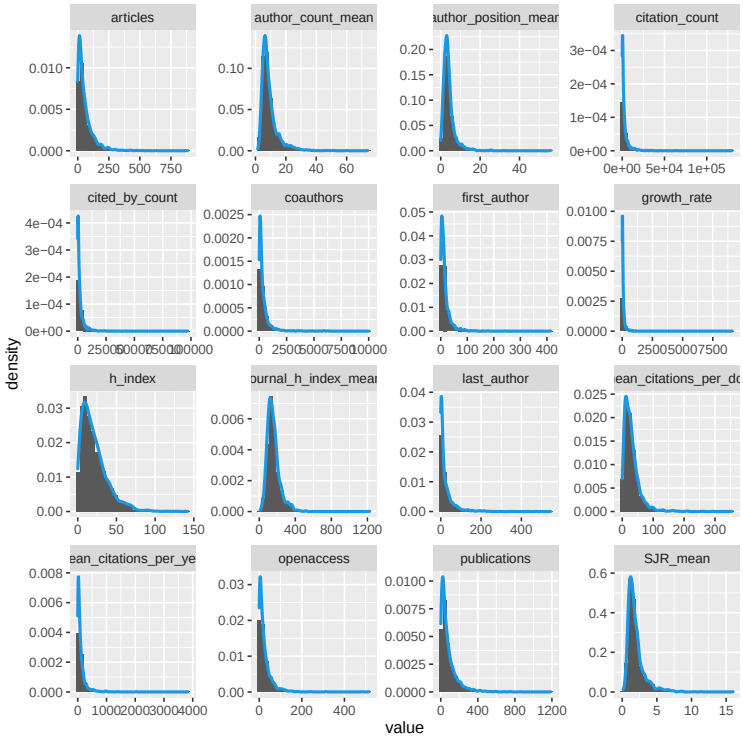


Figure B.1: Histograms of all variables before log transformation.

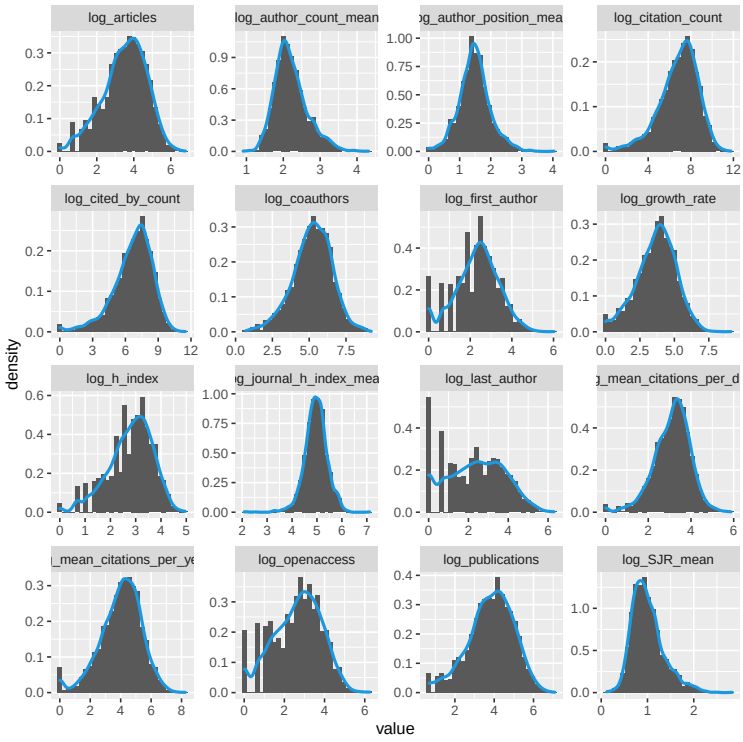


Figure B.2: Histograms of all log-transformed variables using $\log(x+1)$.

Figure B.3: Pairwise scatterplots of all log-transformed bibliometric count variables used as inputs for PCA-C.

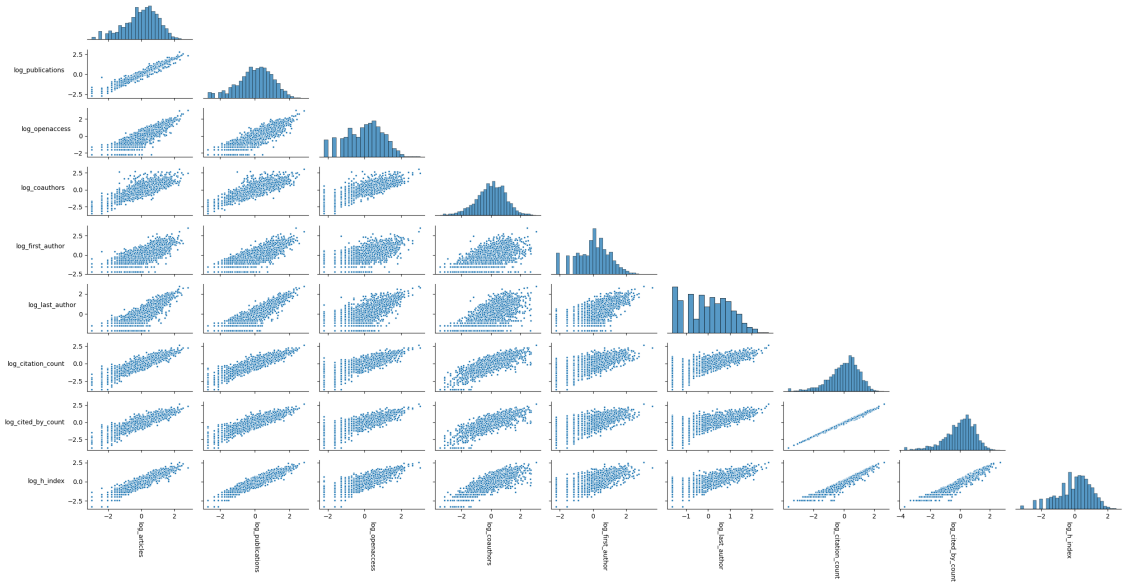


Figure B.4: Pairwise scatterplots of predictions generated by the PCA-C model.

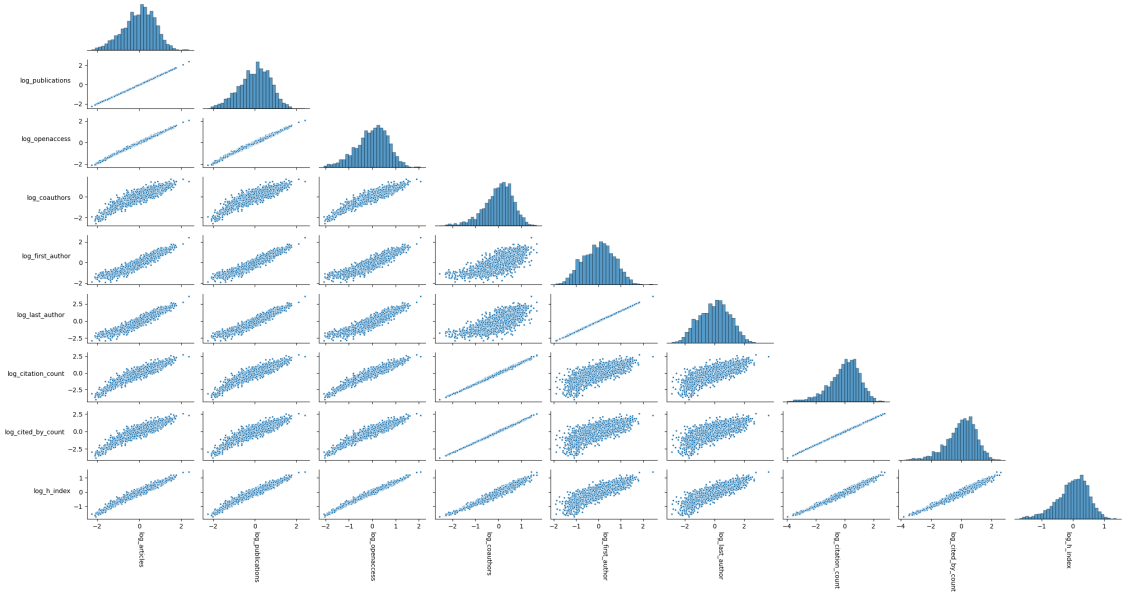


Figure B.5: Pairwise scatterplots of all raw bibliometric count variables used as inputs for PLNPCA.

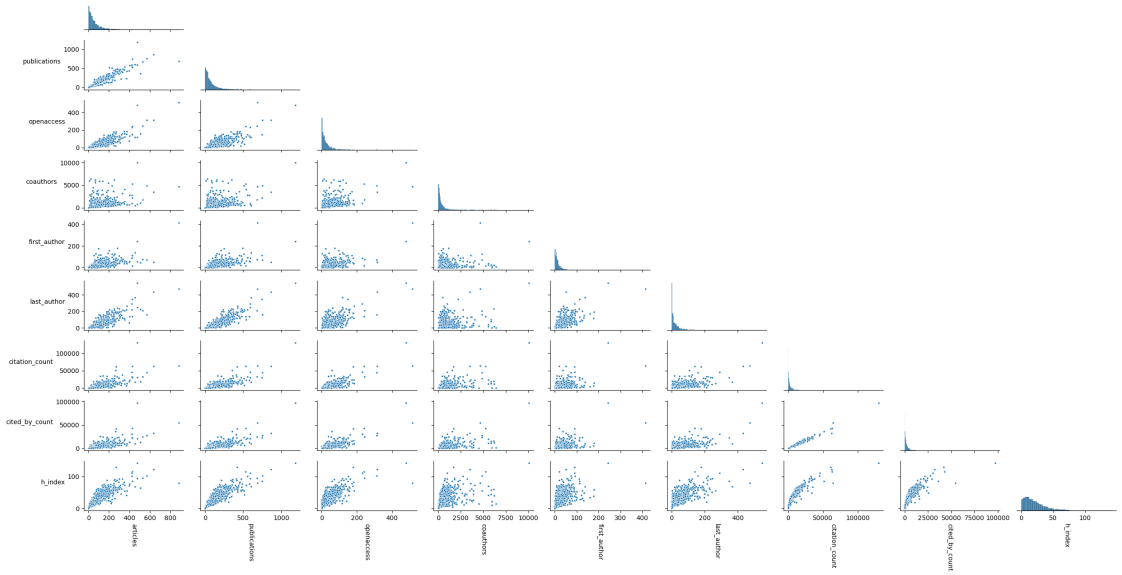


Figure B.6: Pairwise scatterplots of predictions generated by the PLNPCA model.

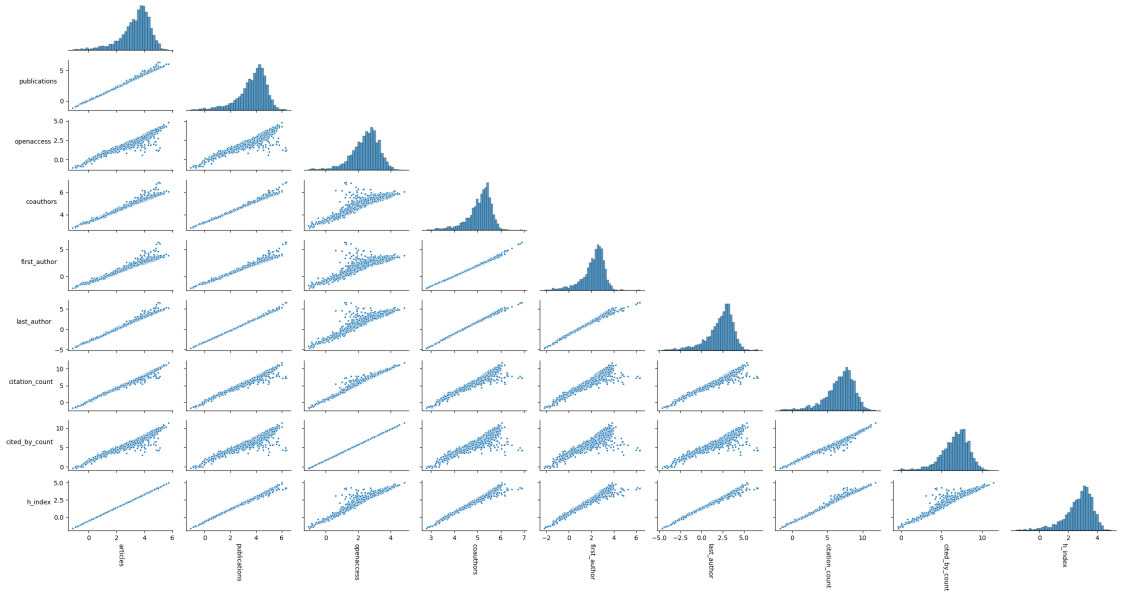


Figure B.7: Pairwise scatterplots of all log-transformed bibliometric variables.

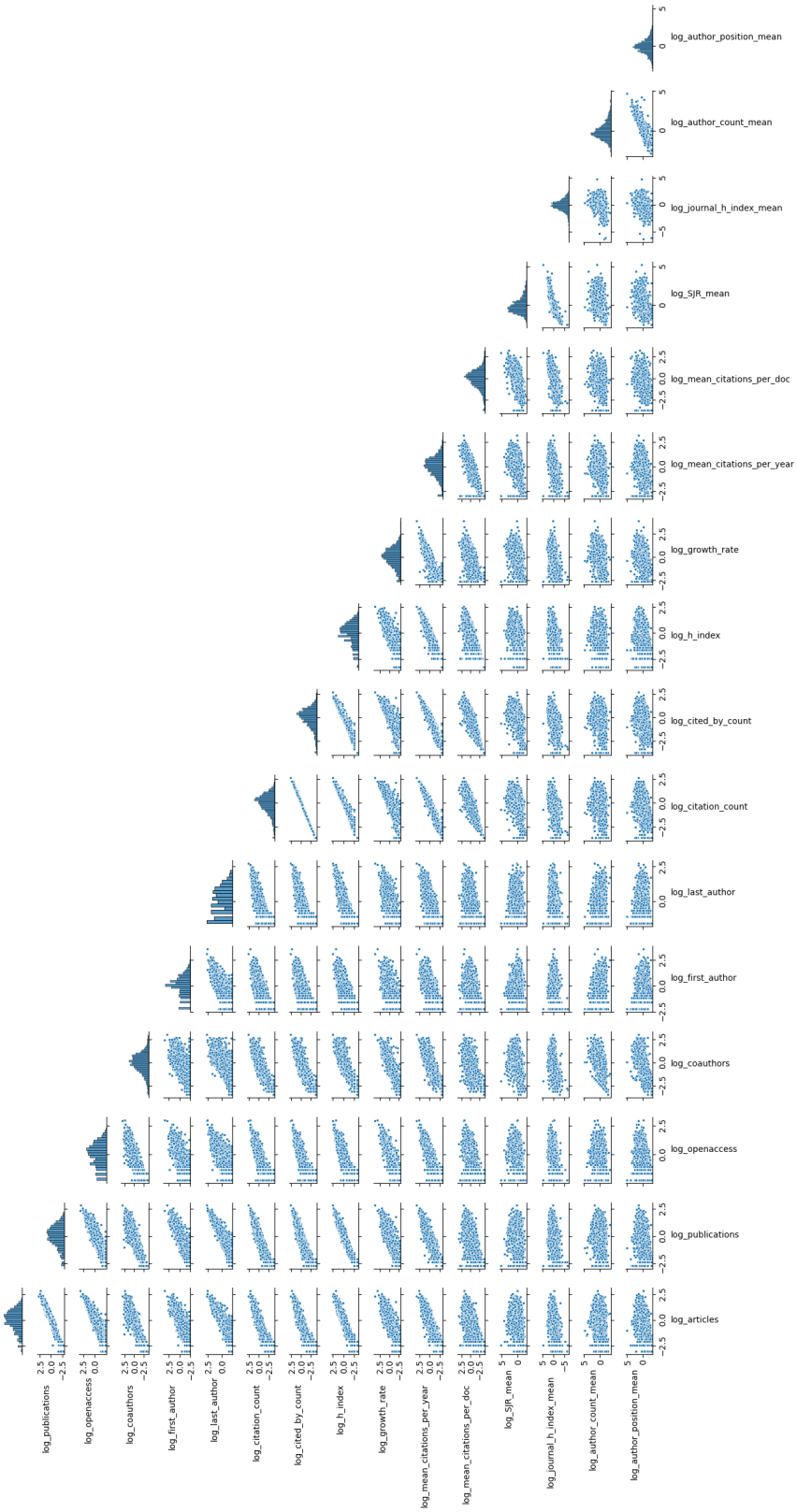


Figure B.8: Pairwise scatterplots of predictions generated by the PCA model.

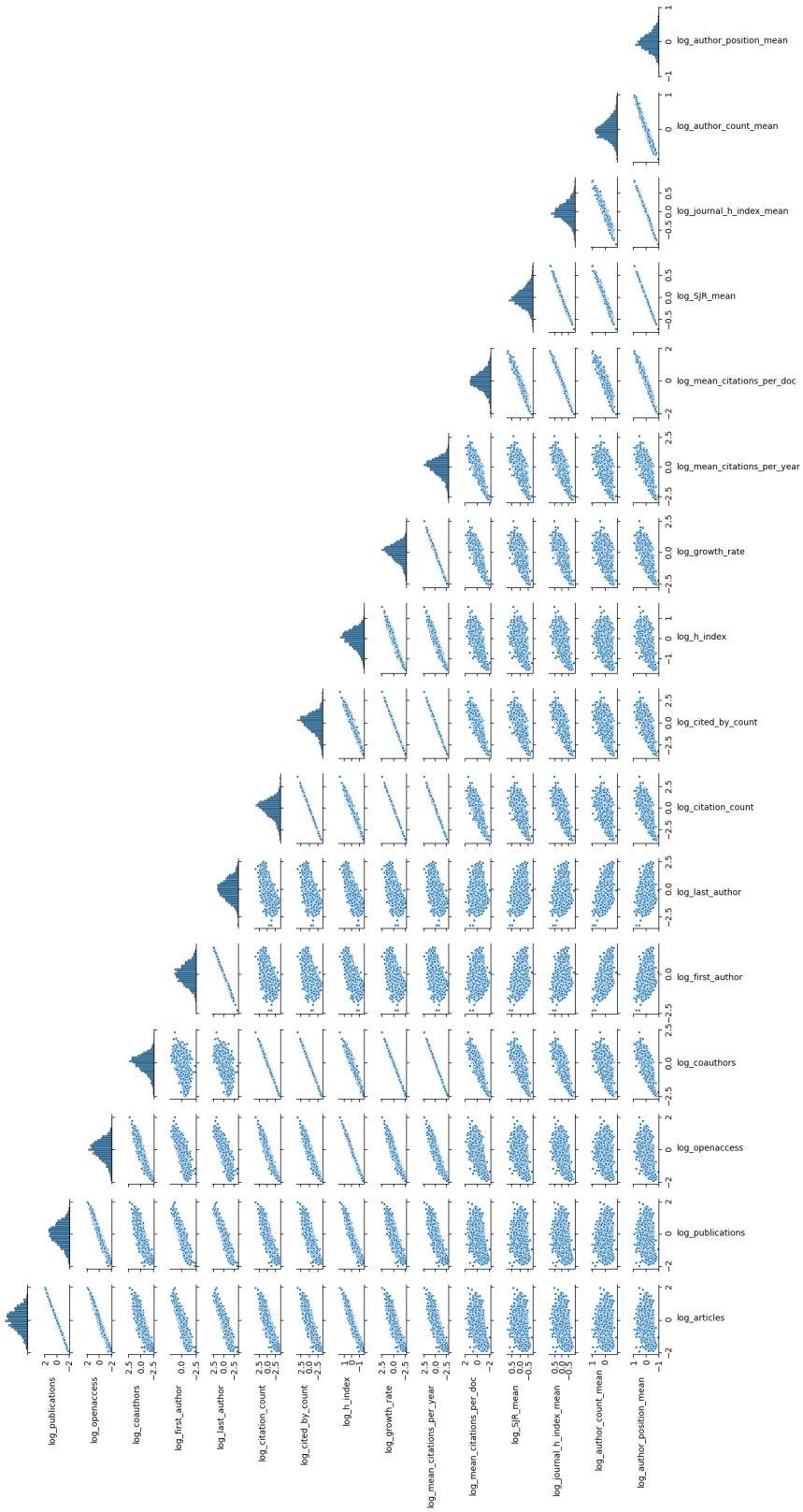


Figure B.9: Pairwise scatterplots of predictions generated by the PWPCA model.

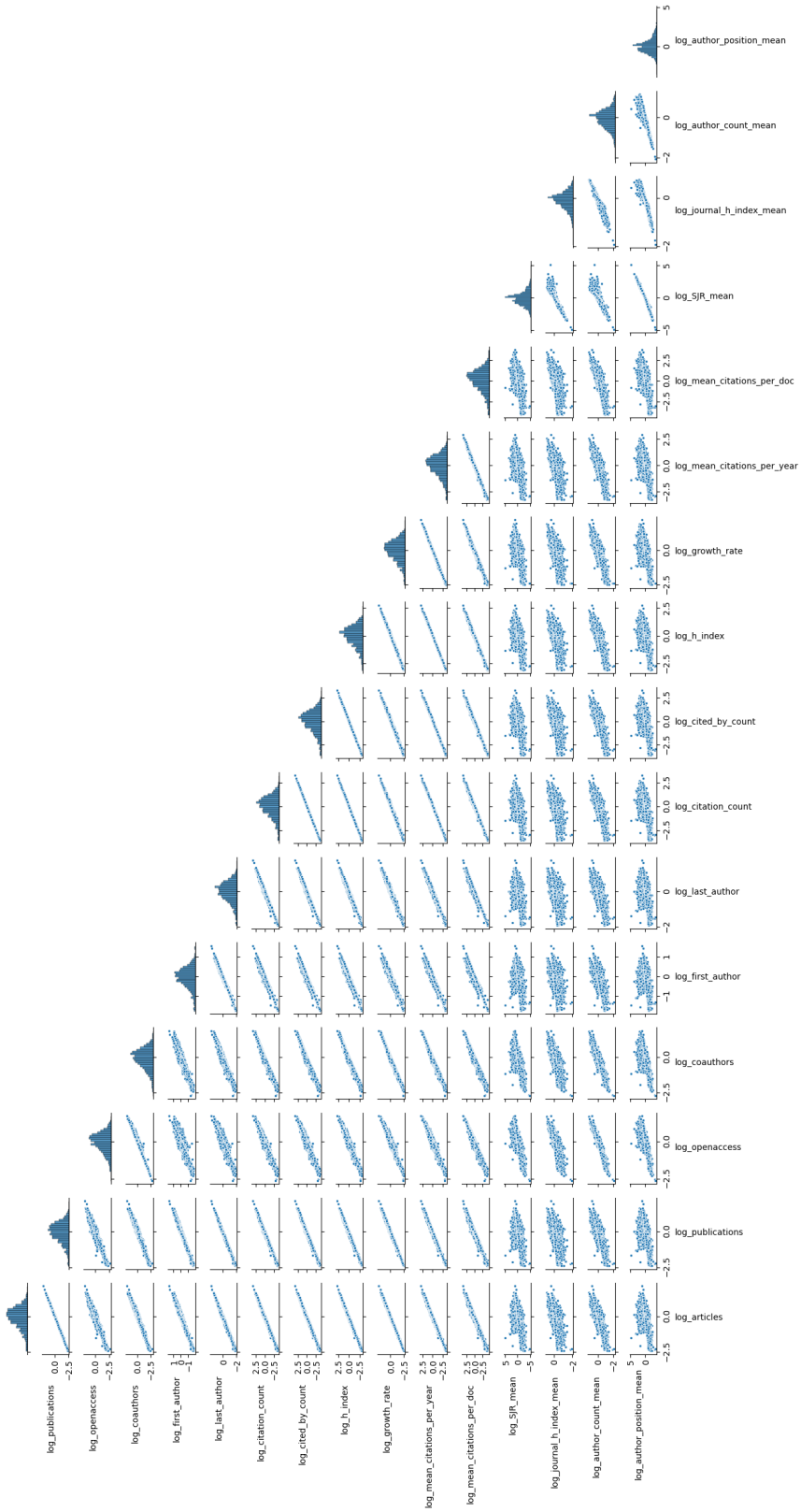


Figure B.10: Pairwise scatterplots of predictions generated by the AE1 model.

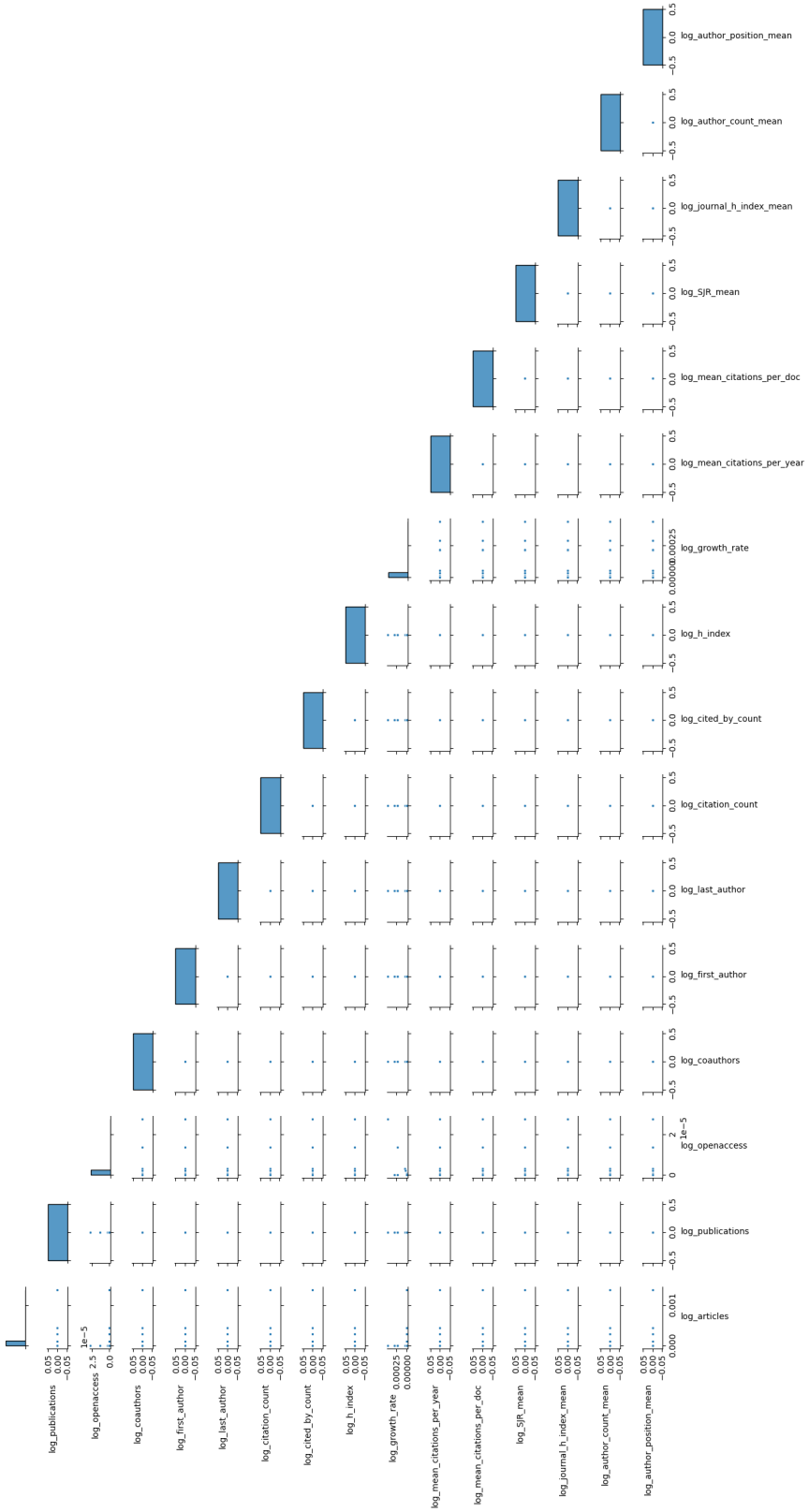
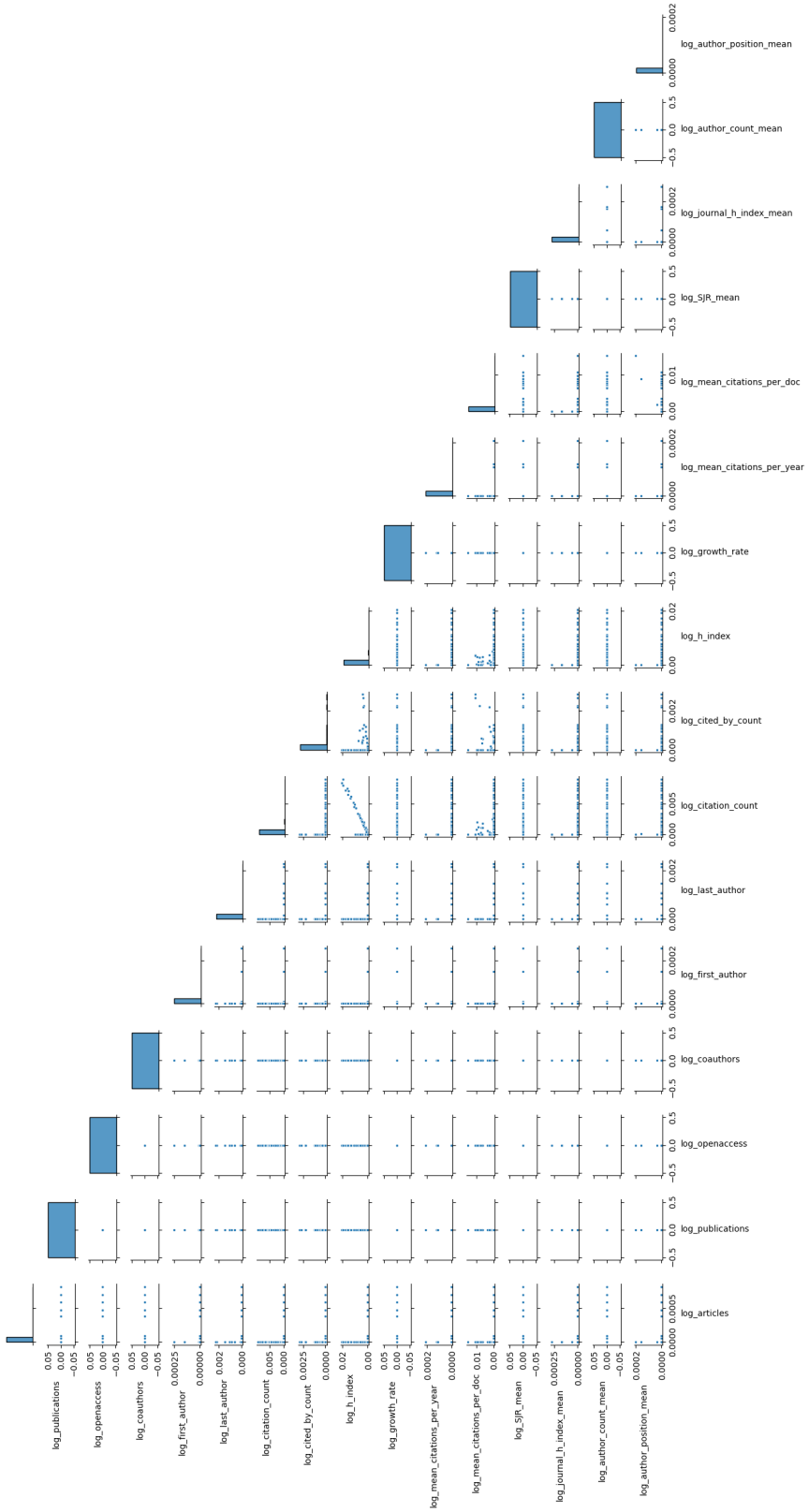


Figure B.11: Pairwise scatterplots of predictions generated by the AE2 model.



Bibliography

[1] ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G. S., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., GOODFELLOW, I., HARP, A., IRVING, G., ISARD, M., JIA, Y., JOZEFOWICZ, R., KAISER, L., KUDLUR, M., LEVENBERG, J., MANÉ, D., MONGA, R., MOORE, S., MURRAY, D., OLAH, C., SCHUSTER, M., SHLENS, J., STEINER, B., SUTSKEVER, I., TALWAR, K., TUCKER, P., VANHOUCKE, V., VASUDEVAN, V., VIÉGAS, F., VINYALS, O., WARDEN, P., WATTENBERG, M., WICKE, M., YU, Y., AND ZHENG, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] ADLER, R., EWING, J., AND TAYLOR, P. Citation statistics: A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS). *Statistical Science 24*, 1 (2009), 1–14.

[3] AITCHISON, J., AND HO, C. The multivariate Poisson-log normal distribution. *Biometrika 76*, 4 (1989), 643–653.

[4] ALLISON, P. D., AND STEWART, J. A. Productivity differences among scientists: Evidence for accumulative advantage. *American Sociological Review* (1974), 596–606.

[5] ANOWAR, F., SADAOUI, S., AND SELIM, B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, kPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). *Computer Science Review 40* (2021), 100378.

[6] ANTHONY, M., AND BARTLETT, P. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[7] ATHREYA, K. B., AND LAHIRI, S. N. *Measure theory and probability theory*. Springer, 2006.

[8] BELKIN, M., AND NIYOGI, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Neural Information Processing Systems* (2001), vol. 14, pp. 585–591.

[9] BISHOP, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

[10] BOYD, S., AND VANDENBERGHE, L. *Convex Optimization*. Cambridge University Press, 2004.

[11] BRADFORD, S. C. *Documentation*. Crosby Lockwood & Son LTD, 1948.

[12] BROOKS, T. A. Evidence for complex citer motivations. *Journal for the American Society for Information Science 37*, 1 (1986), 34–36.

[13] BUCKLAND, M. K., AND HINDLE, A. Documentation notes. *Journal of Documentation 25*, 1 (1969), 52–57.

[14] CAGAN, R. The San Francisco declaration on research assessment (DORA). *Disease Models & Mechanisms 6*, 4 (2013), 869–870.

[15] CERIANI, L., AND VERME, P. The origins of the Gini index: Extracts from variabilità e mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality 10* (2012), 421–443.

[16] CHIQUET, J., MARIADASSOU, M., AND ROBIN, S. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics 12*, 4 (2018), 2674–2698.

[17] CHIQUET, J., MARIADASSOU, M., AND ROBIN, S. The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution 9* (2021), 188.

[18] CHOLLET, F., ET AL. Keras. `https://keras.io`, 2015.

[19] COIFMAN, R. R., AND LAFON, S. Diffusion maps. *Applied and Computational Harmonic Analysis 21*, 1 (2006), 5–30.

[20] CYBENKO, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 2*, 4 (1989), 303–314.

[21] DASGUPTA, A. *Probability for statistics and machine learning: Fundamentals and advanced topics*. Springer Science & Business Media, 2011.

[22] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*, 1 (1977), 1–22.

[23] ECKART, C., AND YOUNG, G. The approximation of one matrix by another of lower rank. *Psychometrika 1*, 3 (1936), 211–218.

[24] FAIRTHORNE, R. A. Empirical distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. *Journal of Documentation 25*, 4 (1969), 319–343.

[25] FRIEDMAN, J. H. Multivariate adaptive regression splines. *The Annals of Statistics 19*, 1 (1991), 1–67.

[26] GARFIELD, E. Citation indexes for science: A new dimension in documentation through association of ideas. *Science 122*, 3159 (1955), 108–111.

[27] GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A., AND RUBIN, D. B. *Bayesian Data Analysis*, 3 ed. CRC Press, 2014.

[28] GHOJOGH, B., GHODSI, A., KARRAY, F., AND CROWLEY, M. Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey. *arXiv preprint arXiv:2101.00734* (2021), 1–16.

[29] GINI, C. *Variabilità e Mutuabilità*. C Cuppini, 1912.

[30] GINI, C. Measurement of inequality of incomes. *The Economic Journal 31*, 121 (1921), 124–126.

[31] GOLDSTEIN, H., AND SPIEGELHALTER, D. J. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 159*, 3 (1996), 385–409.

[32] GOODFELLOW, I., BENGIO, Y., AND COURVILLE, A. *Deep Learning*. MIT press, 2016.

[33] HAM, J., LEE, D. D., MIKA, S., AND SCHÖLKOPF, B. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the Twenty-first International Conference on Machine Learning* (2004), p. 47.

[34] HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H., AND FRIEDMAN, J. H. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, vol. 2. Springer, 2009.

[35] HICKS, D., WOUTERS, P., WALTMAN, L., DE RIJCKE, S., AND RAFOLS, I. Bibliometrics: The Leiden manifesto for research metrics. *Nature 520*, 7548 (2015), 429–431.

[36] HIRSCH, J. E. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences 102*, 46 (2005), 16569–16572.

[37] HORNIK, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*, 2 (1991), 251–257.

[38] HORNIK, K., STINCHCOMBE, M., AND WHITE, H. Multilayer feedforward networks are universal approximators. *Neural Networks 2*, 5 (1989), 359–366.

[39] HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*, 6 (1933), 417.

[40] HUI, A. H.-H. Is inverse matrix convex? Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/454753 (version: 2013-07-29).

[41] INOUYE, D. I., YANG, E., ALLEN, G. I., AND RAVIKUMAR, P. A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics 9*, 3 (2017), e1398.

[42] IOANNIDIS, J. P., KLAVANS, R., AND BOYACK, K. W. Multiple citation indicators and their composite across scientific disciplines. *PLoS Biology 14*, 7 (2016), e1002501.

[43] JOLLIFFE, I. T. *Principal Component Analysis*, 2 ed. Springer, 2002.

[44] KASSAMBARA, A., AND MUNDT, F. *Factoextra: Extract and visualize the results of multivariate data analyses, version 1.0.7*, 2020. https://CRAN.R-project.org/package=factoextra.

[45] KAWAMURA, K. The structure of multivariate Poisson distribution. *Kodai Mathematical Journal 2*, 3 (1979), 337–345.

[46] KINGMA, D. P., AND WELLING, M. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114* (2013), 1–14.

[47] LAWANI, S. M. Bibliometrics: Its theoretical foundations, methods, and applications. *Libri 31*, 4 (1981), 294–315.

[48] LEE, S., AND BOZEMAN, B. The impact of research collaboration on scientific productivity. *Social Studies of Science 35*, 5 (2005), 673–702.

[49] LERMAN, R. I., AND YITZHAKI, S. A note on the calculation and interpretation of the Gini index. *Economics Letters 15* (1984), 363–368.

[50] LOTKA, A. J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16*, 12 (1926), 317–323.

[51] MANDELBROT, B. An informational theory of the statistical structure of language. *Communication Theory 84* (1953), 486–502.

[52] MARDIA, K. V., KENT, J. T., AND BIBBY, J. M. *Multivariate Analysis*. Academic Press, 1979.

[53] MARTIN, B. R., AND IRVINE, J. Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy 12*, 2 (1983), 61–90.

[54] MIASKO, T., AND NOWOTNY, M. *PyJAGS: The Python interface to JAGS*, 2022.

[55] MICROHAUS (HTTPS://STATS.STACKEXCHANGE.COM/USERS/294515/MICROHAUS). Explain the expected value of an indicator variable in section 9.3.1 of PRML. Cross Validated. URL:https://stats.stackexchange.com/q/511163 (version: 2021-02-26).

[56] MOHRI, M., ROSTAMIZADEH, A., AND TALWALKAR, A. *Foundations of Machine Learning*. MIT press, 2018.

[57] MOLINARI, J.-F., AND MOLINARI, A. A new methodology for ranking scientific institutions. *Scientometrics 75*, 1 (2008), 163–174.

[58] MONTUFAR, G. F., PASCANU, R., CHO, K., AND BENGIO, Y. On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems 27* (2014), 1–9.

[59] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[60] PETERSEN, K. B., AND PEDERSEN, M. S. The matrix cookbook. *Technical University of Denmark 7*, 15 (2012), 510.

[61] PLUMMER, M., ET AL. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (2003), no. 125.10 in 124, Vienna, Austria, pp. 1–10.

[62] PRICE, D. D. S. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science 27*, 5 (1976), 292–306.

[63] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[64] ROWEIS, S. T., AND SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 5500 (2000), 2323–2326.

[65] RUDY, J. *Py-earth*, 2013.

[66] RUMELHART, D. E., HINTON, G. E., AND WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature 323*, 9 (1986), 533–536.

[67] SCHÖLKOPF, B., HERBRICH, R., AND SMOLA, A. J. A generalized representer theorem. In *International Conference on Computational Learning Theory* (2001), Springer, pp. 416–426.

[68] SCHÖLKOPF, B., SMOLA, A., AND MÜLLER, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation 10*, 5 (1998), 1299–1319.

[69] SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics* (1978), 461–464.

[70] STACKLIES, W., REDESTIG, H., SCHOLZ, M., WALTHER, D., AND SELBIG, J. pcaMethods: A bioconductor package providing PCA methods for incomplete data. *Bioinformatics 23*, 9 (2007), 1164–1167.

[71] STEPHAN, P., VEUGELERS, R., AND WANG, J. Reviewers are blinkered by bibliometrics. *Nature 544*, 7651 (2017), 411–412.

[72] STEWART, J. A. The Poisson-lognormal model for bibliometric/scientometric distributions. *Information Processing & Management 30*, 2 (1994), 239–251.

[73] SVANBERG, K. The method of moving asymptotes: A new method for structural optimization. *International Journal for Numerical Methods in Engineering 24*, 2 (1987), 359–373.

[74] TENENBAUM, J. B., DE SILVA, V., AND LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 5500 (2000), 2319–2323.

[75] TIPPING, M. E., AND BISHOP, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*, 3 (1999), 611–622.

[76] TOLSGAARD, M. G., ELLAWAY, R., WOODS, N., AND NORMAN, G. Salami-slicing and plagiarism: How should we respond? *Advances in Health Sciences Education 24*, 1 (2019), 3–14.

[77] TORGERSON, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika 17*, 4 (1952), 401–419.

[78] TU, L. W. *An Introduction to Manifolds*, 2 ed. Springer, 2011.

[79] VAN DER MAATEN, L., POSTMA, E., AND VAN DEN HERIK, J. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research 10*, 66-71 (2009), 13.

[80] WALTMAN, L., AND VAN ECK, N. J. The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology 63*, 2 (2012), 406–415.

[81] WICKHAM, H. Reshaping data with the reshape package. *Journal of Statistical Software 21* (2007), 1–20.

[82] WICKHAM, H. *ggplot2: Elegant Graphics for Data Analysis*, 2016. https://ggplot2.tidyverse.org.

[83] WICKHAM, H., AND BRYAN, J. *readxl: Read Excel Files*, 2022. https://readxl.tidyverse.org, https://github.com/tidyverse/readxl.

[84] WU, C. J. On the convergence properties of the EM algorithm. *The Annals of Statistics 11*, 1 (1983), 95–103.

[85] ZIPF, G. K. *Human behavior and the principle of least effort*. Addison-Wesley Press, Inc, 1949.