Bayesian Principal Component Analysis (BCPA) applied to ecological and image processing data

Morgan Southgate, Philip Nguyen

December 16, 2021

1 Introduction

Principal Component Analysis (PCA) is a widely used statistical technique that aims to summarize a data-set in terms of a lower dimensional subspace without losing too much information [1, 2, 18]. It is relatively straightforward to implement using modern programming libraries and makes few assumptions regarding the structure of the data (i.e. it is non-parametric) [11, 18]. PCA is an unsupervised technique, in comparison to related techniques such as latent factor analysis, in that grouping variables are not specified in the model. Because of this attribute, it is an excellent data-exploration technique.

One limitation of this procedure is that it lacks a generative model of the data, and cannot handle missing values [15, 21]. Additionally, the dimensions of the subspace must be specified when reconstructing the data, a non-trivial task when dealing with large datasets. Probabilistic PCA (PPCA) was developed in order to overcome these limitations [21]. This technique is an improvement on normal PCA, but one may also want to incorporate prior information regarding the data in order to improve the accuracy of the estimated data and model, which isn't possible in standard PPCA. Bayesian PCA (BPCA) was formulated to remedy these issues by automatically determining the number of effective principal components while simultaneously yielding a generative story of the data [15, 21].

The purpose of this project is to understand how BPCA works, and to test its utility on two real-world data-sets. The first data-set describes the ecology of the *Adiantum pedatum* complex, a clade of maidenhair ferns. Ecological niche is defined by Hutchinson as the portion of an n-dimensional hypervolume, or niche space, which the species in question occupies [10]; these dimensions are usually interpreted as single ecological variables, but they can also be considered as combinations of variables. As such, this definition enables a quantitative determination of ecological niche using multivariate methods such as PCA. The second data-set is from the MNIST data-base, used for training image processing systems. One of the many applications of PCA is image processing due to the redundancy and high dimensionality of images.

1.1 A Brief Overview of PCA

Here we summarize PCA and refer the interested reader to other works for detailed derivations [11, 18].

Let X be an $n \times p$ matrix that contains n observations or events characterized by p mean-centered variables. The goal of PCA is to find a set of orthogonal basis vectors W (the principal axes) that maximize the variance of the projection of X onto W where $Z = XW$ is the resulting projection, i.e. the principal components (also known as the loadings).

Such a basis is found by computing the eigenvectors and eigenvalues of the covariance matrix of the data. W is a $p \times k$ matrix containing k eigenvectors of the covariance matrix ordered by the magnitude of their corresponding eigenvalues. These eigenvectors are the principal axes that maximize the variance of the projected data Z (while also minimizing the reconstruction error [18]).

To perform dimensionality reduction, the principal components that explain little of the variation in the data can be removed from W, and then the data projected back onto W. In the end we have a lower rank representation of the original data, $ZW^T \approx X$, which captures the most significant patterns. The components to retain can be determined by cross-validation, calculating a Q^2 score for each component that describes the ratio of variance that can be predicted by each component alone.

The steps of PCA can be summarized as follows:

1. Mean-center each feature vector of a data matrix X
2. Obtain the eigenvectors and eigenvalues of the covariance matrix of X (or similarly, perform Singular Value Decomposition on X [18])
3. Construct a projection matrix W from the k eigenvectors that explain most of the variance, i.e. the principal axes
4. Project X onto the principal axes of W to obtain a lower dimensional representation of the data, i.e. the principal components $Z = XW \rightarrow ZW^T = XWW^T \approx X$

The accuracy of the model can be checked, and the mean square error (MSE) calculated, by reconstructing the original variables from the principal components. The reconstructed data can be obtained by mapping the data back to p dimensions with the transposed matrix of the eigenvectors, W ([3]):

$$X_{rec} = ZW^T.$$

1.2 Probabilistic PCA

One limitation of conventional PCA is the lack of a generative model that can explain the observed data [21]. Probabilistic PCA (PPCA) addresses this through the construction of a latent variable model whose maximum likelihood estimation yields the desired principal components. We provide a brief walkthrough of the derivation since it naturally extends to the formulation of Bayesian PCA (BPCA), our main method of interest.

We begin by assuming that the observed p-dimensional data of X was generated by a set of k-dimensional latent variables Z through the linear mapping

$$x_i = Wz_i + \mu + \epsilon$$

where

- W maps the latent observation z_i to x_i
- μ is an offset term
- $\epsilon \sim N(0, \sigma^2 I_p)$ is a noise term where I_p is a p-dimensional identity matrix
- and $z_i \sim N(0, I_k)$ where I_k is a k-dimensional identity matrix.

This leads to a conditional multivariate Gaussian distribution of the form

$$p(X|Z) = N(WZ + \mu, \sigma^2 I_p)$$

and the marginal distribution of X as

$$p(X) = N(\mu, C)$$

where $C = WW^T + \sigma^2 I_p$. From here it can be shown that the log-likelihood is

$$\ln p(X|\mu, W, \sigma^2) = -\frac{N}{2}\{p \ln 2\pi + \ln |C| + Tr[C^{-1}S]\}$$

after plugging in the maximum likelihood estimator (MLE) for μ, which is simply the mean of the data [21]. S is the covariance matrix of X and $Tr(.)$ is the trace operator. This generative process is summarized in Figure [1].

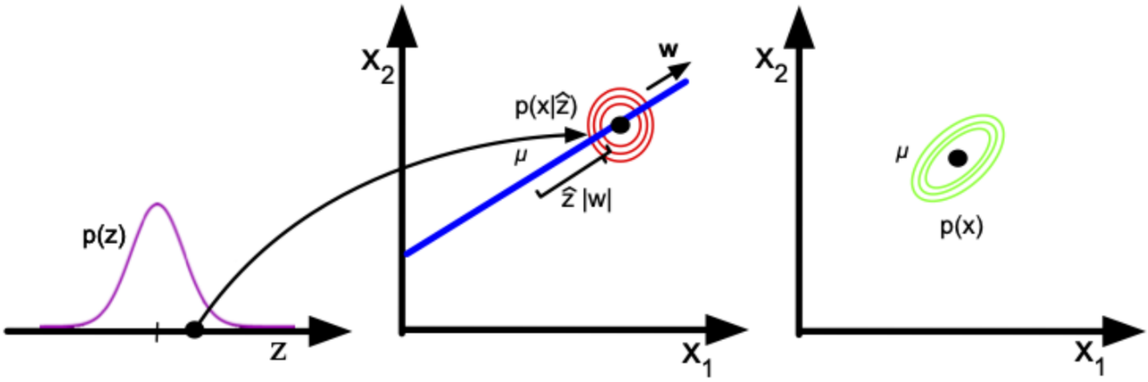


Figure 1: An illustration of the generative process where a single 1-d latent variable is mapped to a 2-d offset. When integrating over the latent variable, one can imagine a Gaussian "cloud" being slid across the principal axis (w in this case) as draws from $p(z)$ are generated, returning an elliptical Gaussian in x space centered on μ. Adapted from Figure 12.9 of [6].

The MLEs for W and σ can be determined using two approaches: expectation maximization (EM) [7] or the analytical solution. The end result is that the maximum likelihood solution for W must satisfy

$$W = U_k(K_k - \sigma^2 I_k)^{1/2} R$$

where U_k is a $p \times k$ matrix whose columns are given by the k eigenvectors of the covariance matrix S associated with the k largest eigenvalues [21, 15]. K_k is a $k \times k$ diagonal matrix of eigenvalues and R is an arbitrary $k \times k$ rotation matrix we can set as $R = I_k$.

The MLE for the variance σ^2 turns out to be

$$\sigma^2 = \frac{1}{p-k} \sum_{i=k+1}^{p} \lambda_i$$

which can be interpreted as the average information associated with the removed dimensions. In the noise free limit where $\sigma^2 = 0$, $W_{\text{MLE}} = U_k K_k^{1/2}$, which is proportional to the standard PCA solution (the columns of W_{MLE} are in the same direction and span the same subspace as the standard PCA solution but have smaller magnitudes).

To use PPCA as an alternative to PCA, one needs to compute the posterior mean $\mathbb{E}[Z|X]$ using Bayes rule [14]. The posterior distribution is given as

$$p(Z|X) = N(Z|M^{-1}W^T(X - \mu), \sigma^2 M^{-1})$$

where $M_{k\times k} = W^TW + \sigma^2 I$. Computing the expectation using the maximum likelihood estimators for μ, σ^2, and W as $\sigma^2 \to 0$, the posterior mean becomes

$$\mathbb{E}[Z|X] = (W^TW)^{-1}W^T(X - \bar{X})$$

which is the orthogonal projection of the data into the latent space. In other words, this is the generative story described by the desired principal components. Additionally, one can use the EM algorithm to efficiently compute W in large datasets by bypassing the need to evaluate the covariance matrix [6].

In the case when $\sigma^2 > 0$, the reconstruction of the data, i.e. $X \approx W \cdot \mathbb{E}[Z|X] + \mu$, is not an orthogonal projection of X and is therefore not optimal in terms of the squared error between the true data and the reconstructed data. However, the optimal reconstruction from the conditional mean can still be obtained as $X \approx W(W^TW)^{-1}M\mathbb{E}[Z|X] + \bar{X}$ using the maximum likelihood estimates [21].

1.3 Bayesian PCA

Conventional PCA and PPCA assume that the desired dimensionality of the principal subspace is known or given, but this must be chosen in practice according to the problem at hand. One could use cross-validation with PPCA to determine the optimal number of dimensions, but this can become computationally costly with larger datasets and probabilistic mixtures of PCA models [4].

BPCA provides an automatic way to determine the effective dimensionality of the principal subspace by utilizing a hierarchical prior over the matrix W [4]. Specifically we set

$$p(W|\alpha) = \prod_{k=1}^{K} (\frac{\alpha_k}{2\pi})^{D/2} \exp\{-\frac{1}{2}\alpha_k {w_k}^T w_k\}$$

where w_k is the kth column of W. Each α_k is found in an iterative fashion as part of the optimization procedure. The values of α_k that are driven to infinity cause the corresponding parameter vector w_k to zero out. Thus the effective dimensionality of the principal subspace is determined by the number of finite α_k values with the corresponding vectors w_k being relevant for modeling the data distribution [6]. This approach makes a trade-off between improving fit to the data and reducing model complexity by tuning the relevant vectors w_k inversely with α_k.

A fully Bayesian approach sets priors over the parameters μ, α and τ, in addition to W [5]. For simplicity we follow the approach wherein μ is instead estimated from the data, placing a prior only over α and τ, though in [5] Bishop specifies broad Gamma distributions over μ as well. Thus in hierarchical form we have the following generative model:

$$\mathbf{x}_n \sim \mathcal{N}(\mathbf{W}\mathbf{z}_n, \tau^{-1}\mathbf{I}_D)$$

$$\mathbf{W}_k|\alpha_k \sim \mathcal{N}(0, \alpha_k^{-1}\mathbf{I}_D)$$

$$\alpha_k \sim \text{Gam}(\alpha_0, \beta_0)$$

$$\tau \sim \text{Gam}(\alpha_0, \beta_0)$$

$$\mathbf{z_n} \sim \mathcal{N}(0, \mathbf{I}_K)$$

$$N = \text{number of data points (indexed by n)}$$

$$K = \text{latent dimension (indexed by k)}$$

$$D = \text{dimension of data point}$$

where α_k determines the relevance of the column w_k of the principal subspace. The generative story of BPCA is summarized in 2.

To obtain the predictive density of the data, we marginalize the parameters out, i.e.

$$p(\tilde{x}|X) = \int\int\int p(\tilde{x}|\mu, W, \sigma^2)p(\mu, W, \sigma^2|X)d\mu dW d\sigma^2.$$

In practice the posterior distribution must be approximated since the analytical solution is intractable. Bishop suggests a local Gaussian (Laplace) approximation, and an MCMC Gibbs sampling approach [4]. A subsequent method using variational inference was developed where a family of distributions $Q(\theta)$ are used to approximate the true posterior distribution via an optimization scheme that minimizes the KL-divergence between the proposed and true posterior distribution [5]. This approach results in a distribution over all parameters of interest, mainly the optimal subspace that we can use to generate predictions and reconstructions of our data $\tilde{x}$.

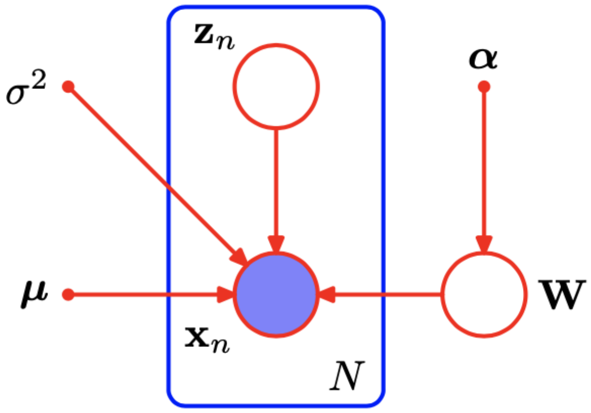


Figure 2: Schematic representation of BPCA as a probabilistic model, adapted from Figure 1 of [6]. The box defines a data-set of N observations of the visible data x_n. This observed data is described by a transformation matrix W which in turn is governed by a set of hierarchical priors α. The observed data are also governed by the observed mean μ and variance σ^2.

1.4 Project Purpose

The purpose of this project is to implement BPCA on two different data-sets in order to establish an understanding of the analysis and how it differs in both process and result from standard PCA and probabilistic PCA.

1.4.1 Maidenhair fern ecology

The *Adiantum pedatum* complex is a clade of maidenhair ferns; in North America, this clade comprises two diploid species, *Adiantum pedatum* L. and *Adiantum aleuticum* (Rupr). Paris as well as *Adiantum viridimontanum* Paris, a fertile hybrid between these two species ([16]). There is a remarkable diversity in ecological niche and geographical distribution among the three species; *A. pedatum* is broadly distributed in rich woods across eastern North America, whereas *A. aleuticum* and *A. viridimontanum* are restricted to a few scattered serpentine habitats in northern VT and Quebec. The three species of *Adiantum* occur on an ecological spectrum defined by soil chemistry; the rich woods habitats of *Adiantum pedatum* are characterized by an abundance of calcium and other important plant nutrients, whereas the habitat of the serpentine maidenhair ferns is characterized by a strikingly low Ca:Mg ratio and the presence of heavy metals including Cd, Cr, Ni, and Zn. The habitats described above represent the extreme ends of the spectruum, but there are many in between.

Previous study by the first author utilized PCA to characterize niche differences among these three closely-related species, and documented niche intermediacy and niche expansion for the hybrid species on a local scale in Vermont [19]. The current version of the *Adiantum* dataset comprises a set of 32 ecological attributes, primarily relating to substrate characteristics, collected 104 survey plots across northeastern North America; each survey plot consists of occurrences of one or two species of maidenhair fern.

1.4.2 MNIST

In addition to the *Adiantum* dataset, we look at the MNIST dataset of handwritten digits to determine how well BPCA is able to generate and reconstruct images. The dataset is comprised of 1797 8×8 images (64 features) of the numbers $0 - 9$, with approximately 180 samples per class. This approach may yield a simpler, more computationally efficient alternative to Variational Autoencoders [13], a popular class of neural networks often used for data compression, representation, and generation. It also provides a point of comparison to how well BPCA performs on the Adiantum dataset.

2 Methodology

2.1 Adiantum

The *Adiantum* dataset consists of 103 observations of 32 quantitative ecological variables, including percent canopy cover, slope, soil layer depth, and soil chemistry. All analysis was done in R ([17]).The pearson correlation coefficient was determined for all variable pairs using the cor() function in R, and one of each pair with an R^2 value greater than 0.75 was removed. This yielded a final data-set comprising 24 quantitative variables. These data were centered and scaled by unit variance using the prep() function in the pcaMethods package.

The standard PCA model was completed using the "svd" option of the pca() function in the pcaMethods package. Components returned was set to 23, and cross-validation completed using Q^2. This metric is calulated by iteratively fitting the model to include one more principal component each time, and the optimum number of PC's to retain is determined by looking at the goodness of fit criteria for each PC ([8]). Data from the standard PCA model were reconstructed using the formula $Xrec = Z * W^T$, and visualized alongside the original centered and scaled variables.

Bayesian PCA was implemented using a stan adaptation [12] of the original model provided by Bishop [4, 5]. This model specifies broad priors over α and τ, with shape and rate parameters of $(1, 1)$ and $(1e^-3, 1e^-3)$, respectively, but does not include a prior on μ. The model was first fit using variational inference following the example in [12], using the vb() function from the rstan package in R. This model fit successfully, but reported pareto-k diagnostic values around 8, and advised use of the classic sampling() function. The stan() function with NUTS sampler was used to fit the models instead, with chains set to 4, and 7000 iterations. Prior values over α and τ were varied systematically to examine the impact of 1) increasing only the prior over τ, 2) increasing only the prior over α, and 3) increasing both parameters. Of these three variations and the original model specification, the best one was selected by reconstructing the variables for each model and plotting them against the original mean-centered, scaled variables.

Using the resulting model, MCMC diagnostics were checked using the bayesplot package. For a subset of the parameters, area plots were made using the mcmc area() function, and trace plots assessing chain convergence were made using the mcmc trace() function. The Potential Scale Reduction Factor (PSRF) was checked for all parameters, and those greater than 1.01 were noted. A pairs plot was used to assess the correlation of a subset of parameters along the principal axes W.

For model diagnostics, a hinton plot was used to characterize the principal components W with the greatest magnitude (8). A hinton plot is essentially a heatmap of W, showing both the direction and magnitude of each variables loading onto each eigenvector. The values of α corresponding to each principal component were evaluated, and those with the lowest value noted. The data were reconstructed using the same formula as for normal PCA results, but were scaled by $1/tau$ to enable direct comparison between the original and reconstructed variables. Using the reconstructed data, a posterior predictive check was conducted by plotting all reconstructed variables against the centered and scaled original variables.

The principal components with the lowest associated alpha value were plotted pairwise using the ggpairs() function of the GGally package, in order to identify patterns of separation by species. Based on the combinations of PC's that appeared meaningful in distinguishing among the three species, a plot of the loadings and scores along each component were visualized.

2.2 MNIST

Following the code and examples in [9], a variational inference approach was used to fit BPCA to the MNIST dataset as an additional frame of reference for BPCA performance. The images were loaded from the Scikit-learn package in Python, then flattened into a 1-D array to be fit by the model. As with the *Adiantum* data, broad Gamma priors were used over α and τ while μ was estimated during optimization using 5000 iterations. Additionally we varied hyperparameter values for α and τ in order to assess the impact of narrow and broad Gamma priors on model fit.

An optimal solution for the parameters was found for each model, typically in under 1000 iterations. To determine model performance on the dataset, we compared the mean-squared error of the reconstructed data using conventional PCA with the effective dimensions found by BPCA, BPCA with the effective dimensions, BPCA with all dimensions of the subspace, and data generated from the fitted BPCA model (drawn from a Multivariate Normal distribution in the effective dimension space) (Table [2]).

In PCA the data are reconstructed by projecting the data onto the optimal transformation matrix W, followed by inverse transforming the projection (the principal components). In BPCA the process is similar except the optimal transformation matrix is the mean of the posterior distribution over W, conditioned on the data. New data points are generated from the BPCA model by drawing values from a Multivariate Normal distribution centered on the mean of the observations with the covariance matrix $C = WW^T + \tau^{-1}I_D$ where W is the posterior mean of W and τ is defined by a prior draw from a Gamma distribution. Additionally, we visually inspected the reconstructed and generated images to gain an intuitive sense of how well the model fits the data.

3 Results

3.1 Adiantum

3.1.1 Standard PCA

The Q^2 results obtained from the cross-validation test indicate that the ratio of variance that can be predicted independently by each PC peaks at PC3, and that the highest range of Q^2 values occur between 2 and 7 (3). Principal components 1-7 were thus selected as the dimensions to retain. The pairwise comparison of these components revealed the greatest amount of niche separation along PC1 and PC2 (17), and so these dimensions were selected for niche comparison. Data reconstruction with 7 principal components shows good structure and variance for most variables (18).

The pairwise plot of the first five principal components shows niche separation primarily along components 1 and 3 (17). A plot of the scores along these components shows a clear pattern of

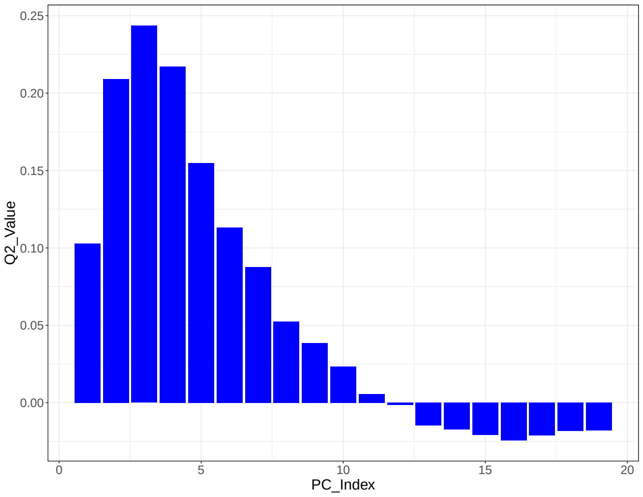


Figure 3: Plot of the Q2 cross-validation scores obtained for each PC

niche separation among the three species, with *Adiantum pedatum* occupying portions of niche space corresponding to habitats with high amounts of Ca, and a deep A soil layer. *Adiantum aleuticum* occupies portions of niche space corresponding to higher levels of soil Mg, and a higher soil pH. *Adiantum viridimontanum* occupies a portion of niche space intermediate between it's two progenitors.

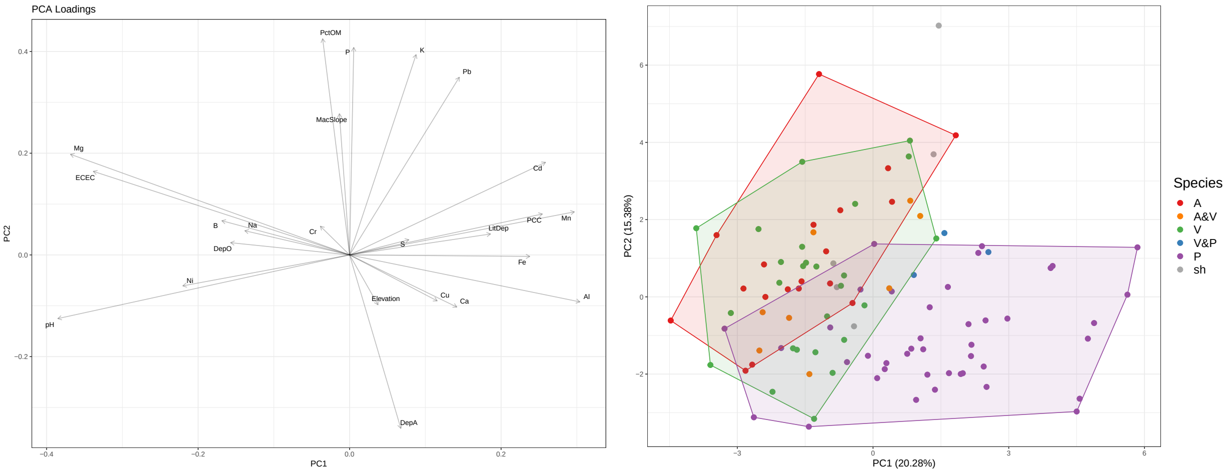
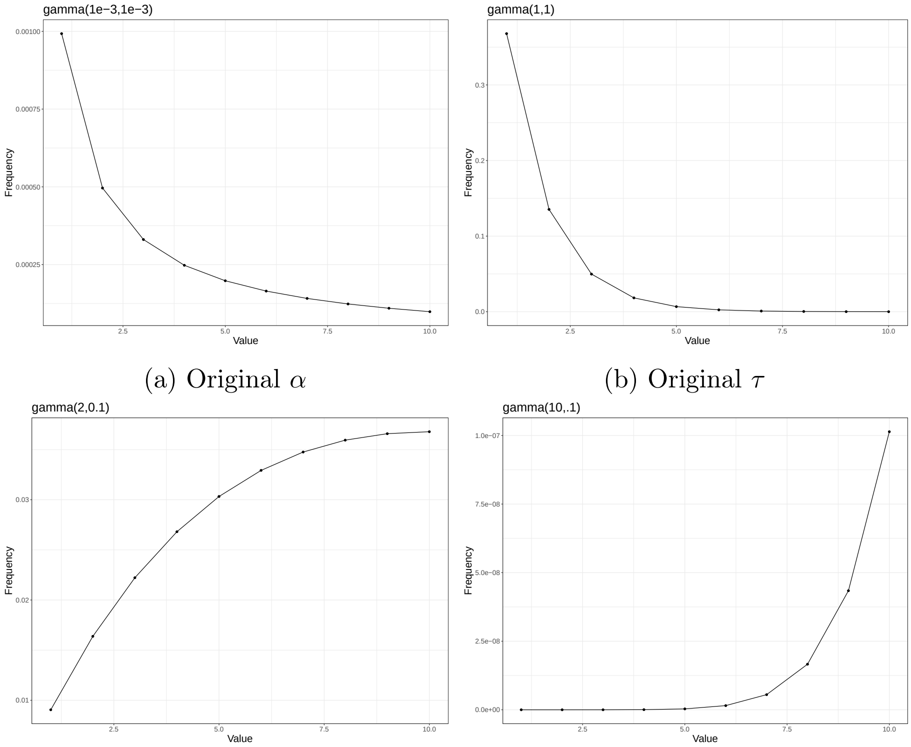


Figure 4: Plots of the loadings (left) and scores (right) along the first and second components derived from standard PCA.

3.1.2 Model Selection

The original priors over α and τ ([4]) are somewhat broad, but have the highest probability in regions of low value parameter space. To explore the possibility that the parameters are more accurately described by a different prior distribution, the shape and rate parameters of the gamma distribution for α and τ were modified from the original values of $gamma(1e-3, 1e-3)$ and $gamma(1, 1)$, respectively (5). The prior over α was increased to $gamma(2, 0.1)$, and the prior over τ was increased to $gamma(10, 0.1)$. The change in both parameters was considered separately and jointly, comprising three models plus the original (Table 1). The variables were reconstructed from each model, scaled by $1/\tau$, and visualized against the original data in order to assess model fit. Model 2, in which the prior over α is unchanged



(a) Original α

(b) Original τ

(c) Modified α

(d) Modified τ

Figure 5: Distribution of the original and modified gamma priors over α and τ; top row shows the original priors, bottom row shows the modified priors, left column shows the priors over α, and right column shows the priors over τ. Specified as in stan ([20]), not by rgamma() in R.

and the prior over τ is increased, was selected as the final model because it shows the best reconstruction of the scale of the original variables (6). The ranges of the reconstructed variables from the original model are much smaller than those of the actual variables (19). The ranges of the reconstructed variables from the third model are comparable to that of the original model (20. The ranges of the reconstructed variables from the fourth model are better than those of the first and third models, but worse than the second (21). Changing the prior over τ increased the estimate of this parameter from 5.02 in the original model to 489.3 in the second model, and 485 in the fourth model. In the third model, where the prior distribution of τ was unchanged from the original, the parameter estimate nevertheless increased to 29.1. Changing the prior over α decreased the estimate of this parameter, from a mean of 90.7 in the original model to 26.3, 25.0, and 23.9 in the second, third, and fourth models respectively (1.

Model Name	Alpha Prior	Sigma Prior	Alpha Post.	Sigma Post.
Original	gamma(1e-3,1e-3)	gamma(1,1)	90.7	5.02
Model 2	gamma(1e-3,1e-3)	gamma(10,0.1)	26.3	489.3
Model 3	gamma(2,0.1)	gamma(1,1)	25.0	29.1
Model 4	gamma(2,0.1)	gamma(10,0.1)	23.9	485.2

Table 1: The impact of prior distribution on model fit. The prior distribution for each parameter is listed for each model, along with the mean posterior parameter estimate

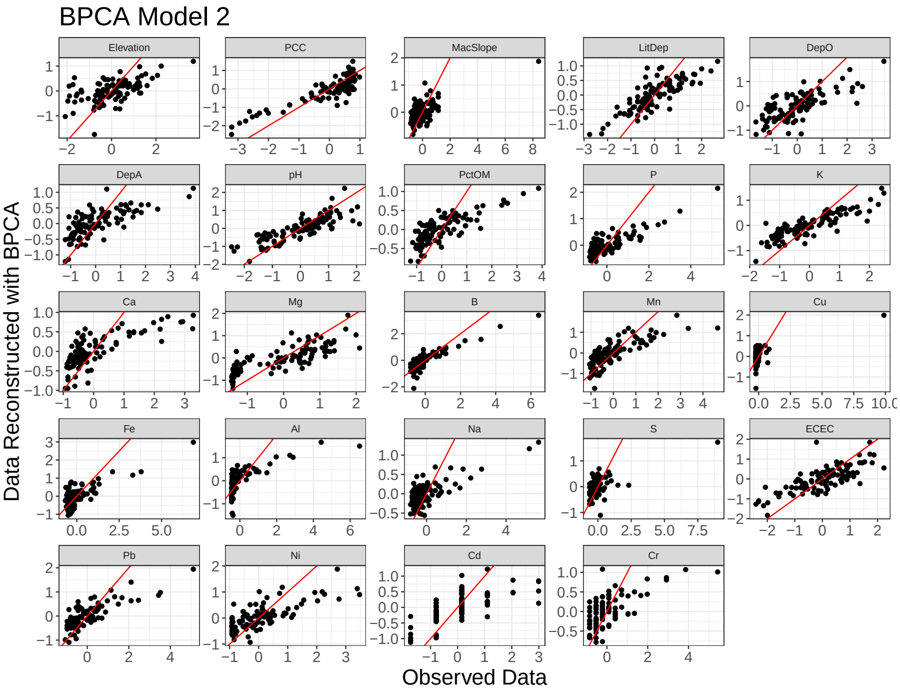


Figure 6: Observed vs reconstructed data for the second model, with τ $gamma(10, 0.1)$, and the prior distribution for α unchanged. An identity line with slope of 1 and intercept of 0 is shown in red.

3.1.3 Stan Diagnostics

Seven generated parameters out of 2,970 had a Rhat value greater than 1.01, with a maximum value of 1.013. Trace plots for the values of W with Rhat greater than 1.01 showed poor chain convergence in comparison to another randomly selected value (Figure [7]).

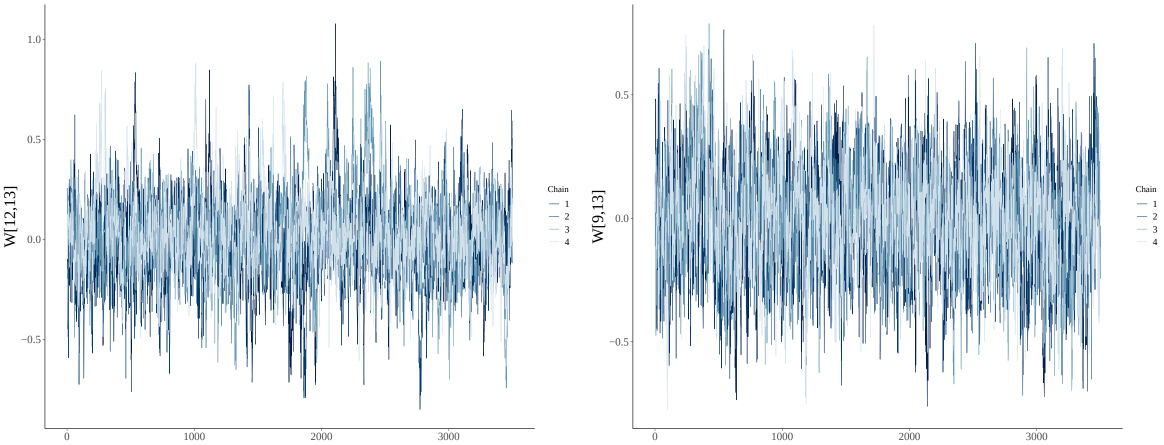


Figure 7: Trace plots of chain convergence. Left: Selected W parameter with Rhat value greater than 1.01. Right: randomly selected W parameter with Rhat less than 1.01.

3.1.4 Model Diagnostics

Analysis of the effective dimensions as quantified by the value of α for each eigenvector was inconclusive, as all were within a similar range of values, and none zeroed out(9). Components 11, 13, 15, and 21 had the highest transformed alpha score. The Hinton diagram was also inconclusive in terms of dimensionality selection, with most latent variables possessing a similar magnitude(8). However, based on the organization of the latent variables by magnitude, the components with the highest α score are mostly contained in the left hand section of the plot, corresponding to the eigenvectors of greatest magnitude.

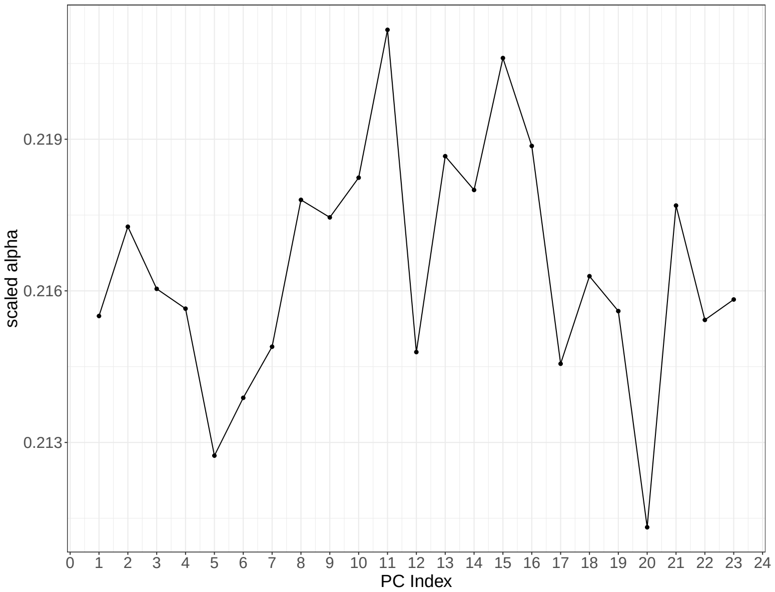


Figure 8: The transformed values of the hyperparameter α corresponding to each column of the matrix W.

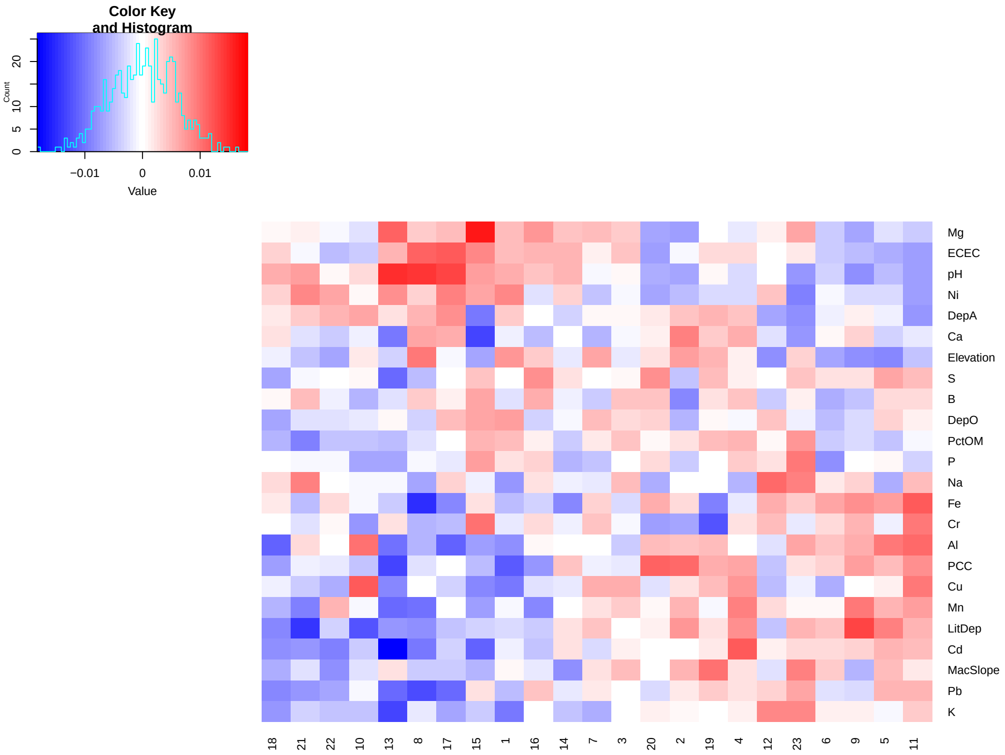


Figure 9: Hinton plot of W, showing the 23 latent variables (columns) and 24 original variables (rows), sorted in both dimensions by the magnitude of the variance. Each cell represents the contribution of each variable to each principal component. Blue represents negative loading of variables, red represents positive loading, and the intensity of the color represents the magnitude.

3.1.5 Model Output

The pairs plot of the components identified as important in the alpha analysis show the most niche separation along components 11 and 13 (10), and so these were selected for further study. A plot of variable contribution to the eleventh and thirteenth principal components(11) shows Cr, Fe, and Al with a strong positive contribution to the first principal component, and Depth A, Ni, and Mg with a strong negative contribution. Along the thirteenth principal component, Ca, P, and Elevation have a moderate negative contribution, and MacSlope has a moderate positive contribution.

The plot of scores along the eleventh and thirteenth principal components show a slight pattern of niche separation among the three species(11, though the niche of *Adiantum pedatum* as represented encompasses the niches of both other species. Some occurrences of *Adiantum pedatum* occupy portions of niche space corresponding to habitats with a higher PCC (percent canopy cover), and higher concentrations of soil Al and Fe, than the two serpentine maidenhair ferns. Occurrences of *Adiantum aleuticum* occupy portions of niche space corresponding to habitats with a higher concentration of soil Mg, Ni, and ECEC. and higher concentrations of Ca and K in the soil. Occurrences of *Adiantum viridimontanum* overlap with those of both its progenitors in niche space.

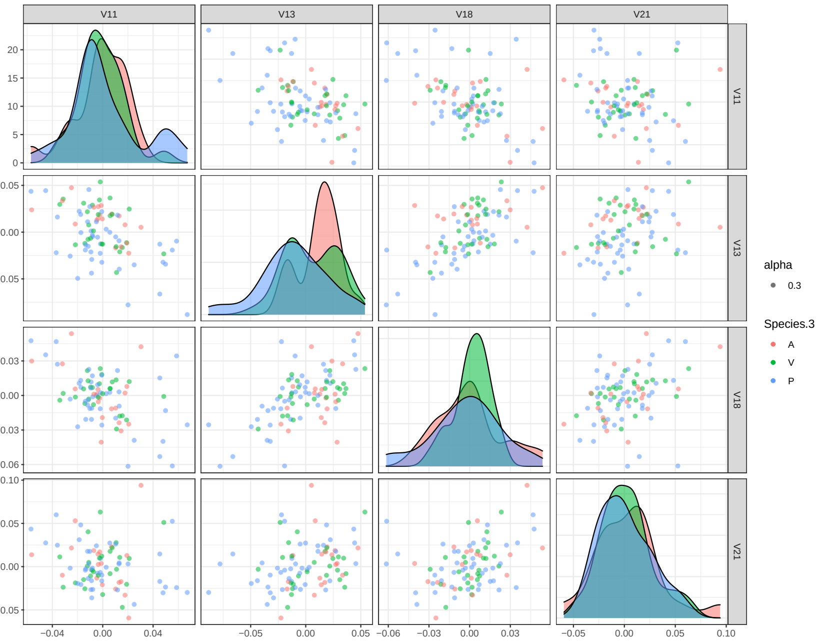


Figure 10: Pairs plot of the principal components with the highest transformed alpha score.

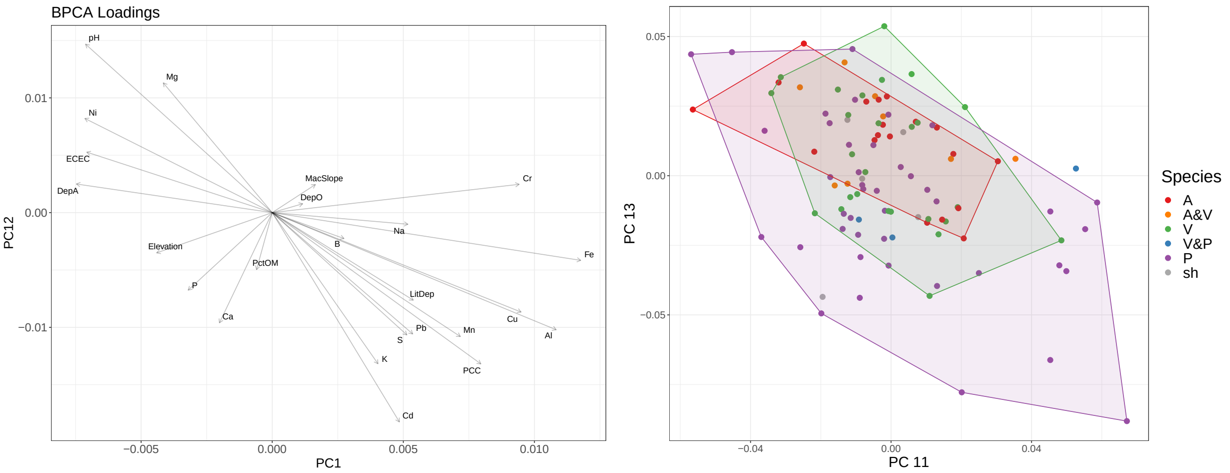


Figure 11: Plots of the loadings (left) and scores (right) along principal components 11 and 13 derived from BPCA.

3.1.6 Model Comparison

The data reconstruction on the basis of 7 principal components for standard PCA and all components for BPCA showed a much comparable model fit (12). The MSE for standard PCA was -3.15e-16, whereas the MSE for BPCA was 2.19e-05. However, PCA seemed to yield a more accurate fit for some variables in particular.

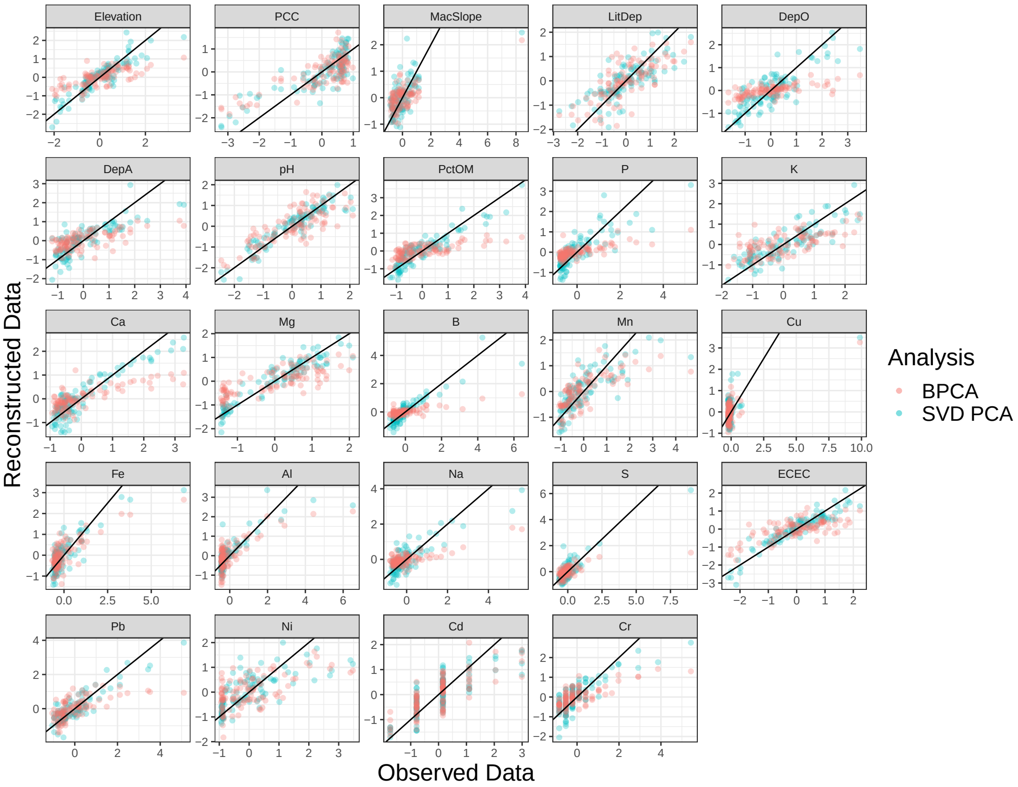


Figure 12: Observed vs reconstructed variables for both PCA (blue) and BPCA (pink). An identity line with slope of 1 and intercept of 0 is shown in red.

3.2 MNIST

A subset of the original, reconstructed, and generated handwritten digits using the original model (see Table 2) are shown below. The top row consists of the original images (row 1), followed by the conventional PCA reconstruction using the effective dimensions found by BPCA (row 2), a BPCA reconstruction using the effective dimensions (row 3), a BPCA reconstruction using all the estimated columns of the principal subspace (row 4), and the images generated from the fitted BPCA model (row 5, also using the effective number of dimensions). Target labels are provided for the first four rows of images since we are approximating the original data given W, while the generated dataset on the last row do not since each image is a draw from a Multivariate Normal distribution.

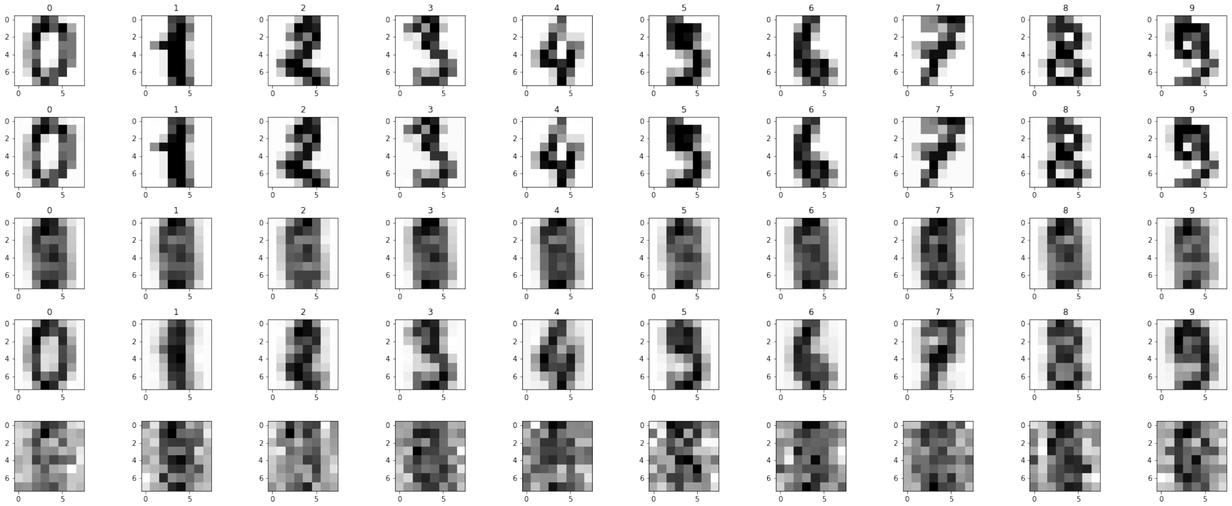


Figure 13: Reconstructed and generated handwritten digits (Original model)

The PCA reconstruction using the effective dimensions suggested by BPCA has the lowest mean-squared error (Figure [14]) compared to the BPCA reconstructions and and looks the most visually similar to the original images.

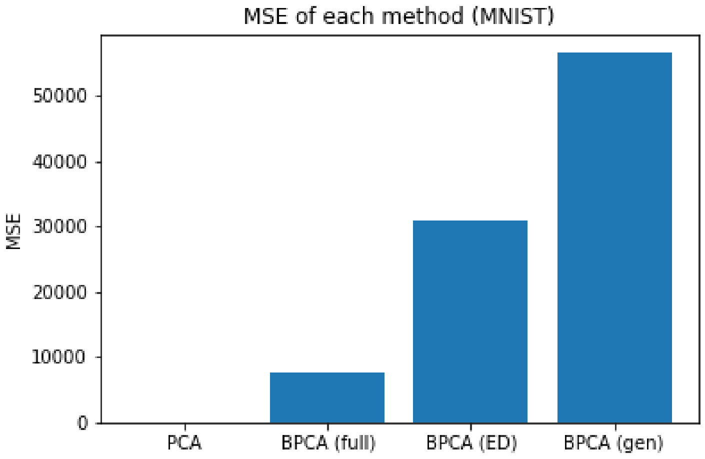


Figure 14: Mean-squared error of Original model

Trying different non-informative priors [2] yields similar results (Model 2, 3, 4). However, when we restrict the priors to a tighter range, the MSE for each BPCA method reduces dramatically (Model 5, 6, 7). We also see that the MSE for the PCA reconstruction increases as we use a smaller number of effective dimensions while the MSE for the BPCA reconstruction and generation decreases.

Model 7 has the lowest MSE for BPCA using the full-rank reconstruction as well as the lowest effective number of dimensions. The full-rank BPCA reconstruction here outperforms the PCA reconstruction with the effective number of dimensions. Figure [15] shows the original and reconstructed images using model 7. We can see that there are less disparities between the original images and the BPCA reconstructed/generated images.

*Note that for gamma(a,b), a=*shape* and b=*scale* (1/*rate*)				**MSE**			
Model	**Alpha**	**Tau**	**ED**	**PCA (ED)**	**BPCA (Full)**	**BPCA (ED)**	**BPCA (gen)**
Original	gamma(0.001, 0.001)	gamma(1, 1)	50	15.27	7799.04	30976.28	56673.90
Model 2	gamma(1, 0.1)	gamma(1, 1)	46	106.37	18141.88	33260.034	122558.11
Model 3	gamma(0.001, 0.001)	gamma(1.25, 0.3)	50	15.27	7874.68	30861.18	56752.24
Model 4	gamma(1, 0.1)	gamma(1.25, 0.3)	48	47.71	18515.74	33099.31	126959.27
Model 5	gamma(1000, 500)	gamma(1000, 500)	34	910.82	2465.77	11336.08	43369.41
Model 6	gamma(1000, 1000)	gamma(1000, 1000)	31	1254.32	1427.72	9814.82	40598.61
Model 7	gamma(10000, 10000)	gamma(10000, 10000)	30	1380.37	1272.05	9831.99	37101.20
Model 8	gamma(10000, 10000)	gamma(1, 1)	49	15.27	7839.96	30938.90	38924.30

Table 2: PCA and BPCA performance as measured by the mean-squared error between the original standardized data and the generated or reconstructed data.

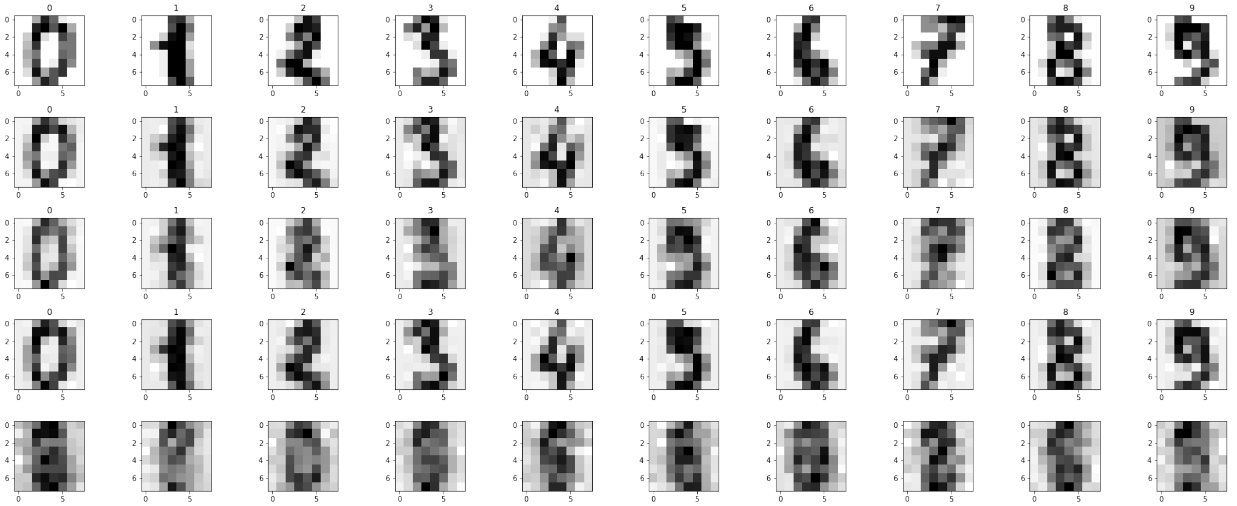


Figure 15: Reconstructed and generated handwritten digits (model 7)

Again the top row consists of the original images (row 1), followed by the conventional PCA reconstruction using the effective dimensions found by BPCA (row 2), a BPCA reconstruction using the effective dimensions (row 3), a BPCA reconstruction using all the estimated columns of the principal subspace (row 4), and the images generated from the fitted BPCA model (row 5, also using the effective number of dimensions).

4 Discussion

4.1 Summary

The original BPCA model, as specified by [4], yielded a poor model fit in comparison to PCA for both the *Adiantum* and MNIST data-sets. However, changing the prior distribution over the parameters yielded a better model fit for both data-sets, specifically when the shape parameter was large and the rate parameter small (giving us a narrow range of values) for the gamma priors over α and τ.

4.2 Adiantum Results

Restricting the prior distribution over τ to higher parameter values yielded a better model fit for the *Adiantum* dataset than the original BPCA model. Using this altered model, the model fit for standard

PCA and BPCA are comparable. However, PCA has a slightly lower MSE than BPCA. This can be seen in terms of how the plot of the reconstructed vs observed data shows a slightly tighter correlation for PCA than BPCA (12). The scale of the BPCA reconstruction, though close to that of PCA, appears to be slightly off for certain variables. A better specification of τ has the potential to yield better results.

For both BPCA and PCA models, the data reconstruction was more accurate for some variables than others, including Elevation, LitDep, pH, ECEC, K, and Cr. Of these variables, all but Cr have distributions that could somewhat accurately be approximated with a normal distribution. This result highlights the importance of the underlying data distribution in shaping the model fit for both normal and bayesian PCA. BPCA assumes that the data and latent variables are generated from multivariate normal distributions. Few of the variables in the *Adiantum* dataset are normally distributed, and some exhibit a logarithmic and even bimodal distribution (16). However, Nounou et al. ([15]) note even variables that are not normally distributed can often still be approximated with a Gaussian density.

In terms of model output, the current implementation of BPCA describes less niche separation among the three *Adiantum* species than does normal PCA. A better model fit would likely reveal similar patterns as described by PCA.

4.3 MNIST

PCA outperformed BPCA when using the original priors (gamma(shape=0.001, scale=0.001)) over α and τ. However we see improvement for BPCA with a narrower range of prior values as indicated by the MSE of the reconstructions and generations (model 5, 6, 7). We expected the MSE of the PCA reconstruction to increase as the number of effective dimensions decreases since PCA yields a deterministic solution where each additional dimension increases the rank of the reconstructed data matrix, thus returning more information. The situation for BPCA is more complicated since we see that decreasing the effective number of dimensions actually decreases the MSE of the reconstruction when priors were narrow. This held for both the full-rank BPCA reconstruction as well as the lower dimensional counterpart using the effective number of dimensions found by the BPCA model.

These effects were driven by the priors we used. A broad, unspecific range of priors returned suboptimal reconstructions and data generations from BPCA. This seems to be due to the noise term τ in the estimated covariance matrix, i.e. $C = WW^T + \tau^{-1}I_D$. When τ is larger the covariance matrix from BPCA is noisier and looks less like the cleaner covariance matrix of PCA. On the other hand, when τ is drawn from a narrow range of values, the estimated BPCA covariance matrix more accurately describes the patterns in the data and closely resembles the PCA estimated covariance matrix. It also appears that W isn't sensitive to the prior values of α in comparison to τ.

4.4 Future Directions

Implementation of BPCA is challenging, but promising; these preliminary results show the potential of utilizing the generative story of BPCA as a probabilistic model. There are several lines of inquiry that may yield more precise and accurate results. First, troubleshooting the dimensionality aspect of BPCA, in order to gain a better sense of which components to retain, would likely increase the utility of this technique, especially for the *Adiantum* dataset. Improved dimensionality selection in the BPCA implementation would likely help to identify which components describe meaningful variation, and which describe noise. Second, setting a prior over μ, as well as over σ and α, has the potential to yield more accurate estimates of the posterior distribution. A prior over μ could potentially also function to improve the accuracy of how the reconstructed variables are re-scaled to match the scale of the original ones. Third, there is likely valuable information in the data that could be utilized in order to set more accurate priors over μ and σ. Nounou et al. (2002) describe a parametric approach for estimating the prior empirically which takes the range and distribution of the underlying measurements into account [15]. This approach could facilitate an iterative model fit where the variance and distribution of each variable is taken into account individually. Fourth, it seems possible that the BPCA model could be fit to assume a non-normal distribution. However, evaluating the effects of such shifts on the assumptions of principal component analysis presents a complex problem. Would using a non-normal distribution for the model violate the required orthogonality of the principal axes? Such inquiries are beyond the scope of this project. Fifth, there are versions of factor analysis and machine learning models that assume a nonlinear relationship between the latent variables and the output data that certainly warrant further exploration [13, 14] and may yield a better generative story.

5 Personal Statements

Morgan: I brought the idea for the project as an important next step in my thesis research. I had previous knowledge of PCA, which provided a foundation from us to work from. I focused on developing a work-flow in R based on the stan model provided by [12], and comparing it with the results obtained via standard PCA. Additionally, I provided the context in terms of 1) the data-set in question, 2) how multivariate analysis relates to ecological theory, and 3) what the implications of our findings are for my study system.

Phil: I dug into and summarized the mechanics of PCA, PPCA, and BPCA. I helped formalize the analyses and suggested graphics to incorporate. I implemented a Python workflow to contrast our Stan MCMC approach with and ran our analyses on and discussed the results of the MNIST dataset.

References

[1] amoeba (https://stats.stackexchange.com/users/28666/amoeba). *Making sense of principal component analysis, eigenvectors and eigenvalues.* Cross Validated. eprint: https://stats.stackexchange.com/q/140579.

[2] Hervé Abdi and Lynne J. Williams. “Principal component analysis”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* (2010). https://doi.org/10.1002/wics.101.

[3] amoeba. *How to reverse PCA and reconstruct original variables from principal components.* website. 2014.

[4] Christopher M Bishop. “Bayesian PCA”. In: *Advances in neural information processing systems* (1999), pages 382–388.

[5] Christopher M Bishop. “Variational principal components”. In: (1999).

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[7] Jason Brownlee. “A Gentle Introduction to Expectation-Maximization”. In: (2019). https://machinelearningmastery.com/expectation-maximization-em-algorithm/.

[8] José Camacho and Alberto Ferrer. “Cross-validation in PCA models with the element-wise k-fold (ekf) algorithm: theoretical aspects”. In: *Journal of Chemometrics* 26.7 (2012), pages 361–373.

[9] Manish Reddy Vuyyuru Feiyu Chen Jianzhun Du. *Probabilistic and Bayesian PCA: Numerical Methods Project.* 2018.

[10] G Evelyn Hutchinson. “Cold spring harbor symposium on quantitative biology”. In: *Concluding remarks* 22 (1957), pages 415–427.

[11] Ian T. Jolliffe and Jorge Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A* (2016). https://doi.org/10.1098/rsta.2015.0202.

[12] Joseph H. Sakaya Suleiman A. Khan. *Probabilistic Factor Analysis Methods*. website. 2017. eprint: https://stats.stackexchange.com/q/140579.

[13] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2014. arXiv: 1312.6114 [stat.ML].

[14] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. probml.ai.

[15] Mohamed N. Nounou, Bhavik R. Bakshi, Prem K. Goel, and Xiaotong Shen. “Bayesian principal component analysis”. In: *Journal of Chemometrics* 16.11 (2002), pages 576–595. https://doi.org/https://doi.org/10.1002/cem.759.

[16] Cathy A Paris. “Adiantum viridimontanum, a new maidenhair fern in eastern North America”. In: *Rhodora* (1991), pages 105–121.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. https://www.R-project.org/.

[18] Jonathon Shlens. “A Tutorial on Principal Component Analysis”. In: (2014). arXiv: 1404.1100 [cs.LG].

[19] Morgan W Southgate, Nikisha R Patel, and David S Barrington. “Ecological outcome of allopolyploidy in Adiantum (Pteridaceae): Niche intermediacy and expansion into novel habitats”. In: *Rhodora* 121.986 (2019), pages 108–135.

[20] STAN. *Stan Reference Manual: Gamma Distribution*. website.

[21] Michael E Tipping and Christopher M Bishop. “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pages 611–622.

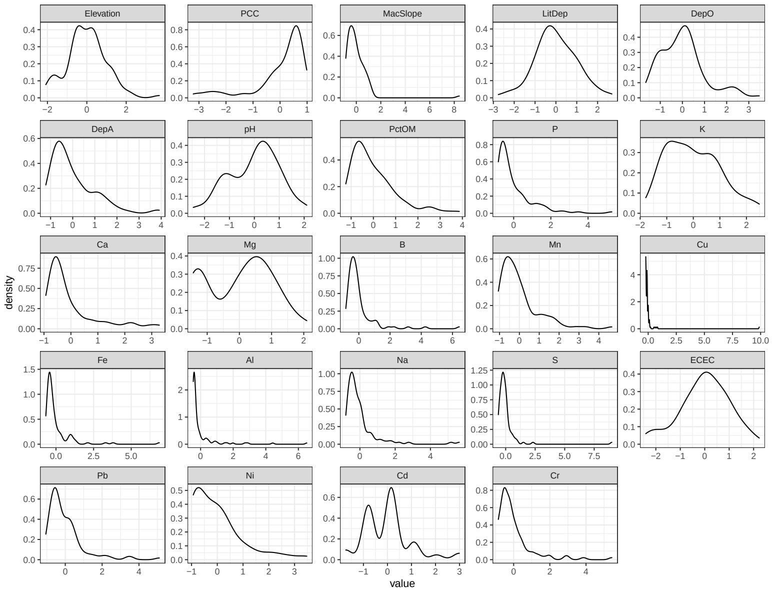


Figure 16: Distribution of the 24 centered and scaled ecological variables comprising the *Adiantum* data-set.

6 Supplementary Information

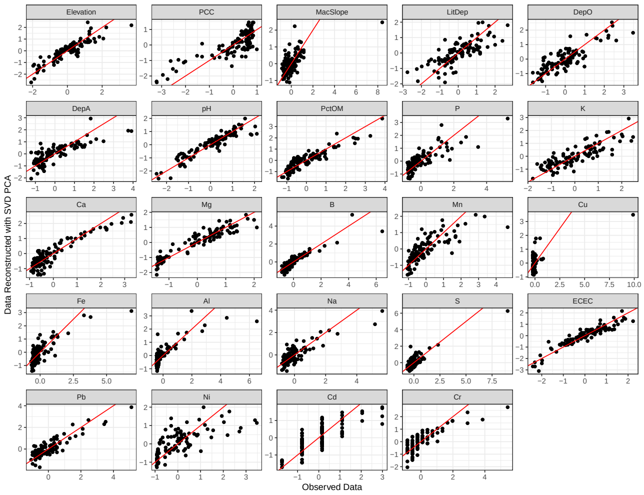


Figure 18: Observed vs reconstructed data for standard PCA. Line with slope of 1 and intercept of 0 is shown in red.

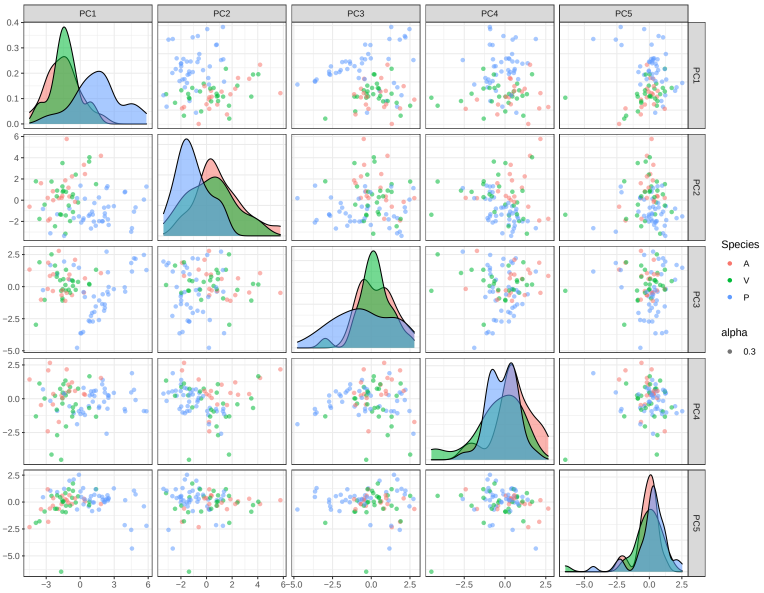


Figure 17: Pairwise plot of the first five principal components obtained from SVD PCA

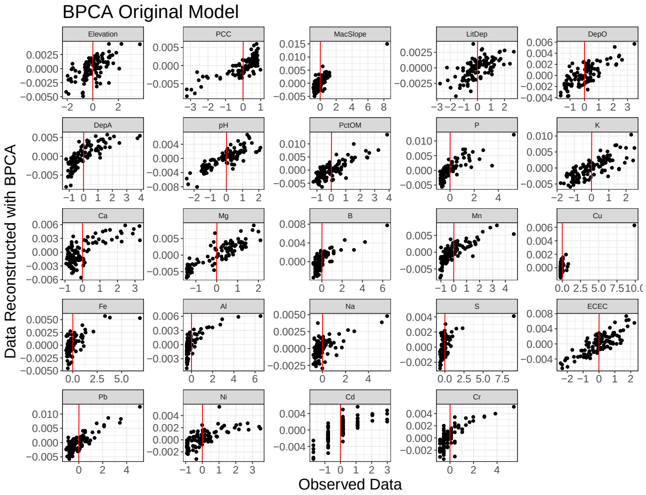


Figure 19: Observed vs reconstructed data for the original model as specified by Bishop, with α $gamma(1e-3, 1e-3)$ and σ^2 $gamma(1,1)$. An identity line with slope of 1 and intercept of 0 is shown in red.

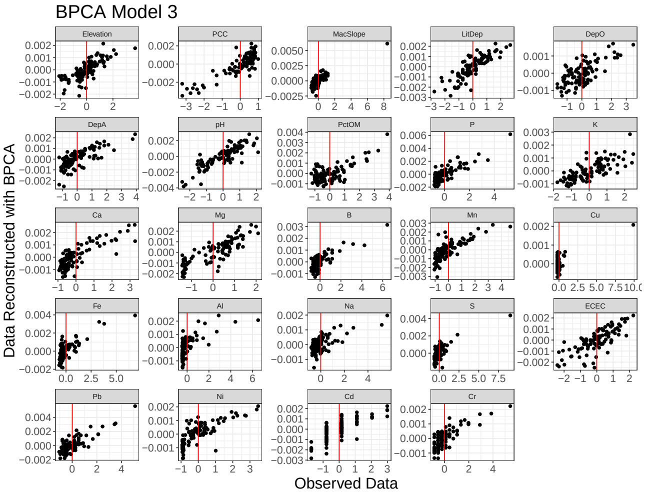


Figure 20: Observed vs reconstructed data over σ as the original model and a prior α distribution of gamma(2,0.1). An identity line with slope of 1 and intercept of 0 is shown in red.

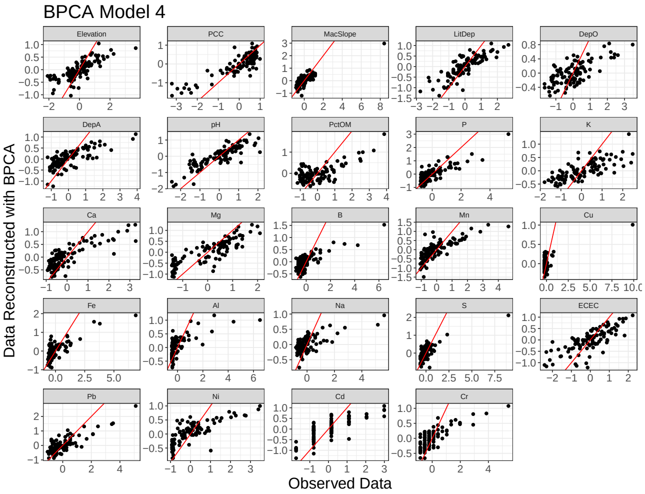


Figure 21: Observed vs reconstructed data for the fourth model, σ $gamma(10, 0.1)$ and α $gamma(2, 0.1)$. An identity line with slope of 1 and intercept of 0 is shown in red.