University of Colorado Anschutz Medical Campus

Kendra Ferrier
Pre-Doctoral Student
Human Medical Genetics & Genomics Program
University of Colorado-Anschutz Medical Campus
https://github.com/k2ferrier/
kendra.ferrier@cuanschutz.edu

Final Project

MATH 3191 : Prof. Moses Khan
Wednesday, December 12, 2019

Contents

Introduction

Currently, the strongest risk factor for melanoma is the number of nevi (moles) a person has [9]. The majority of nevi develop during childhood and adolescence and the number of nevi developed is influenced by a variety of environmental, behavioral, genetic, and phenotypic factors [2, 14]. While the relationship of environmental and behavioral factors such as UV exposure and use of sun protection, respectively, have clear relationships with nevus development, the impact of impact of each factor in conjunction on nevus development needs further investigation.

In a longitudinal observational study on mole development in children from Colorado, data was collected on the child's mole count, base skin color, eye color, hair color, ethnicity (Hispanic or non-Hispanic), number of seaside vacations since birth, and OCA2 genotype. OCA2, a gene important for melanin synthesis, has been associated with heritability in nevus count, skin color, and eye color [3, 8]. More specifically, individuals homozygous for the g allele demonstrated a significant correlation between having higher nevi counts, blonde hair, and blue or green eyes. Whereas individuals homozygous for the a allele tended to have fewer nevi, darker hair, and brown eyes [8, 12]. In addition to collinearity between genotype, eye color, and hair color, there is also a colinear relationship between base skin color and ethnicity. While a linear regression model excluding potential collinear variables can be conducted, there may be variation that can be explained by the factors such as eye color or hair color that is not explained by OCA2 genotype. Therefore, a method for estimating the coefficients of collinear variables is necessary to appropriately analyze this dataset.

Methods for Regression of Collinear Data

Principal component analysis (PCA) is a powerful technique used in many fields to reduce high-dimensional data into fewer variables that can best explain the variance in the data. This is achieved by performing an orthogonal transformation on a set of variables that are likely correlated to yield a set of linearly uncorrelated variables (principal components, or PCs) [13]. The first PC describes the greatest amount of variability in the data as possible, and each subsequent PC describes the next highest variance orthogonal to the preceding component. The output of PCA is an orthogonal basis set of vectors (each comprised of a linear combination of variables with n observations) which can be used to predict future data, or as covariates within linear models to account for undefined structures that would confound regression analysis. For example, in a genetic case-control disease study, differences in allele frequency due to genetic ancestry can cause spurious associations if genetic ancestry is not accounted for in the model [11]. Genetic ancestry explains much of the variance in genetic data and thus will be represented in the first several PCs in a PCA of the dataset. These PCs can then be used as a proxy for genetic ancestry and added as covariates to the model, thus correcting for ancestry as confounding factor [10]. In addition to PCs being useful as covariates for linear regression models, variations on PCA, such as principal component regression (PCR) can be used to address multicollinearity.

PCR is a regression analysis method based on PCA that uses PCA for estimating unknown regression coefficients in a model. In PCR, the PCs of the explanatory variable are used as regressors instead of regressing the dependent variable on the explanatory variable directly (as is done in a standard linear regression model) [7]. One of the limitations of PCA based dimensionality reduction is that if two or more variables in the dataset are highly correlated, the columns corresponding to these variables become linearly dependent and the matrix itself becomes rank deficient. Collinearity is handled in PCR by excluding low-variance (nearly or exactly zero eigenvalue) PCs in the regression part of the process. Depending on the purpose of the analysis, other low-variance PCs can also be excluded in the regression to reduce the effective number of parameters defining the model and mitigating potential overfitting.

However, there is a quite significant limitation to PCR in that it does not consider the response variable when deciding which principle components to drop (in the case of there being collinearity), i.e. just because a factor has a small eigenvalue does not mean it is not a strong predictor of the response variable [6]. This means that when PCs are dropped the coefficient estimates become biased, more difficult to interpret (because they are weighted averages), and less accurate for predicting new data [1]. It is possible for important explanatory variables to be given small eigenvalues (and vice versa with unimportant variables) and estimated coefficients in the wrong direction. In summary, PCR is an effective and easy method for identifying collinearity, though other methods should be used to estimate the coefficients, such as a partial least squares (PLS) regression or a multivariate linear regression using only the linearly independent variables.

PLS regression is similar to a PCR, except that instead of finding hyperplanes of the highest variance between the response and independent variables, it analyses the relationship between the response and independent variables as matrices [4]. More specifically, the PLS regression finds the multidimensional direction of the independent variables that explains the greatest amount of multidimensional variation direction in the response variable matrix [5]. Since this model incorporates the response variable in the selection of PCs that explain the most variance, it is the most appropriate for PCA-like regression of datasets that are rank deficient or have collinearity among variables, and for providing accurate coefficient estimates.

Materials & Methods

Materials

Summary of Moles Cohort

genotype	hispanic	gender	n	eye color: BG	eye color: Br	eye color: H	hair color: B	hair color: R	hair color: Br	hair color: Bl	mean base skin color	mean vacations	mean mole count
aa	N	F	11	0	9	2	1	0	9	1	11.70	3.73	44.36
aa	N	M	17	0	15	2	3	3	11	0	10.72	3.47	48.47
aa	Y	F	6	1	5	0	0	0	5	1	12.58	2.17	25.33
aa	Y	M	8	0	8	0	0	0	5	3	11.39	1.38	31.00
ga	N	F	80	7	34	39	33	6	40	1	10.96	2.80	49.30
ga	N	M	61	9	20	32	27	4	30	0	10.17	3.02	54.49
ga	Y	F	15	2	13	0	3	1	10	1	11.88	2.53	29.40
ga	Y	M	14	0	12	2	0	1	12	1	12.18	2.50	35.50
gg	N	F	108	95	0	13	70	7	31	0	10.70	2.64	52.46
gg	N	M	87	69	0	18	58	3	26	0	9.99	2.47	58.97
gg	Y	F	4	3	0	1	1	1	1	1	12.92	1.00	60.25
gg	Y	M	3	2	0	1	2	0	1	0	11.62	1.00	30.00

Table 1: Characteristics of the subset of children analyzed from the longitudinal observational study of mole development in children from Colorado. Characteristics include OCA2 genotype (gg, ag, or aa), hispanic (Y/N), gender (M/F), eye color (BG = Blue, Green, or a combination of both, Br = Brown, H = Hazel), hair color (B = Blonde, R = Red, Br = Brown, Bl = Black), base skin color (colorimeter measurement of skin pigmentation), number of seaside vacations, and number of moles at age 6.

This study was conducted using a subset of data from a longitudinal observational study of children from Colorado followed from 6 to 10 years of age. This subset includes annual measurements of mole count, OCA2 status, eye color, base skin color, hair color, ethnicity (Hispanic or not Hispanic),and reported number of seaside vacations since birth for 472 children. Nevi $\geq$5mm is size were counted during a skin examination by health care providers trained annually by the study's lead dermatologist. OCA2 status was measured by

PCR amplification and genotyping of rs12913832 from saliva DNA samples. Eye color, base skin color, hair color, and ethnicity were recorded during the baseline visit at age 6. Number of seaside vacations since birth was assessed via an annual survey of the children's parents where 'waterside location' was explicitly defined.

For the purposes of this analysis, only the mole count and number of seaside vacations reported from the baseline visit at 6 years of age was used. In PCA, genotype is treated as a continuous variable, whereas in PCR and PLS the dummy variables are treated as categorical. A summary of the subsection of the cohort used in this analysis can be viewed in Table 1.

PCA

The first step to principal component analysis is to calculate the covariance matrix of the data. Analysis of variances can be used to answer questions about whether the variable(s) in X have a relationship with Y.

To begin, the mean of all the m variables from an $m \times n$ matrix, A, is calculated and stored in a single vector in $\mathbb{R}^m$, u:

$$\vec{\mu} = \left(\frac{1}{(n-1)}\right)(\vec{x}_1 + \vec{x}_2 + \ldots + \vec{x}_n)$$

Then, the data can be re-centered/normalized by subtracting the mean vector from the matrix of observations:

$$B = \left[\vec{x}_1 - \vec{\mu} \;\middle|\; \ldots \;\middle|\; \vec{x}_n - \vec{\mu}\right]$$

From the re-centered matrix, B, the covariance matrix, C, can be computed:

$$C = \frac{1}{(n-1)} BB^T$$

C will be an $m \times m$ symmetric matrix (based on the properties of matrix multiplication) that contains the variance for each variable along the diagonal entries and the covariances between variables in the other entries of the matrix . Once the covariance matrix has been calculated, spectral theory can be applied to the matrix to calculate eigenvalues and produce orthogonal vectors. Spectral theory states that for a symmetric matrix, S, there exists real eigenvalues and nonzero orthogonal vectors such that:

$$S\vec{v}_i = \lambda_i \vec{v}_i \text{ for } i = 1, 2, \ldots, n$$

These orthogonal vectors/eigenvectors are the principal components of the dataset.

PCR

Principal component regression can generally be explained in three steps: First, perform PCA (detailed above), then use ordinary least squares regression to regress the dataset with principal components as covariates to create a vector of estimated regression coefficients (if a PC contains extremely small eigenvalues with a high variance inflation factor, the PC can be omitted to address multicollinearity), and lastly, transform the estimated coefficients to the scale of the original covariates by multiplying the vector by the PCA loadings. This vector of coefficients represents the weights that reflect the covariance between predictor variables.

The general framework for a linear regression in matrix format is described in equation 1 below where the vector of Y's represents the measured values for the response variables, the matrix with x's represent the

measured values for each predictor variable (each column represents a variable, each row an observed value for an individual), the vector of β's is the estimated coefficients determined by the linear regression, and the vector of ϵ's are the error terms. In this analysis, the Y vector corresponds to the mole count for each individual, the X matrix are the observations for genotype (factored, with genotype gg as the intercept), gender, Hispanic status, eye color, hair color, base skin color, and number of seaside vacations, and the values of the β vector are being estimated by the regression . Equation 2 is an equivalent representation of a linear regression model in non-matrix format.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_3 \end{pmatrix} \tag{1}$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{2}$$

PLS

PLS is and PCR are similar in that both produce factor scores from linear combinations of the original variables. The difference between them is in how the scores are factored. The general model for partial least squares regression decomposes the matrices of responses, Y, and the predictors X, into projections V and W, respectively, which have maximal covariance between them. V and W are then transformed by the orthogonal loading matrices R and P, respectively. Error terms, D and E, are also added to both the response and predictor models. The resulting vector of coefficients from this method represents the weights that reflect the covariance structure between the predictor and response variables.

$$X = WP^T + E$$

$$Y = VR^T + D$$

Performing PCA, PCR, and PLS

Data manipulation and linear regression analysis was conducted using R v3.6.1. to assess the contribution of different factors in the mole counts of 6 year-old children from Colorado. PCA was conducted using the princomp function from the package stats v3.6.1 (cor = T) /citestats. PCR and PLS was performed with the pcr and plsr functions from the pls package v.2.7-2 /citepls. Default settings were used for both analysis, with the addition of scale = T for PCR. Code for the entirety of this analysis can be accessed at `https://github.com/k2ferrier/MATH3191/blob/master/Final_project_lin_alg.Rmd`.

Results

PCA

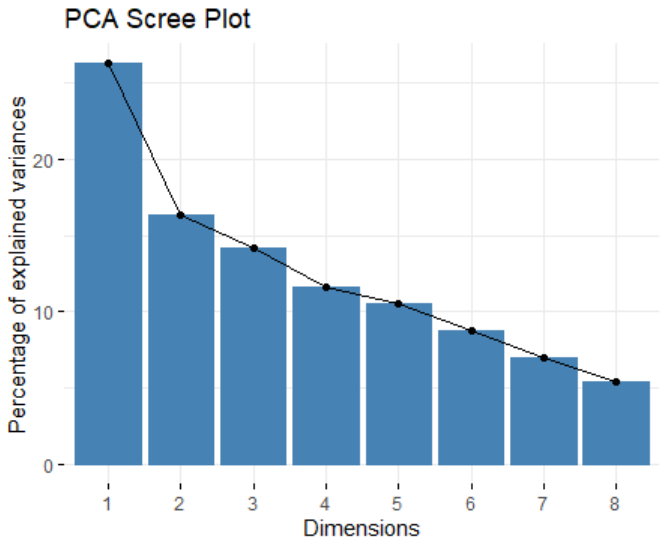

Figure 1: Scree plot of percent variation explained as a function of number of PCs included from PCA of the entirety of the moles dataset (including the variable mole count).

From the PCA biplot of PCs 1 and 2 in Fig. 2, genotype corresponds to three different clusters where the red (0 in the figure legend), green (1), and blue (2) correlate to the genotypes gg, ga, and gg, respectively. Additionally, we can see that eye color, genotype, and hair color all have similar direction and magnitude, though genotype shows the greatest magnitude of the three. Similarly, Hispanic status and base skin color display a similar direction and magnitude of effect, where Hispanic status has the larger magnitude in PC1 and base skin color has the larger magnitude in PC2. These groupings are unsurprising given their known biological relationship and provide further evidence supporting their collinear association. Gender and number of seaside vacations both show independent direction and magnitude of effect in these PCs.

Despite the presence of multicollinearity, the scree plot describing the percent variance explained as a function of number of PCs shows that all PCs have non-zero contributions in explaining the variance (Fig. 1). This validates further investigation of the contribution of each factor in explaining mole count in principal component and partial least squares regression.

Results [PCA] continued on next page...

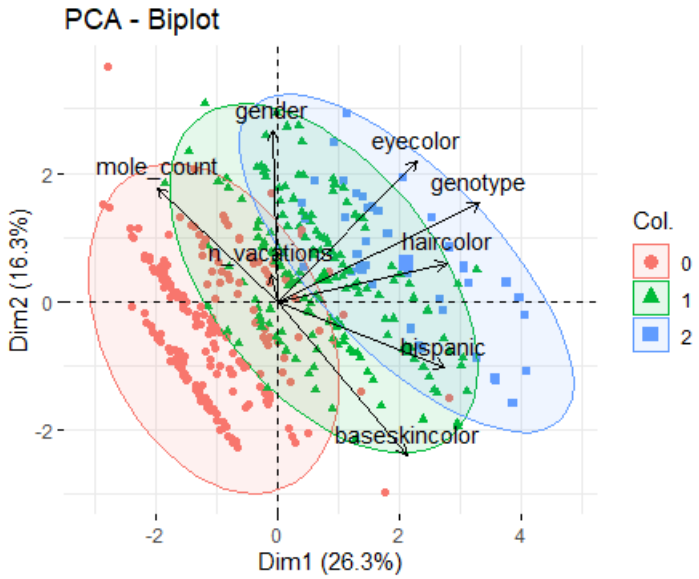

Figure 2: Biplot of PC1 and PC2 from PCA of the moles dataset. The ellipses show the groupings based on genotype with red (0) = genotype gg, green (1) = genotype = ga, and blue (2) = genotype aa. The heterozygous genotype ga overlaps with genotypes gg and aa, but the homozygous genotypes (gg and aa) show distinct groupings.

PCR

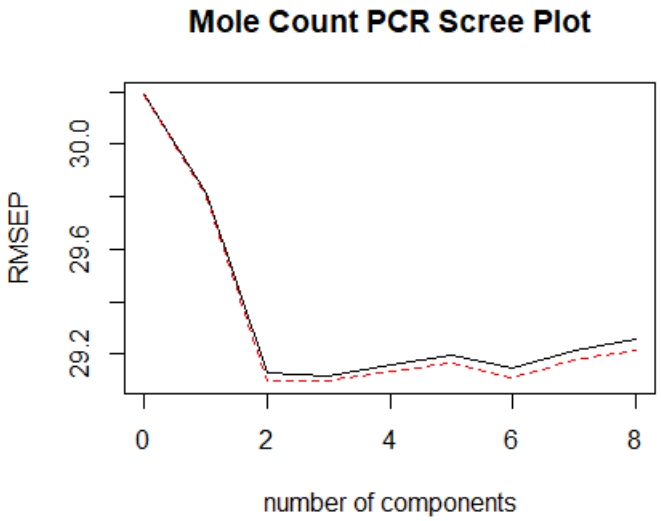

Figure 3: Root mean squared error of prediction (RMSEP) plot of the PCR of the moles dataset with mole count as the response variable. The point with the lowest RMSEP represents the number of components with the highest ability to predict mole count. Lowest RMSEP = 2.

Results [PCR] continued on next page...

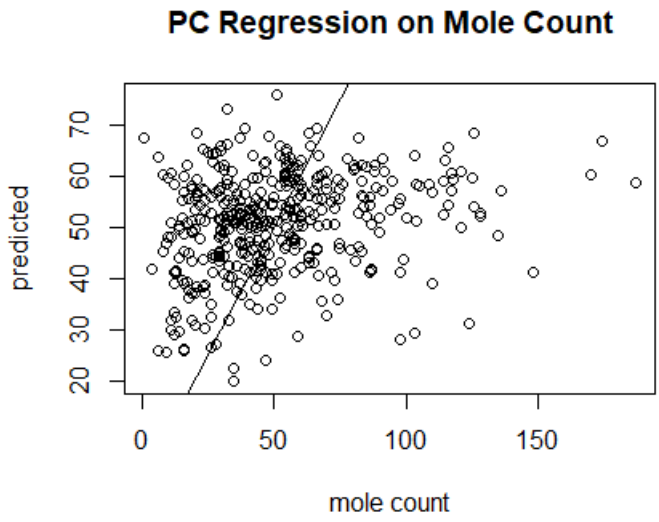

Figure 4: Scatterplot of the observed mole count measurements versus the cross-validated predictions with the regression line from PCR analysis of the moles dataset.

In PCR, the variable genotype can be treated as a factor to see if there is an effect from allele type on mole count. This brings the total number of variables in the model to 8 (the group mole_count is not included as it is the response variable). Figure 3 shows a plot of the root mean squared error of prediction as a function of the number of PCs included. The lowest point, which corresponds to the highest predictive ability of the model, is at two PCs, suggesting that only the first two PCs are necessary to explain most of the variation in number of moles. The scatterplot of measured mole counts versus cross-validated predictions shows the regression line going through densest part of the data, though it does not predict higher mole counts well (Fig. 4). The estimated coefficients for each variable based on the PCR can be seen in Table 2.

PLS

In the PLS model, the variable genotype is also treated as a factor. The root mean squared error of prediction as a function of PCs for the PLS regression shows its lowest point at 5 PCs (Fig. 5). In contrast to the PCR method, PLS suggests that there are three more PCs necessary to have the highest prediction ability. Considering that the PCA step in PCR does not consider the response variable in calculating the covariance matrix, the difference in number of PCs is unsurprising. Figure 6 shows the scatterplot of measured mole counts versus cross-validated predictions for the PLS regression, which looks very similar to that of the PCR analysis. The consistent lack of fit between these models suggests that there are likely factors not included in the model that would better help to describe the variation for higher mole counts.

Results [PLS] continued on next page...

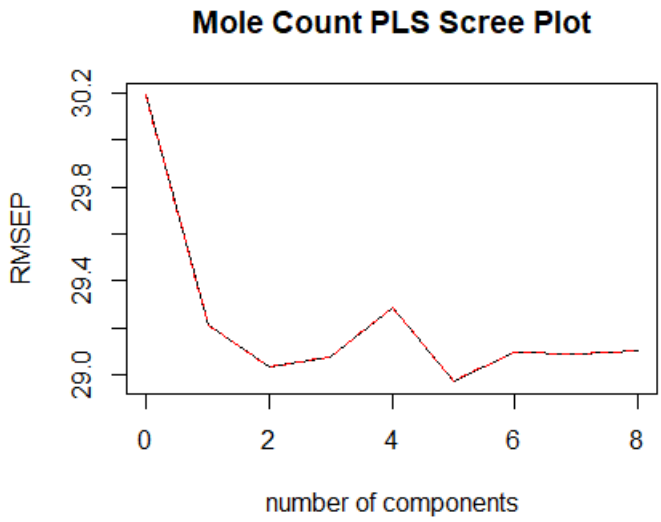

Figure 5: Root mean squared error of prediction (RMSEP) plot of the PLS of the moles dataset with mole count as the response variable. The point with the lowest RMSEP represents the number of components with the highest ability to predict mole count. Lowest RMSEP = 5.

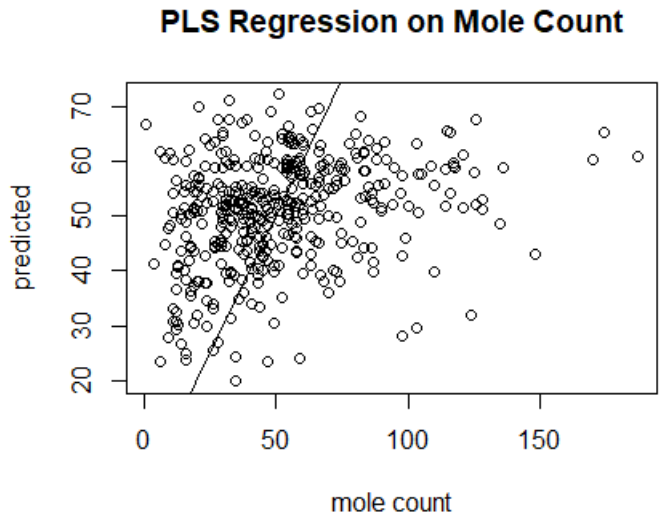

Figure 6: Scatterplot of the observed mole count measurements versus the cross-validated predictions with the regression line from PLS analysis of the moles dataset.

The estimated coefficients for each variable based on the PLS regression can be seen in Table 2. While the estimated magnitude between the two models is quite different, the direction of effect is consistent between models. Given that the PLS incorporates the response variable in its covariance matrix, the estimates from the PLS model are likely more accurate. To confirm this, an ordinary least squares regression was performed using the same response and predictor variables and the estimated coefficients match those of the PLS.

Discussion

Principal component analysis (PCA) is largely known for its role as a data reduction technique. Here, principal components are used in two different methods, principal component regression (PCR) and partial least squares (PLS) regression, to address the problem of multicollinear or rank deficient data. This is particularly useful for high dimensional data sets which will often exhibit multicollinearity between sets of columns. Although the moles dataset used as an example in this study was not high-dimensional, it did have highly correlated variables (eye color, hair color, and OCA2 genotype, and base skin color and Hispanic ethnicity). While biological evidence supported the multicollinearity of these variables, the relationship between the variables was further validated through PCA. As expected, when PCA was conducted on the data 5 groupings of the variables occurred between the first two PCs: (1) Eye color, hair color, and OCA2 genotype, (2) base skin color and Hispanic, (3) Gender, (4) number of seaside vacations, (5) mole count. The variables eye color, hair color, and base skin color could have been excluded and an ordinary lease squares regression performed to assess the contribution of each remaining variable on number of moles a child has. However, there is likely variation in mole type that can be explained by eye color, hair color, and base skin color that are not due to their association with genotype or Hispanic status, respectively. Therefore, a regression that utilizes principal components can be used to address the question of interest while keeping the collinear columns in the analysis.

Summary of Estimated Coefficient

Variable	*PCR*	*PLS*	*OLS*
genotype ga	2.07	4.20	4.20
genotype gg	3.32	6.64	6.64
gender	1.43	2.87	2.87
eye color	-1.12	-1.35	-1.35
hair color	-1.14	-1.14	-1.14
base skin color	-5.75	-3.50	-3.50
hispanic	-3.47	-10.64	-10.64
n vacations	3.66	1.46	1.46

Table 2: Summary of the estimated coefficients for each predictor variable of mole count (excluding genotype gg, which was the intercept term) from each type of regression tested. PCR = principal component regression, PLS = partial least squares, and OLS = ordinary least squares regression.

The first method of using PCs in a linear regression that was tested on the moles dataset was a principal component regression, in which the PCs from a PCA of all the variables except the response variable are used as covariates in the linear model. The use of select PCs from a PCA in a linear regression can be beneficial when there is an unknown structure underlying the data that can act as a confound, such as genetic ancestry in a genome wide association study. However, PCR is limited in its use for estimating coefficients of predictor variables. In PCR, the covariance matrix used to calculate the PCs uses only the distance between the predictor variables and does not consider the association between the predictors and the response. Thus, while the PCs constructed by PCR will be able to explain the variation in the data (as can be seen by a similar fit of the regression lines to the data between PCR and PLS methods), the estimated coefficients will be inaccurate in representing the relationship between the predictors and the response variable.

Discussion continued on next page...

PLS regression is a more appropriate method of using factor scores for estimating the coefficients of predictor variables. Both PLS and PCR are methods that use factor scores determined by linear combinations of the original variables to construct PCs that represent the variability in the data. However, PLS constructs a variance-covariance matrix using information on the distance between the predictors and between the predictors and the response. The incorporation of the response variable in the construction of PCs in PLS makes it better suited for estimating the contribution of each predictor variable in explaining the variation of the response variable. This is most easily seen in the RSMEP plots (Fig 3 & 5) where PCR suggests only 2 principal components are necessary to describe the majority of the variation in the data, while PLS suggests that 5 PCs are necessary. Furthermore, the estimated coefficients from PLS regression have considerably different magnitudes compared with the estimated coefficients from PCR. In contrast with coefficients estimated using an ordinary least squares regression, PLS is more accurate in representing the relationship between the predictors and the response (the estimated coefficients were the same), as was anticipated.

Conclusion

Principal components have a variety of applications in the analysis of high-dimensional data. While principal component analysis is most well-known for its data-reduction capabilities, it is also useful as a method for addressing datasets with multicollinearity. In the construction of PCs, linear combinations of the data are used to create factor scores, which ultimately result in orthogonal vectors (PCs) that represent the amount of variation explained by each variable. This means that even if the original data contains variables that are collinear, the PCs will be independent of each other.

Once PCs have been constructed from the data, they can be used in linear regression models to help account for underlying structure in the data that would otherwise confound the relationship between the main predictor variable and the response, or to estimate the coefficients of each variable. If PCs are being used to account for unknown structure in the data, a linear regression using the first several PCs as covariates is appropriate. However, principal component regression is not appropriate for estimating coefficients of each predictor variable as the response variable is not incorporated in the calculation of the covariance matrix, and thus the relationship between the predictor and the response variable is not represented. Instead PLS, which is similar to PCR, is an appropriate method for estimating coefficients as the relationship between the response and predictors is incorporated in the creation of the PCs. In summary, PCA is a powerful, versatile tool for handling high-dimensional and collinear data, though depending on the purpose of PCs in downstream analysis, careful consideration should be used when determining how to construct them.

References

[1] Heidi Artigue and Gary Smith. The principal problem with principal components regression. *Cogent Mathematics Statistics*, (just-accepted):1622190, 2019.

[2] Jürgen Bauer, Petra Büttner, Tine Sander Wiecker, Heike Luther, and Claus Garbe. Risk factors of incident melanocytic nevi: a longitudinal study in a cohort of 1,232 young german children. *International journal of cancer*, 115(1):121–126, 2005.

[3] David L Duffy, Gu Zhu, Xin Li, Marianna Sanna, Mark M Iles, Leonie C Jacobs, David M Evans, Seyhan Yazar, Jonathan Beesley, and Matthew H Law. Novel pleiotropic risk loci for melanoma and nevus density implicate multiple biological pathways. *Nature communications*, 9(1):4774, 2018.

[4] Michael Haenlein and Andreas M. Kaplan. A beginner's guide to partial least squares analysis. *Understanding Statistics*, 3(4):283–297, 2004.

[5] Inge S Helland. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, pages 97–114, 1990.

[6] Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.

[7] RX Liu, Jian Kuang, Qi Gong, and XL Hou. Principal component regression analysis with spss. *Computer methods and programs in biomedicine*, 71(2):141–147, 2003.

[8] Jonas Mengel-From, Terence H Wong, Niels Morling, Jonathan L Rees, and Ian J Jackson. Genetic determinants of hair and eye colours in the scottish and danish populations. *BMC genetics*, 10(1):88, 2009.

[9] Catherine M Olsen, Heidi J Carroll, and David C Whiteman. Estimating the attributable fraction for cancer: a meta-analysis of nevi and melanoma. *Cancer prevention research*, 3(2):233–245, 2010.

[10] Gina M Peloso and Kathryn L Lunetta. Choice of population structure informative principal components for adjustment in a case-control study. *BMC genetics*, 12(1):64, 2011.

[11] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.

[12] Sri N Shekar, David L Duffy, Tony Frudakis, Richard A Sturm, Zhen Z Zhao, Grant W Montgomery, and Nicholas G Martin. Linkage and association analysis of spectrophotometrically quantified hair color in australian adolescents: the effect of oca2 and herc2. *Journal of Investigative Dermatology*, 128(12):2807–2814, 2008.

[13] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[14] Philippa Youl, Joanne Aitken, Nicholas Hayward, David Hogg, Ling Liu, Norman Lassam, Nicholas Martin, and Adele Green. Melanoma in adolescents: a case-control study of risk factors in queensland, australia. *International journal of cancer*, 98(1):92–98, 2002.