The Number of Linearly Independent Vectors in Spectral Databases

Carlos Sáenz, Begoña Hernández, Coro Alberdi, Santiago Alfonso,
and José Manuel Diñeiro

Departamento de Física, Universidad Pública de Navarra,
Campus Arrosadia, 31006 Pamplona, Spain

Abstract. Linear dependence among spectra in spectral databases affects the eigenvectors obtained from principal component analysis. This affects the values of usual spectral and colorimetric metrics. The effective dependence is proposed as a tool to quantify the maximum number of linearly independent vectors in the database. The results of the proposed algorithm do not depend on the selection of the first seed vector and are consistent with the results based on reduction of the bivariate coefficient of determination.

Keywords: Spectral databases, effective dependence, linear correlation, collinearity.

1 Introduction

Spectral databases are used in many applications within the context of spectral colour science. Dimensionality reduction techniques like principal component analysis (PCA), incomplete component analysis (ICA) and others are used to describe spectral information with a reduced number of basis functions. Applications of these techniques are found in many fields and require a detailed evaluation of their performance. Testing the performance of these methods usually involve spectral databases from two complementary but different points of view. The set of basis functions or vectors are obtained from a particular spectral database, called the Training set, using some specific spectral or colorimetric metrics. Then the performance of the basis functions in order to reconstruct spectral or colorimetric information is checked with the help of a second spectral database, the Test set. Numerical results depend on the used databases [1] and metrics, in this scenario some authors recommend the simultaneous use of several metrics to evaluate the quality of the data reconstruction [2,3].

Spectral databases may differ because of the measurement technique, wavelength limits, wavelength interval or number of data points in their spectra. Even more important differences are found because of the origin of the samples used to construct the database. Some databases have been obtained from color atlases or color collections, others correspond to measurements of natural objects or to samples specifically created with some purpose. Recently the principal characteristics of some frequently used spectral databases have been reviewed [4].

Some of the most frequently used spectral databases, like Munsell or NCS, have been measured in collections of color samples. These color collections have been constructed according to some specific colorimetric or perceptual criteria, say uniformly distributed samples in the color space. No spectral criteria were used in their construction. In fact, we do not actually possess a criterion that allows us to talk for instance about uniformly distributed spectra.

In this work we will analyze the possibility of using the linear dependence between spectra as a measure of the amount of spectral information contained in the database. A parameter of this kind, independent of particular choices of spectral or colorimetric measures, could be a valuable indicator of the 'spectral diversity' within the database.

2 Spectral Databases and Linear Dependence

2.1 Effect of Linear Dependence in $RMSE$ and ΔE^*

Let us suppose that we have a spectral database formed by q spectra $\mathbf{r}_i$, $i = 1, 2, \dots, q$, representing the reflectance factor $\mathbf{r}$ of q samples measured at n wavelengths. In general any spectrum $\mathbf{r}_j$ can be obtained from the other spectra $\mathbf{r}_i$, $i \neq j$ in the database as:

$$\mathbf{r}_j = \sum_{\substack{i=1 \\ i \neq j}}^q w_i \mathbf{r}_i + \mathbf{e}_j = \hat{\mathbf{r}}_j + \mathbf{e}_j \quad (1)$$

Where w_i are the appropriate weights. In (1) the vector $\hat{\mathbf{r}}_j$ is the estimated value of $\mathbf{r}_j$ that can be obtained from the remaining vectors in the database and $\mathbf{e}_j = \mathbf{r}_j - \hat{\mathbf{r}}_j$ is an error term. Respect to the spectral information in $\mathbf{r}_j$ the error term $\mathbf{e}_j$ represents the intrinsic information contained in $\mathbf{r}_j$ that can not be reproduced by the rest of spectra. In general, an accepted measure of the spectral similarity/difference is the $RMSE_j$ value between the original and estimated vectors defined as

$$RMSE_j = \sqrt{\frac{1}{n} \sum_{k=1}^n (\mathbf{r}_{kj} - \hat{\mathbf{r}}_{kj})^2} = \sqrt{\frac{1}{n} \sum_{k=1}^n \mathbf{e}_{kj}^2} \quad (2)$$

Where the index k identifies each of the n measured wavelengths. If we are interested in colorimetric information, the tristimulus values must be also computed. For a given illuminant $\mathbf{S}$, the X_i tristimulus value of $\mathbf{r}_j$ is:

$$X_j = K \sum_{k=1}^n r_{kj} S_k \bar{x}_k \quad (3)$$

Where K is a normalization factor, r_{kj} the reflectance factor at wavelength k and $\bar{x}$ the color matching function. The tristimulus Y_j and Z_j are obtained using the color matching functions $\bar{y}$ and $\bar{z}$ respectively. Substituting (1) in (3)

the tristimulus value X_j can be obtained as a function of the tristimulus values of the other spectra in the database as:

$$\begin{aligned}
 X_j &= K \sum_{k=1}^n \hat{r}_{kj} S_k \bar{x}_k + K \sum_{k=1}^n e_{kj} S_k \bar{x}_k = \\
 &= \sum_{\substack{i=1 \\ i \neq j}}^q w_i X_i + K \sum_{k=1}^n e_{kj} S_k \bar{x}_k = \\
 &= \hat{X}_j + X e_j
 \end{aligned} \tag{4}$$

Which is an obvious consequence of the linearity of (3). In this expression $\hat{X}_j$ is the estimation of X_j that we can obtain solely with the vectors in the database and $X e_j$ is the tristimulus value associated to the error term $\mathbf{e}_j$. Therefore the tristimulus values of $\mathbf{r}_j$ are a linear combination of the tristimulus values or the other spectra in the database plus an extra term that depends on $\mathbf{e}_j$.

If the error term $\mathbf{e}_j$ is sufficiently small, in the sense that all e_{kj} are small, then $RMSE_j$ will be also small. Furthermore, $X e_j$ will be also small and $X_j \approx \hat{X}_j$. The same argument can be extended to the other tristimulus values. If true and estimated tristimulus values are very similar, then ΔE^* color differences between true and estimated spectra are expected to be small too.

All these arguments are well known and linear models are extensively used in spectral and color reproduction and estimation; where in general spectra are reconstructed using a limited number of basis vectors. An interesting and ever present problem is that there is no evident relationship between the spectral reconstruction accuracy, measured with $RMSE$ or other spectral metric, and the color reproduction accuracy determined with a particular color difference measure. This means that we do not have a clear criterion to quantify what does mean a sufficiently small error term $\mathbf{e}_j$. Furthermore colorimetric results are sensible to the illuminant $\mathbf{S}$ used in the calculations.

When the error term $\mathbf{e}_j$ vanishes in (1) then $\mathbf{r}_j$ is an exact linear combination of other spectra in the database. In this case we have $RMSE_j = 0$ and identical tristimulus values in (4) and therefore color differences between the original and reconstructed spectra vanish. It could be said that in this situation $\mathbf{r}_j$ does not provide additional spectral or color information respect to the remaining vectors in the database.

In general the number of spectra q in the database is higher than the number of sampled wavelengths n . If $\mathbf{X}$ is the $n \times q$ matrix where each column is a spectrum of the database, then upper limit to the number of linearly independent vectors in $\mathbf{X}$ is $rank(\mathbf{X}) = \min(n, q) = n$ assuming that $q > n$.

PCA is affected by collinearity [5] and the effect on the basis vectors can be noticeable. Since only few basis vectors are usually retained, the spectral and colorimetric reconstruction accuracy will be also affected. In order to show this effect we have performed the following experiment that resembles the standard Training and Test databases approach. We have used the Munsell colors measured by the Joensuu Color Group [6]. The Munsell dataset consists in 1269 reflectance factor spectra measured with a Perkin-Elmer lambda 9 UV/VIS/NIR

spectrophotometer at 1 nm intervals between 380 and 800 nm. We have randomly split the Munsell database in two, a Training database $\mathbf{A}$ with q_A spectra and a Test database $\mathbf{B}$ with q_B spectra. Then we have randomly selected a single vector from $\mathbf{A}$ to serve as seed vector in order to generate linearly dependent vectors. We have iteratively added to $\mathbf{A}$ vectors proportional to the seed vector, thus increasing q_A in each iteration. The proportionality constant has been uniformly sampled in the range $[0,1]$. After every iteration we have used PCA to obtain the first n_b eigenvectors. Using these eigenvectors we have obtained $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$, the estimations of $\mathbf{A}$ and $\mathbf{B}$. The process has been repeated for different random partitions of the Munsell database.

The effect that the addition of such vectors has on the first two principal components can be seen in Fig. 1 for an example with $q_A = 10$. The seed spectrum (reduced by a factor two) has been also included for comparison. Due to the randomness of the multiplicative constant, eigenvectors do not always evolve in the same direction, being these changes rather unpredictable even though we are modifying the database in the simplest way. A similar situation is found for the other eigenvectors. We can also see the effect on the $RMSE$ and ΔE^* between original and reconstructed data sets in Fig. 2. As the number of linearly dependent vectors in the Training set $\mathbf{A}$ increases, eigenvectors evolve to explain the resulting changes in the correlation matrix. This produces the reduction in the mean $RMSE$ and ΔE^* values between $\mathbf{A}$ and $\hat{\mathbf{A}}$. On the contrary, maximum $RMSE$ and ΔE^* differences increase slightly because the reconstruction accuracy of the original vectors in the database deteriorates accordingly. Respect to the Test database, the new vectors added to $\mathbf{A}$ do not improve either the mean nor the maximum $RMSE$ and ΔE^* values between $\mathbf{B}$ and $\hat{\mathbf{B}}$ and these parameters are roughly constant during all the process. In the presence of linear correlation, the minimization of $RMSE$ or ΔE^* in the Training database does not guarantee optimal results in the Test database. Similar conclusions are obtained if we repeat the process for different initial $\mathbf{A}$ and $\mathbf{B}$ sets although details may differ, sometimes substantially.

2.2 Effective Dependence

In the previous examples the collinearity within the data is *a priori* known, by construction. In a real situation collinearity will be distributed over the entire sample set in an unknown manner. Therefore it is interesting to possess a measure of the amount of collinearity or linear dependence between variables for the entire spectral set. Although bivariate correlation is accurately defined through the Pearson correlation coefficient we do not have a single, widely accepted measure of linear dependence in the case of multivariate data.

In a recent paper Peña and Rodríguez [7] have proposed two new descriptive measures for multivariate data: the effective variance and the effective dependence. Their main objective was to define a dependence measure that could be used to compare data sets with different number of variables. In particular, if $\mathbf{X}$ is the $n \times p$ matrix having p variables and n observations of each variable, then the effective dependence $D_e(\mathbf{X})$ is defined as:

Fig. 1. Changes in the first (top) and second (bottom) eigenvectors after de addition of 1,2,10,20,30 and 40 vectors proportional to a single seed vector belonging to the original set. The seed vector (dark line) has been reduced by a factor 2.

$$D_e(\mathbf{X}) = 1 - |\mathbf{R}_\mathbf{X}|^{1/p} \quad (5)$$

Where $|\mathbf{R}_\mathbf{X}|$ is the determinant of the correlation matrix $\mathbf{R}_\mathbf{X}$ of $\mathbf{X}$. Authors demonstrate that $D_e(\mathbf{X})$ satisfies the main properties of a dependence measure and of particular interest in our discussion:

- a) $0 \leq D_e(\mathbf{X}) \leq 1$, and $D_e(\mathbf{X}) = 1$ if and only if we can find a vector $\mathbf{a} \neq 0$ and $\mathbf{b}$ such $\mathbf{a}'\mathbf{X} + \mathbf{b} = 0$. This means that $D_e(\mathbf{X}) = 1$ implies that there exists

Fig. 2. *RMSE* (top) and ΔE^* (bottom) as a function of the number linearly dependent vectors added to the training database. Solid lines are mean values and dot dashed lines maximum values. Letters A and B refer to the training and test databases respectively. All parameters have been normalized to the first value.

collinearity within the data. Also $D_e(\mathbf{X}) = 0$ if and only if the covariance matrix of $\mathbf{X}$ is diagonal.

- b) Let $\mathbf{Z}' = [\mathbf{X}'\mathbf{Y}']$ be a random vector of dimension $p + q$ where $\mathbf{X}$ and $\mathbf{Y}$ are random variables of dimension p and q respectively, then $D_e(\mathbf{Z}) \geq D_e(\mathbf{X})$ if and only if $D_e(\mathbf{Y} : \mathbf{X}) > D_e(\mathbf{X})$ where $D_e(\mathbf{Y} : \mathbf{X})$ is the additional correlation introduced by $\mathbf{Y}$. Analogously $D_e(\mathbf{Z}) \leq D_e(\mathbf{X})$ if and only if $D_e(\mathbf{Y} : \mathbf{X}) < D_e(\mathbf{X})$.

Fig. 3. The value of R^2 of the spectrum removed from the training database (solid line) and of the $D_e(\mathbf{X})$ (dot dashed line) as a function of the number of the remaining spectra q . The arrow marks the point where $D_e(\mathbf{X})$ starts to decrease.

We now propose to use the effective dependence to find the number of linearly independent vectors in the database. We have investigated two different approaches that we will analyze independently.

2.3 Backward Method: Reduction of Bivariate Correlation

In this method we start with the entire spectral database and calculate the pair wise coefficients of determination R_{ij}^2 with $i \neq j$ between all possible pairs of spectra within the database. Then the spectrum having the maximum R_{ij}^2 is removed and the process repeated for the remaining spectra.

Fig. 3 shows the $\max(R_{ij}^2)$ value of the removed spectrum during the entire process for a random subset of 400 spectra from the Munsell database. The value of $D_e(\mathbf{X})$ after each iteration is also shown. The reduction process starts at the rightmost value in the figure ($n = 400$) and continues to the left. We can observe that in this example the first occurrence of $D_e(\mathbf{X}) < 0$ happens where the number of remaining spectra in the database is $q_1 = 120$ and $\max(R_{ij}^2) = 0.9349$.

Further reduction in q also implies a reduction in the value of the effective dependence. Notice that the effective dependence decreases monotonically with the number of remaining spectra in the database q .

Since R_{ij}^2 is a bivariate statistic we can not assume that this procedure is the most effective to reduce global collinearity within the database. Therefore the value q_1 must be regarded as a lower limit to the number of linearly independent vectors in the original spectral database.

Fig. 4. The effective dependence as a function of the spectra in the data base

2.4 Forward Method: $D_e(\mathbf{X})$ Minimization

The second approach is based in the properties of the effective dependence and consists in finding the subset of spectra of the original database that minimizes $D_e(\mathbf{X})$ and maximizes the number of spectra. The algorithm begins with a single spectrum, the seed spectrum. Then the value of $D_e(\mathbf{X})$ resulting after the addition of a second spectrum is computed for all remaining spectra in the database. The spectrum providing the minimum increment to $D_e(\mathbf{X})$ is retained increasing the number of spectra in one. The process is repeated, adding new vectors, until $D_e(\mathbf{X}) = 1$ is obtained. Let it be q_2 the number of spectra in the optimized set immediately before $D_e(\mathbf{X}) = 1$.

In order to apply this method, we must select an initial spectrum, the seed spectrum, from the data set. Lacking of a good reason to choose a particular one we have repeated the process using all vectors as seed vectors. In principle this would led to different solutions, having different number of spectra q_2 . The solution or solutions having maximum q_2 inform us about the maximum number of independent vectors in the original dataset.

We have performed the experiment over the same subset of the preceding section, with 400 vectors. In Fig. 4 we show the evolution of the effective dependence during the construction of the 'optimized' sets. The 400 curves corresponding to the 400 possible seed vectors have been plotted. It can be seen that the rate of change in the effective dependence depends only slightly on the seed vector and $D_e(\mathbf{X})$ values rapidly converge in all cases, giving very similar number of vectors q_2 in the optimized set. In particular, for this dataset, we have obtained $q_2 = 133$ vectors in 338 cases and $q_2 = 134$ vectors in 62 cases. This suggests that the choice of the initial seed vector is of little relevance. This fact is of practical importance since the forward algorithm is time consuming. Therefore, for large

databases the algorithm could be used for a small random subset of seed spectra. We have also tested the possibility that a random set having $q = q_2$ spectra could exhibit less collinearity ($D_e(\mathbf{X}) < 1$) than the 'optimized' set. We have created 5000 random sets with $q=133$ vectors taken from the original dataset and in all cases the value $D_e(\mathbf{X}) = 1$ was obtained.

As expected, q_2 is greater than q_1 and both much larger than the usual number of basis vectors that are retained in practical applications. In fact the 'optimized' data sets are optimized solely in terms of the effective dependence measure. This does not necessarily mean that they provide a better starting point to apply standard dimensionality reduction techniques.

3 Conclusions

Most spectral databases are affected by collinearity. This produces a bias in the basis vectors obtained from statistical methods like principal component analysis. This bias need not to be a drawback, since it accounts for the distributional properties of the original data, which may be necessary for the particular application. However collinearity may affect the results when different spectral databases, with different origin, are compared.

The effective dependence provides a measure of the degree of collinearity within a spectral database. The maximum number of spectra that can be retained before the effective dependence becomes unity inform us about the quantity of independent information contained. The properties of the effective dependence allow a forward construction algorithm that gives solution having a number of vectors that are almost independent on the seed vector used to start the process. The results obtained are in agreement with the simpler and more intuitive backward algorithm based in the removal of those spectra having high bivariate correlations.

Several practical aspects need further investigation: the properties of the optimized sets with regard to the spectral and colorimetric reconstruction, the relationship between the effective dependence and the number of sampled wavelengths or how to use the 'effective number of spectra' to compare different spectral data sets.

References

1. Sáenz, C., Hernández, B., Alberdi, C., Alfonso, S., Diñero, J.M.: The effect of selecting different training sets in the spectral and colorimetric reconstruction accuracy. In: Ninth International Symposium on Multispectral Colour Science and Application, MCS 2007, Taipei, Taiwan (2007)
2. Imai, F.H., Rosen, M.R., Berns, R.S.: Comparative study of metrics for spectral match quality. In: Cgiv 2002: First European Conference on Colour in Graphics, Imaging, and Vision, Conference Proceedings, pp. 492–496 (2002)
3. Viggiano, J.S.: Metrics for evaluating spectral matches: A quantitative comparison. In: Cgiv 2004: Second European Conference on Color in Graphics, Imaging, and Vision - Conference Proceedings, pp. 286–291 (2004)

4. Kohonen, O., Parkkinen, J., Jaaskelainen, T.: Databases for spectral color science. *Color Research and Application* 31(5), 381–390 (2006)
5. Jolliffe, I.T.: *Principal component analysis*, 2nd edn. Springer series in statistics. Springer, New York (2002)
6. Spectral Database, University of Joensuu Color Group,
<http://spectral.joensuu.fi>
7. Peña, D., Rodríguez, J.: Descriptive measures of multivariate scatter and linear dependence. *Journal of Multivariate Analysis* 85(2), 361–374 (2003)