

# Kernel Based Subspace Projection of Near Infrared Hyperspectral Images of Maize Kernels

Rasmus Larsen<sup>1</sup>, Morten Arngren<sup>1,2</sup>, Per Waaben Hansen<sup>2</sup>,  
and Allan Aasbjerg Nielsen<sup>3</sup>

<sup>1</sup> DTU Informatics, Technical University of Denmark  
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark  
{r1,ma}@imm.dtu.dk

<sup>2</sup> FOSS Analytical AS, Slangerupgade 69, DK-3400 Hillerød, Denmark  
pwh@foss.dk

<sup>3</sup> DTU Space, Technical University of Denmark  
Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, Denmark  
aa@space.dtu.dk

**Abstract.** In this paper we present an exploratory analysis of hyperspectral 900-1700 nm images of maize kernels. The imaging device is a line scanning hyper spectral camera using a broadband NIR illumination. In order to explore the hyperspectral data we compare a series of subspace projection methods including principal component analysis and maximum autocorrelation factor analysis. The latter utilizes the fact that interesting phenomena in images exhibit spatial autocorrelation. However, linear projections often fail to grasp the underlying variability on the data. Therefore we propose to use so-called kernel version of the two afore-mentioned methods. The kernel methods implicitly transform the data to a higher dimensional space using non-linear transformations while retaining the computational complexity. Analysis on our data example illustrates that the proposed kernel maximum autocorrelation factor transform outperform the linear methods as well as kernel principal components in producing interesting projections of the data.

## 1 Introduction

Based on work by Pearson [1] in 1901, Hotelling [2] in 1933 introduced principal component analysis (PCA). PCA is often used for linear orthogonalization or compression by dimensionality reduction of correlated multivariate data, see Jolliffe [3] for a comprehensive description of PCA and related techniques.

An interesting dilemma in reduction of dimensionality of data is the desire to obtain simplicity for better understanding, visualization and interpretation of the data on the one hand, and the desire to retain sufficient detail for adequate representation on the other hand.

Schölkopf et al. [4] introduce kernel PCA. Shawe-Taylor and Cristianini [5] is an excellent reference for kernel methods in general. Bishop [6] and Press et al. [7] describe kernel methods among many other subjects.

The kernel version of PCA handles nonlinearities by implicitly transforming data into high (even infinite) dimensional feature space via the kernel function and then performing a linear analysis in that space.

The maximum autocorrelation factor (MAF) transform proposed by Switzer [11] defines maximum spatial autocorrelation as the optimality criterion for extracting linear combinations of multispectral images. Contrary to this PCA seeks linear combinations that exhibit maximum variance. Because the interesting phenomena in image data often exhibit some sort of spatial coherence spatial autocorrelation is often a better optimality criterion than variance. A kernel version of the MAF transform has been proposed by Nielsen [10].

In this paper we shall apply kernel MAF as well as kernel PCA and ordinary PCA and MAF to find interesting projections of hyperspectral images of maize kernels.

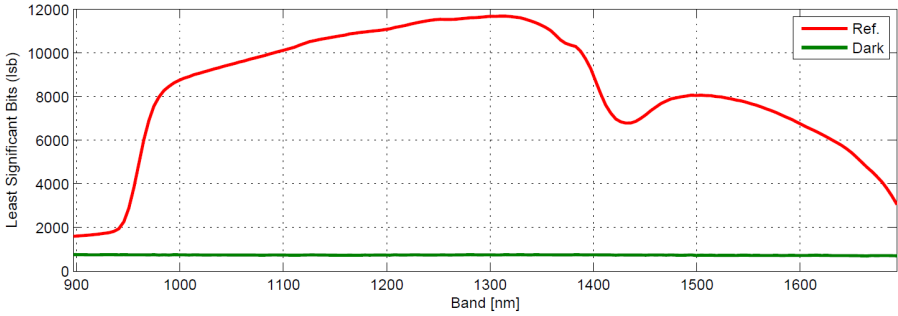
## 2 Data Acquisition

A hyperspectral line-scan NIR camera from Headwall Photonics sensitive from 900-1700nm was used to capture the hyperspectral image data. A dedicated NIR light source illuminates the sample uniformly along the scan line and an advanced optic system developed by Headwall Photonics disperses the NIR light onto the camera sensor for acquisition. A sledge from MICOS GmbH moves the sample past the view slot of the camera allowing it to acquire a hyperspectral image. In order to separate the different wavelengths an optical system based on the Offner principle is used. It consists of a set of mirrors and gratings to guide and spread the incoming light into a range of wavelengths, which are projected onto the InGaAs sensor.

The sensor has a resolution of 320 spatial pixels and 256 spectral pixels, i.e. a physical resolution of  $320 \times 256$  pixels. Due to the Offner dispersion principle (the convex grating) not all the light is in focus over the entire dispersed range. This means that if the light were dispersed over the whole 256 pixel wide sensor the wavelengths at the periphery would be out of focus. In order to avoid this the light is only projected onto 165 pixels instead and the top 91 pixels are disregarded. This choice is a trade-off between spatial sampling resolution and focus quality of the image.

The camera acquires 320 pixels and 165 bands for each frame. The pixels are represented in 14 bit resolution with 10 effective bits In Fig. 1 average spectra for a white reference and dark background current images are shown. Note the limited response in the 900-950 nm range.

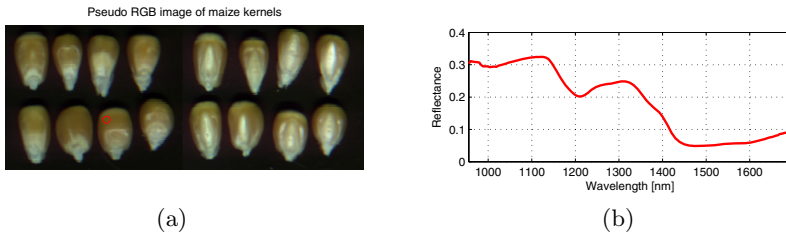
Before the image cube is subjected to the actual processing a few pre-processing steps are conducted. Initially the image is corrected for the reference light and dark background current. A reference and dark current image are acquired and the mean frame is applied for the correction. In our case the hyperspectral data are kept as reflectance spectra throughout the analysis.



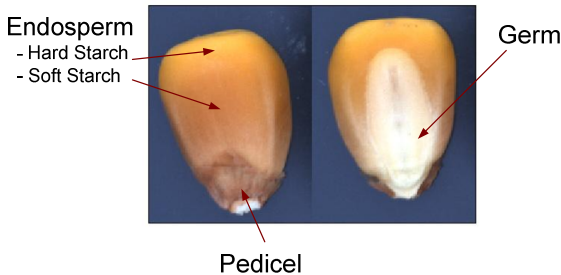
**Fig. 1.** Average spectra for white reference and dark background current images

### 2.1 Grain Samples Dataset

For the quantitative evaluation of the kernel MAF method a hyperspectral image of eight maize kernels is used as the dataset. The hyperspectral image of the maize samples are comprised of the front and back-side of the kernels on a black background (NCS-9000) appended as two separate cropped images as depicted in Fig. 2(a). In Fig. 2(b) an example spectrum is shown. The kernels are not



**Fig. 2.** (a) Front (left) and back (right) images of eight maize kernels on a dark background. The color image is constructed as an RGB combination of NIR bands 150, 75, and 1; (b) reflectance spectrum of the pixel marked with red circle in (a).



**Fig. 3.** Maize kernel constituents front- and backside (pseudo RGB)

fresh from harvest and hence have a very low water content and are in addition free from any infections. Many cereals in general share the same compounds and basic structure. In our case of maize a single kernel can be divided into many different constituents on the macroscopic level as illustrated in Fig. 3.

In general, the structural components of cereals can be divided into three classes denoted *Endosperm*, *Germ* and *Pedicle*. These components have different functions and compounds leading to different spectral profiles as described below.

*Endosperm.* The endosperm is the main storage for starch ( $\sim 66\%$ ), protein ( $\sim 11\%$ ) and water ( $\sim 14\%$ ) in cereals. Starch being the main constituent is a carbohydrate and consists of two different glucans named *Amylose* and *Amylopectin*. The main part of the protein in the endosperm consists of *zein* and *glutenin*. The starch in maize grains can be further divided into a soft and a hard section depending on the binding with the protein matrix. These two types of starch are typically mutually exclusive, but in maize grain they both appear as a special case as also illustrated in figure 3.

*Germ.* The germ of a cereal is the reproductive part that germinates to grow into a plant. It is the embryo of the seed, where the scutellum serves to absorb nutrients from the endosperm during germination. It is a section holding proteins, sugars, lipids, vitamins and minerals [13].

*Pedicle.* The pedicle is the flower stalk and has negligible interest in terms of production use. For a more detailed description of the general structure of cereals [12].

### 3 Principal Component Analysis

Let us consider an image with  $n$  observations or pixels and  $p$  spectral bands organized as a matrix  $X$  with  $n$  rows and  $p$  columns; each column contains measurements over all pixels from one spectral band and each row consists of a vector of measurements  $x_i^T$  from  $p$  spectral bands for a particular observation  $X = [x_1^T x_2^T \dots x_n^T]^T$ . Without loss of generality we assume that the spectral bands in the columns of  $X$  have mean value zero.

#### 3.1 Primal Formulation

In ordinary (primal also known as R-mode) PCA we analyze the sample variance-covariance matrix  $S = X^T X / (n - 1) = 1 / (n - 1) \sum_{i=1}^n x_i x_i^T$  which is  $p$  by  $p$ . If  $X^T X$  is full rank  $r = \min(n, p)$  this will lead to  $r$  non-zero eigenvalues  $\lambda_i$  and  $r$  orthogonal or mutually conjugate unit length eigenvectors  $u_i$  ( $u_i^T u_i = 1$ ) from the eigenvalue problem

$$\frac{1}{n-1} X^T X u_i = \lambda_i u_i. \quad (1)$$

We see that the sign of  $u_i$  is arbitrary. To find the principal component scores for an observation  $x$  we project  $x$  onto the eigenvectors,  $x^T u_i$ . The variance of these

scores is  $u_i^T S u_i = \lambda_i u_i^T u_i = \lambda_i$  which is maximized by solving the eigenvalue problem.

### 3.2 Dual Formulation

In the dual formulation (also known as Q-mode analysis) we analyze  $XX^T/(n-1)$  which is  $n$  by  $n$  and which in image applications can be very large. Multiply both sides of Equation 1 from the left with  $X$

$$\frac{1}{n-1}XX^T(Xu_i) = \lambda_i(Xu_i) \quad \text{or} \quad \frac{1}{n-1}XX^T v_i = \lambda_i v_i \tag{2}$$

with  $v_i$  proportional to  $Xu_i$ ,  $v_i \propto Xu_i$ , which is normally not normed to unit length if  $u_i$  is. Now multiply both sides of Equation 2 from the left with  $X^T$

$$\frac{1}{n-1}X^T X(X^T v_i) = \lambda_i(X^T v_i) \tag{3}$$

to show that  $u_i \propto X^T v_i$  is an eigenvector of  $S$  with eigenvalue  $\lambda_i$ . We scale these eigenvectors to unit length assuming that  $v_i$  are unit vectors  $u_i = X^T v_i / \sqrt{(n-1)\lambda_i}$ .

We see that if  $X^T X$  is full rank  $r = \min(n, p)$ ,  $X^T X/(n-1)$  and  $XX^T/(n-1)$  have the same  $r$  non-zero eigenvalues  $\lambda_i$  and that their eigenvectors are related by  $u_i = X^T v_i / \sqrt{(n-1)\lambda_i}$  and  $v_i = Xu_i / \sqrt{(n-1)\lambda_i}$ . This result is closely related to the Eckart-Young [8,9] theorem.

An obvious advantage of the dual formulation is the case where  $n < p$ . Another advantage even for  $n \gg p$  is due to the fact that the elements of the matrix  $G = XX^T$ , which is known as the Gram<sup>1</sup> matrix, consist of inner products of the multivariate observations in the rows of  $X$ ,  $x_i^T x_j$ .

### 3.3 Kernel Formulation

We now replace  $x$  by  $\phi(x)$  which maps  $x$  nonlinearly into a typically higher dimensional feature space. The mapping by  $\phi$  takes  $X$  into  $\Phi$  which is an  $n$  by  $q$  ( $q \geq p$ ) matrix, i.e.  $\Phi = [\phi(x_1)^T \phi(x_2)^T \dots \phi(x_n)^T]^T$  we assume that the mappings in the columns of  $\Phi$  have zero mean. In this higher dimensional feature space  $C = \Phi^T \Phi / (n-1) = 1/(n-1) \sum_{i=1}^n \phi(x_i) \phi(x_i)^T$  is the variance-covariance matrix and for PCA we get the primal formulation  $1/(n-1)\Phi^T \Phi u_i = \lambda_i u_i$  where we have re-used the symbols  $\lambda_i$  and  $u_i$  from above. For the corresponding dual formulation we get re-using the symbol  $v_i$  from above

$$\frac{1}{n-1}\Phi\Phi^T v_i = \lambda_i v_i. \tag{4}$$

As above the non-zero eigenvalues for the primal and the dual formulations are the same and the eigenvectors are related by  $u_i = 1/(\sqrt{(n-1)\lambda_i}) \Phi^T v_i$ , and  $v_i = \Phi u_i / \sqrt{(n-1)\lambda_i}$ . Here  $\Phi\Phi^T$  plays the same role as the Gram matrix above and has the same size, namely  $n$  by  $n$  (so introducing the nonlinear mappings in  $\phi$  does not make the eigenvalue problem in Equation 4 bigger).

<sup>1</sup> Named after Danish mathematician Jørgen Pedersen Gram (1850-1916).

**Kernel Substitution.** Applying kernel substitution also known as the kernel trick we replace the inner products  $\phi(x_i)^T \phi(x_j)$  in  $\Phi\Phi^T$  with a kernel function  $\kappa(x_i, x_j) = \kappa_{ij}$  which could have come from some unspecified mapping  $\phi$ . In this way we avoid the explicit mapping  $\phi$  of the original variables. We obtain

$$Kv_i = (n - 1)\lambda_i v_i \quad (5)$$

where  $K = \Phi\Phi^T$  is an  $n$  by  $n$  matrix with elements  $\kappa(x_i, x_j)$ . To be a valid kernel  $K$  must be symmetric and positive semi-definite, i.e., its eigenvalues are non-negative. Normally the eigenvalue problem is formulated without the factor  $n - 1$

$$Kv_i = \lambda_i v_i. \quad (6)$$

This gives the same eigenvectors  $v_i$  and eigenvalues  $n - 1$  times greater. In this case  $u_i = \Phi^T v_i / \sqrt{\lambda_i}$  and  $v_i = \Phi u_i / \sqrt{\lambda_i}$ .

**Basic Properties.** Several basic properties including the norm in feature space, the distance between observations in feature space, the norm of the mean in feature space, centering to zero mean in feature space, and standardization to unit variance in feature space, may all be expressed in terms of the kernel function without using the mapping by  $\phi$  explicitly [5,6,10].

**Projections onto Eigenvectors.** To find the kernel principal component scores from the eigenvalue problem in Equation 6 we project a mapped  $x$  onto the primal eigenvector  $u_i$

$$\begin{aligned} \phi(x)^T u_i &= \phi(x)^T \Phi^T v_i / \sqrt{\lambda_i} = \phi(x)^T [\phi(x_1) \phi(x_2) \cdots \phi(x_n)] v_i / \sqrt{\lambda_i} \\ &= [\kappa(x, x_1) \kappa(x, x_2) \cdots \kappa(x, x_n)] v_i / \sqrt{\lambda_i}, \end{aligned} \quad (7)$$

or in matrix notation  $\Phi U = K V \Lambda^{-1/2}$  ( $U$  is a matrix with  $u_i$  in the columns,  $V$  is a matrix with  $v_i$  in the columns and  $\Lambda^{-1/2}$  is a diagonal matrix with elements  $1/\sqrt{\lambda_i}$ ), i.e., also the projections may be expressed in terms of the kernel function without using  $\phi$  explicitly. If the mapping by  $\phi$  is not column centered the variance of the projection must be adjusted, cf. [5,6].

Kernel PCA is a so-called memory-based method: from Equation 7 we see that if  $x$  is a new data point that did not go into building the model, i.e., finding the eigenvectors and -values, we need the original data  $x_1, x_2, \dots, x_n$  as well as the eigenvectors and -values to find scores for the new observations. This is not the case for ordinary PCA where we do not need the training data to project new observations.

**Some Popular Kernels.** Popular choices for the kernel function are stationary kernels that depend on the vector difference  $x_i - x_j$  only (they are therefore invariant under translation in feature space),  $\kappa(x_i, x_j) = \kappa(x_i - x_j)$ , and homogeneous kernels also known as radial basis functions (RBFs) that depend on the Euclidean distance between  $x_i$  and  $x_j$  only,  $\kappa(x_i, x_j) = \kappa(\|x_i - x_j\|)$ . Some of the most often used RBFs are ( $h = \|x_i - x_j\|$ )

- multiquadric:  $\kappa(h) = (h^2 + h_0^2)^{1/2}$ ,
- inverse multiquadric:  $\kappa(h) = (h^2 + h_0^2)^{-1/2}$ ,
- thin-plate spline:  $\kappa(h) = h^2 \log(h/h_0)$ , or
- Gaussian:  $\kappa(h) = \exp(-\frac{1}{2}(h/h_0)^2)$ ,

where  $h_0$  is a scale parameter to be chosen. Generally,  $h_0$  should be chosen larger than a typical distance between samples and smaller than the size of the study area.

## 4 Maximum Autocorrelation Factor Analysis

In maximum autocorrelation factor (MAF) analysis we maximize the autocorrelation of linear combinations,  $a^T x(r)$ , of zero-mean original (spatial) variables,  $x(r)$ .  $x(r)$  is a multivariate observation at location  $r$  and  $x(r + \Delta)$  is an observation of the same variables at location  $r + \Delta$ ;  $\Delta$  is a spatial displacement vector.

### 4.1 Primal Formulation

The autocovariance  $R$  of a linear combination  $a^T x(r)$  of zero-mean  $x(r)$  is

$$R = \text{Cov}\{a^T x(r), a^T x(r + \Delta)\} \tag{8}$$

$$= a^T \text{Cov}\{x(r), x(r + \Delta)\}a \tag{9}$$

$$= a^T C_\Delta a \tag{10}$$

where  $C_\Delta$  is the covariance between  $x(r)$  and  $x(r + \Delta)$ . Assuming or imposing second order stationarity of  $x(r)$ ,  $C_\Delta$  is independent of location,  $r$ . Introduce the multivariate difference  $x_\Delta(r) = x(r) - x(r + \Delta)$  with variance-covariance matrix  $S_\Delta = 2 S - (C_\Delta + C_\Delta^T)$  where  $S$  is the variance-covariance matrix of  $x$  defined in Section 3. Since

$$a^T C_\Delta a = (a^T C_\Delta a)^T \tag{11}$$

$$= a^T C_\Delta^T a \tag{12}$$

$$= a^T (C_\Delta + C_\Delta^T) a / 2 \tag{13}$$

we obtain

$$R = a^T (S - S_\Delta / 2) a. \tag{14}$$

To get the autocorrelation  $\rho$  of the linear combination we divide the covariance by its variance  $a^T S a$

$$\rho = 1 - \frac{1}{2} \frac{a^T S_\Delta a}{a^T S a} \tag{15}$$

$$= 1 - \frac{1}{2} \frac{a^T X_\Delta^T X_\Delta a}{a^T X^T X a} \tag{16}$$

where the  $n$  by  $p$  data matrix  $X$  is defined in Section 3 and  $X_\Delta$  is a similarly defined matrix for  $x_\Delta$  with zero-mean columns.  $C_\Delta$  above equals  $X^T X_\Delta / (n-1)$ . To maximize  $\rho$  we must minimize the Rayleigh coefficient  $a^T X_\Delta^T X_\Delta a / (a^T X^T X a)$  or maximize its inverse.

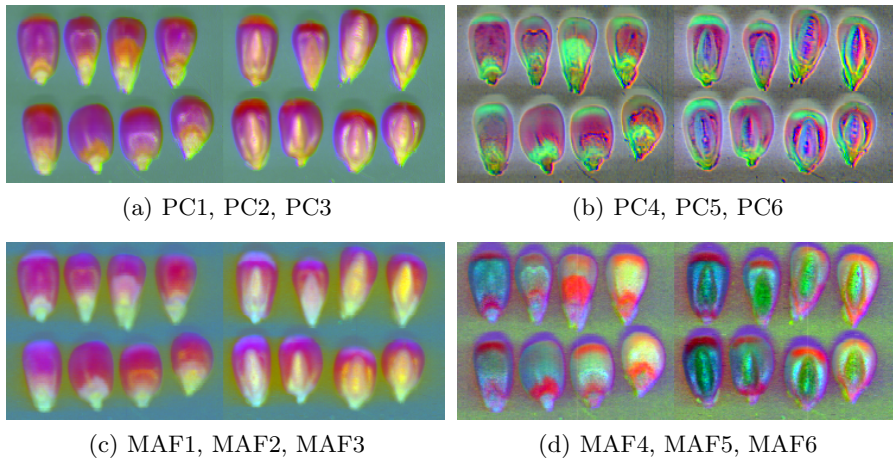
Unlike linear PCA, the result from linear MAF analysis is scale invariant: if  $x_i$  is replaced by some matrix transformation  $Tx_i$  corresponding to replacing  $X$  by  $XT$ , the result is the same.

## 4.2 Kernel MAF

As with the principal component analysis we use the kernel trick to obtain an implicit non-linear mapping for the MAF transform. A detailed account of this is given in [10].

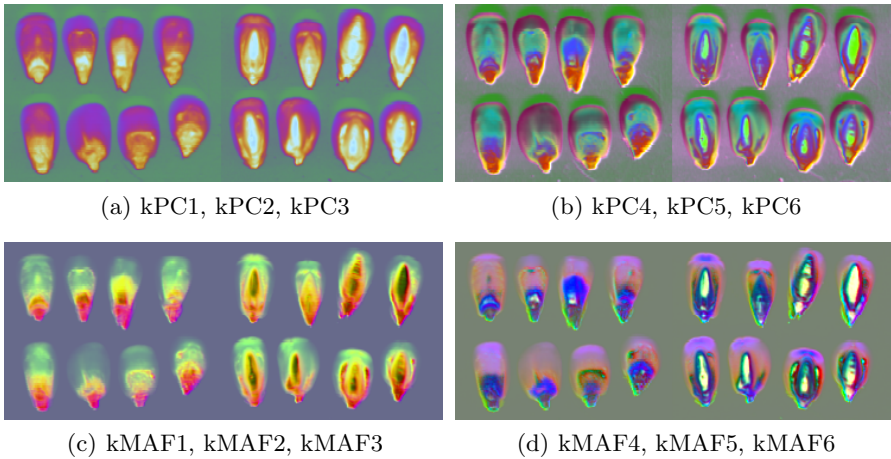
## 5 Results and Discussion

To be able to carry out kernel MAF and PCA on the large amounts of pixels present in the image data, we sub-sample the image and use a small portion termed the training data only. We typically use in the order  $10^3$  training pixels (here  $\sim 3,000$ ) to find the eigenvectors onto which we then project the entire image termed the test data kernelized with the training data. A Gaussian kernel  $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$  with  $\sigma$  equal to the mean distance between the training observations in feature space is used.



**Fig. 4.** Linear principal component projections of front and back sides of 8 maize kernels shown as RGB combination of factors (1,2,3) and (4,5,6) (two top panels), and corresponding linear maximum autocorrelation factor projections (bottom two panels)





**Fig. 5.** Non-linear kernel principal component projections of front and back sides of 8 maize kernel shown as RGB combination of factors (1,2,3) and (4,5,6) (two top panels), and corresponding non-linear kernel maximum autocorrelation factor projections (bottom two panels)

In Fig. 4 linear PCA and MAF components are shown as RGB combination of factors (1,2,3) and (4,5,6) are shown. The presented images are scaled linearly between  $\pm 3$  standard deviations. The linear transforms both struggle with the background noise, local illumination and shadow effects, i.e., all these effects are enhanced in some of the first 6 factors. Also the linear methods fail in labeling the same kernel parts in same colors. On the other hand the kernel based factors shown in Fig. 5 have a significantly better ability to suppress background noise, illumination variation and shadow effect. In fact this is most pronounced in the kernel MAF projections. When comparing kernel PCA and kernel MAF the most striking difference is the ability of the kernel MAF transform to provide same color labeling of different maize kernel parts across all grains.

## 6 Conclusion

In this preliminary work on finding interesting projections of hyperspectral near infrared imagery of maize kernels we have demonstrated that non-linear kernel based techniques implementing kernel versions of principal component analysis and maximum autocorrelation factor analysis outperform the linear variants by their ability to suppress background noise, illumination and shadow effects. Moreover, the kernel maximum autocorrelation factors transform provides a superior projection in terms of labeling different maize kernels parts with same color.

## References

1. Pearson, K.: On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2(3), 559–572 (1901)
2. Hotelling, H.: Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24, 417–441, 498–520 (1933)
3. Jolliffe, I.T.: *Principal Component Analysis*, 2nd edn. Springer, Heidelberg (2002)
4. Schölkopf, B., Smola, A., Müller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10(5), 1299–1319 (1998)
5. Shawe-Taylor, J., Cristianini, N.: *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge (2004)
6. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer, Heidelberg (2006)
7. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: *Numerical Recipes: The Art of Scientific Computing*, 3rd edn. Cambridge University Press, Cambridge (2007)
8. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. *Psychometrika* 1, 211–218 (1936)
9. Johnson, R.M.: On a theorem stated by Eckart and Young. *Psychometrika* 28(3), 259–263 (1963)
10. Nielsen, A.A.: Kernel minimum noise fraction transformation (2008) (submitted)
11. Switzer, P.: Min/Max Autocorrelation factors for Multivariate Spatial Imagery. In: Billard, L. (ed.) *Computer Science and Statistics*, pp. 13–16 (1985)
12. Hosney, R.C.: *Principles of Cereal Science and Technology*. American Association of Cereal Chemists (1994)
13. Belitz, H.-D., Grosch, W., Schieberle, P.: *Food Chemistry*, 3rd edn. Springer, Heidelberg (2004)