A statistical approach of identifying indexes crucial to characterizing Chinese yams in terms of shape

Koki Kyo,[1] **Mitsuru Hachiya**[2]

[1]Obihiro University of Agriculture and Veterinary Medicine *
[2]National Agricultural and Food Research Organization †

Abstract

We propose a statistical approach with which to identify the indexes crucial to classifying Chinese yams by shape. The proposed approach comprises two steps. The first step is to estimate the diameters indicating the shapes of yams along their length, and the second step is to characterize the shapes of sample yams employing principal component analysis and correlation analysis. It is found that the ratio of the weight to length of yams can be used as a crucial index for the purpose of classification of yams.

Keywords: Bayesian model, Chinese yam, Shape estimation and characterization

1. Introduction

Before Chinese yams are planted, it is necessary to cut them into chunks as seeds (see Fig. 1). To guarantee the seedlings grow equally, the chunks of yam must be cut equally. Nowadays, seed yams are mainly produced manually, resulting in excessive personnel costs and time. A mechanization of the production of seed yams is thus desired to improve the efficiency of operation.

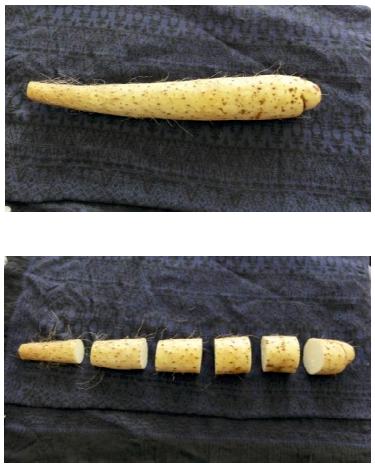

Fig. 1: A chinese yam (top) is cut into chunks as seeds (below)

A key problem in designing a machine for cutting yams is how to estimate the shapes of objective yams exactly and quickly. A series of methods with which to estimate the shapes of yams was reported by Kyo et al. (2012). However, the statistical models used to estimate the shapes of yams did not consider the differences in certain yam indexes, such as the length,

*Department of Agro-Environmental Science, Hokkaido 080-8555, Japan
†Institute of Agricultural Machinery, Saitama 331-8537, Japan

weight and functions of the two. If we can classify the objective yams using key indexes and estimate the shapes using different models for each class, then the estimation of yam shape can be improved.

In this paper, as a part of prospective work, we propose a statistical approach with which to identify the crucial indexes used to classify yams. The proposed approach comprises two steps. The main step is estimation of diameters indicating the shapes of yams along their lengths, and the second step is characterization of the shapes of sample yams using methods of principal component analysis (PCA) (see, for example, Srivastava (2002) for the detail of PCA) and correlation analysis.

2. Estimating diameters for a sample of yams

2.1. Model

The model considered here is for a sample of yams of size M. We take N points in equal intervals along the length of each sample. For the i-th yam in the sample with $i = 1, 2, \ldots, M$, we consider the model for the observation of the diameter at the j-th point:

$$\begin{aligned} &y_{ij} = d_{ij} + \epsilon_{ij}, \qquad (1)\\ &\epsilon_{ij} \sim \mathrm{N}(0,\ \sigma^2)\ (j = 1, 2, \ldots, N), \end{aligned}$$

where y_{ij}, d_{ij} and ϵ_{ij} are respectively the observation, the true value and the measurement error for the diameter. Note that the observations for y_{ij} are obtained rarely, so we take y_{ij} as a real observation if there is an observation near the j-th point, and we treat it as a missing observation if not. We aim to estimate d_{ij} from y_{ij} using the model in Eq. (1).

Here, taking a Bayesian approach, we treat d_{ij} as a random variable. It is assumed to be distributed with stochastic difference equations that are called smoothness priors (Kitagawa and Gersch, 1996). For a fixed value of i, we express the smoothness priors for d_{ij} by a second-order stochastic difference equation as

$$\begin{aligned} &d_{ij} = 2d_{i(j-1)} - d_{i(j-2)} + v_{ij}, \qquad (2)\\ &v_{ij} \sim \mathrm{N}(0,\ \tau_i^2)\ (j = 1, 2, \ldots, N). \end{aligned}$$

In Eqs. (1) and (2), ϵ_{ij} and v_{ij} are white-noise sequences on j and are independent of each other. Here, we treat σ^2 and τ_i^2 as unknown parameters.

We now put

$$\begin{aligned} \boldsymbol{z}_{ij} &= \begin{bmatrix} d_{ij} \\ d_{i(j-1)} \end{bmatrix},\\ \boldsymbol{F} &= \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix},\\ \boldsymbol{G} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ \boldsymbol{H} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^{\mathrm{t}}. \end{aligned}$$

The model in Eqs. (1) and (2) can then be expressed by a state space model as

$$\begin{aligned} \boldsymbol{z}_{ij} &= \boldsymbol{F}\boldsymbol{z}_{j-1} + \boldsymbol{G}v_{ij}, \qquad (3)\\ y_{ij} &= \boldsymbol{H}\boldsymbol{z}_{ij} + \epsilon_{ij}. \qquad (4) \end{aligned}$$

In the state space model comprising Eqs. (3) and (4), the true diameter d_{ij} is included in the state vector $\boldsymbol{z}_{ij}$, and its estimate can thus be obtained from the estimate of $\boldsymbol{z}_{ij}$. Moreover, the variances σ^2 and τ_i^2 can be estimated with the maximum likelihood method.

2.2. Estimation of diameters

For a fixed value of i, let $\boldsymbol{z}_{i0}$ denote the initial value of the state and $Y_{im} = \{y_{in}; n = 1, 2, \ldots, m\}$ denote a set of observations up to the point

m. Assume that $\boldsymbol{z}_{i0} \sim \mathrm{N}(\boldsymbol{z}_{i0|0}, \boldsymbol{C}_{i0|0})$. It is well known that the distribution $f(\boldsymbol{z}_{ij}|Y_{im})$ for the state $\boldsymbol{z}_{ij}$ conditionally on Y_{im} is Gaussian, and it is thus only necessary to obtain the mean $\boldsymbol{z}_{ij|m}$ and covariance matrix $\boldsymbol{C}_{ij|m}$ of $\boldsymbol{z}_{ij}$ with respect to $f(\boldsymbol{z}_{ij}|Y_{im})$.

When the values of σ^2 and τ_i^2, the initial distribution $\mathrm{N}(\boldsymbol{z}_{i0|0}, \boldsymbol{C}_{i0|0})$, and a set of observations Y_{iN} up to the point N are given, the estimates for the state $\boldsymbol{z}_{ij}$ can be obtained using the well-known Kalman filter (for $j = 1, 2, \ldots, N$) and fixed-interval smoothing (for $j = N-1, N-2, \ldots, 1$) recursively (see, for example, Anderson and Moore (1979), and Kitagawa and Gersch (1996)):

[**Kalman filter**]

$$
\begin{aligned}
\boldsymbol{z}_{ij|j-1} &= \boldsymbol{F}\boldsymbol{z}_{i(j-1)|j-1}, \\
\boldsymbol{C}_{ij|j-1} &= \boldsymbol{F}\boldsymbol{C}_{i(j-1)|j-1}\boldsymbol{F}^{\mathrm{t}} \\
&\quad +\tau_i^2 \boldsymbol{G}\boldsymbol{G}^{\mathrm{t}}. \\
\boldsymbol{L}_{ij} &= \boldsymbol{C}_{ij|j-1}\boldsymbol{H}^{\mathrm{t}} \\
&\quad \times (\boldsymbol{H}\boldsymbol{C}_{ij|j-1}\boldsymbol{H}^{\mathrm{t}} + \sigma^2)^{-1}, \\
\boldsymbol{z}_{ij|j} &= \boldsymbol{z}_{ij|j-1} \\
&\quad +\boldsymbol{L}_{ij}(y_{ij} - \boldsymbol{H}\boldsymbol{z}_{ij|j-1}), \\
\boldsymbol{C}_{ij|j} &= (\boldsymbol{I} - \boldsymbol{L}_{ij}\boldsymbol{H})\boldsymbol{C}_{ij|j-1}.
\end{aligned}
$$

[**Fixed-Interval Smoothing**]

$$
\begin{aligned}
\boldsymbol{P}_{ij} &= \boldsymbol{C}_{ij|j}\boldsymbol{F}^{\mathrm{t}}\boldsymbol{C}_{i(j+1)|j}^{-1}, \\
\boldsymbol{z}_{ij|N} &= \boldsymbol{z}_{ij|j} + \boldsymbol{P}_{ij}(\boldsymbol{z}_{i(j+1)|N} \\
&\quad -\boldsymbol{z}_{i(j+1)|n}), \\
\boldsymbol{C}_{ij|N} &= \boldsymbol{C}_{ij|j} + \boldsymbol{P}_{ij}(\boldsymbol{C}_{i(j+1)|N} \\
&\quad -\boldsymbol{C}_{i(j+1)|j})\boldsymbol{P}_{ij}^{\mathrm{t}}.
\end{aligned}
$$

Here, $\boldsymbol{I}$ denotes an identity matrix. Note that the calculation in the step of the filter (the last three lines in the Kalman filter) will be skipped when y_{ij} is a missing observation.

The posterior distribution of $\boldsymbol{z}_{ij}$ can then be given by $\boldsymbol{z}_{ij|N}$ and $\boldsymbol{C}_{ij|N}$, and, subsequently, the estimates for the true diameters d_{ij} can be obtained because the state space model described by Eqs. (3) and (4) incorporates d_{ij} in the state vector $\boldsymbol{z}_{ij}$. Hereafter, the estimates of d_{ij} are denoted $\widehat{d}_{ij}$.

2.3. Estimation of variances

For a fixed value of i, when the observations $Y_{iN} = \{y_{i1}, y_{i2}, \ldots, y_{iN}\}$ are given, a likelihood function for the variances σ^2 and τ_i^2 is defined approximately by

$$
\begin{aligned}
& f(Y_{iN}|\sigma^2, \tau_i^2) \qquad (5) \\
&\quad = \prod_{m=1}^{N} f_m(y_{im}|Y_{i(m-1)}; \sigma^2, \tau_i^2),
\end{aligned}
$$

where $f_m(y_{im}|Y_{i(m-1)}; \sigma^2, \tau_i^2)$ is the conditional density function of y_{im} given the past history

$$
Y_{i(m-1)} = \{\ldots, y_{i(m-2)}, y_{i(m-1)}\}
$$

with $Y_{i0} = \{\ldots, y_{i(-1)}, y_{i0}\} = \Phi$ being an empty set; then

$$
f_1(y_{i1}|Y_{i0}; \sigma^2, \tau_i^2) = f_1(y_{i1}|\sigma^2, \tau_i^2).
$$

As given by Kitagawa and Gersch (1996), under the use of a Kalman filter, the conditional density $f_m(y_{im}|Y_{i(m-1)}; \sigma^2, \tau_i^2)$ is as a normal density given by

$$
\begin{aligned}
& f_m(y_{im}|Y_{i(m-1)}; \sigma^2, \tau_i^2) \\
&\quad = \frac{1}{\sqrt{2\pi w_{im|m-1}}} \\
&\quad \times \exp\left\{ -\frac{(y_{im} - \widehat{y}_{im|m-1})^2}{2w_{im|m-1}} \right\},
\end{aligned}
$$

where $\widehat{y}_{im|m-1}$ is the one-step ahead prediction for y_{im} and $w_{im|m-1}$ is the variance of the predictive error, which are respectively given by

$$
\begin{aligned}
\widehat{y}_{im|m-1} &= \boldsymbol{H}\boldsymbol{z}_{im|m-1}, \\
w_{im|m-1} &= \boldsymbol{H}\boldsymbol{C}_{im|m-1}\boldsymbol{H}^{\mathrm{t}} + \sigma^2.
\end{aligned}
$$

Table 1: Summary of the first three PCs

	Comp.1	Comp.2	Comp.3
Standard deviation	50.2797	24.2002	11.7821
Proportion of Variance	0.7185	0.1665	0.0395
Cumulative Proportion	0.7185	0.8850	0.9245

Thus, the estimates of σ^2 and τ_i^2 can be obtained using the maximum likelihood method. Concretely, firstly for a given value of σ^2, we can obtain the estimate $\widehat{\tau}_i^2$ of τ_i^2 for $i = 1, 2, \ldots, M$ by maximizing $f(Y_{iN}|\sigma^2, \tau_i^2)$ in Eq. (5) numerically. The estimate $\widehat{\sigma}^2$ for σ^2 is then obtained by maximizing

$$\bar{\ell}(\sigma^2) = \frac{1}{M}\sum_{i=1}^{M} \log f(Y_{iN}|\sigma^2, \tau_i^2),$$

which is the average of the log-likelihood, similar to the case in the numerical method. By applying the results of $\widehat{\sigma}^2$ and $\widehat{\tau}_i^2$ to the above algorithms of the Kalman filter and fixed-interval smoothing, we can obtain the final estimates $\widehat{d}_{ij}$ of d_{ij} from the results of $\boldsymbol{z}_{ij|N}$.

3. Indentifying indexes crucial to the purpose

From the results of the above section, we can obtain the estimates $\widehat{d}_{ij}$ for the true diameter at the j-th point of the i-th yam. Here we use $\widehat{d}_{ij}$ for $i = 1, 2, \ldots, M;\ j = 1, 2, \ldots, N$ as a set of multivariate data for N variables and M individuals. Note that $N = 100$ and $M = 111$ for the present data analyzed below.

To reduce the dimensions of the data, we employ PCA. PCA is a statistical tool that generates principal components (PCs) according to a set of orthogonal transformations of the data

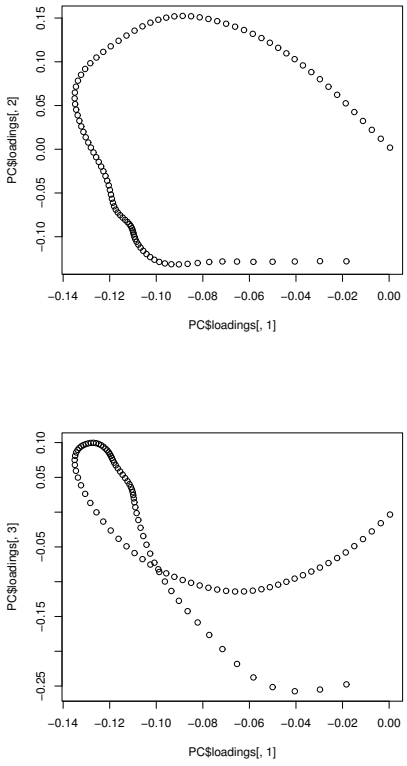

Fig. 2: Scatter diagrams for the PC loadings: first vs. second (top) and first vs. third (below)

matrix. Table 1 summarizes the PCA for the first three PCs.

From Table 1 we find that the cumulative proportion is about 0.92 for the first three PCs; i.e., about 92% of total variance in the data can be captured by the first three PCs having the largest variances.

Table 2: Correlation coefficients for the indexes and PCs

	Weight	Length	W/L	Comp.1	Comp.2	Comp.3
Weight	1.0000					
Length	0.7396	1.0000				
W/L	0.8401	0.2718	1.0000			
Comp.1	-0.7648	-0.2231	-0.9483	1.0000		
Comp.2	-0.1275	-0.1858	-0.0116	0.0000	1.0000	
Comp.3	-0.2843	-0.3640	-0.0964	0.0000	0.0000	1.0000

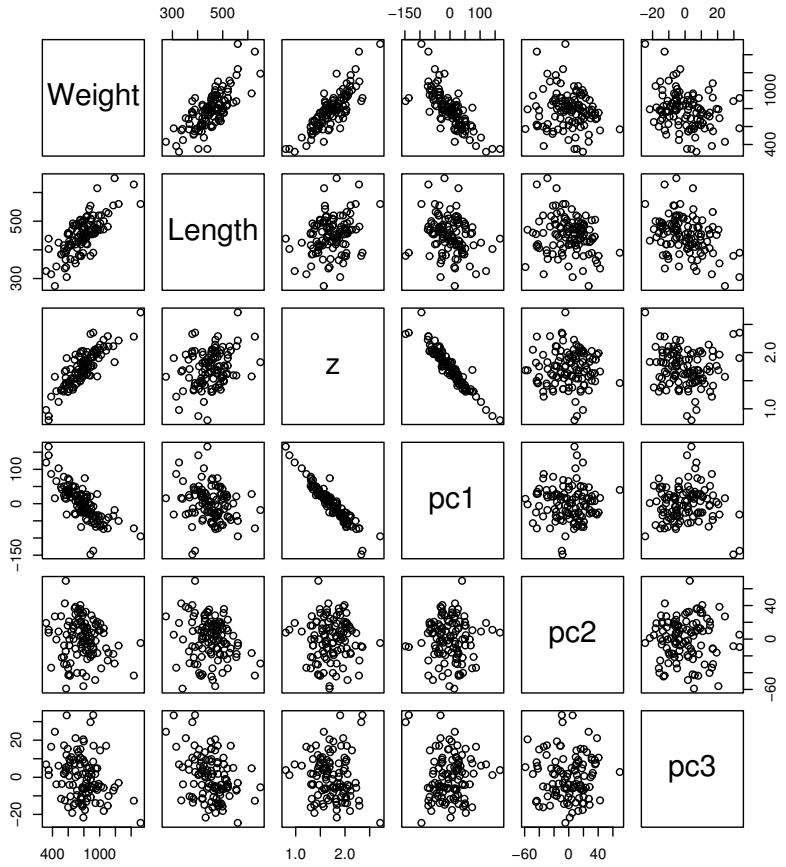

Fig. 3: Matrix of scatter diagrams for the indexes and PCs

Figure 2 plots scatter diagrams for the PC loadings. The diagram at top is for the first vs. the second PC loadings, and that below is for the first vs. the third PC loadings. The figure shows that the patterns of the PC loadings are unique and interesting.

From the results of the above PCA,

we obtain three PCs that can be applied as latent indexes in the indication of the characteristics of the shapes of the sample yams. However, these PCs cannot be measured beforehand, and thus cannot serve to improve the model. Thus, for yams for which the shapes need to be predicted, we have to take variables that can be measured prior to the estimation of diagrams as the crucial indexes in characterizing the shape. We consider here the length (L), weight (W) and $z = W/L$ as candidates of the crucial indexes.

To compare the performances of these indexes, we calculate coefficients of correlation between the indexes and the first three PCs. The results are presented in Table 2.

Figure 3 is a matrix of scatter diagrams for the indexes and PCs.

Table 2 and Fig. 3 show that there is very strong negative correlation between $z = W/L$ and the first component. Thus, $z = W/L$ can be taken as a crucial index for characterizing yams in terms of shape.

4. Conclusions

In this paper, we proposed a statistical approach with which to identify the indexes crucial for classifying yams. In the proposed approach, we firstly applied a Bayesian modeling method to estimate the diameters according to a sample of 111 yams. Then, to characterize the shapes of sample yams, we used PCA together with correlation analysis. We found that the ratio of the weight to length of yams can be used as a crucial index for the purpose of classification of Chinese yams.

References

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, New Jersey, 1979.

[2] G. Kitagawa and W. Gersch, *Smoothness Priors Analysis of Times Series*, Springer-Verlag, New York, 1996.

[3] K. Kyo et al., Development of methods for estimating the shape of Chinese yams, In *Proceedings of Annual Conference of Societies related to Statistics in Japan*, page 236, 2012 (in Japanese).

[4] M. S. Srivastava, *Methods of Multivariate Statistics*, John Wiley and Sons, New York, 2002.