Improving Tone Prediction Accuracy in Calibration for Color Electrophotography Part II – Principal Component Regression

Yan-Fu Kuo[a], Chao-Lung Yang [b], George T.-C. Chiu[a], Yuehwern Yih[b], and Jan P. Allebach[c]; a: School of Industrial Engineering, b: School of Mechanical Engineering, c: School of Electrical and Computer Engineering, Purdue University, W Lafayette, IN 47907

Abstract

This paper presents results of improved tone prediction accuracy in calibration through a principal component analysis based regression approach for color electrophotography (EP). During calibration, multiple color patches of the same primary color at different halftone levels are printed on a belt with black exterior and are measured using on-board sensors. Regression models are often developed to predict the primary color tone values on output media from these on-board sensor measurements. The prediction accuracy of the regression models directly impact the quality and consistency of the calibration. Analyses have revealed a high degree of correlation among the color patch measurements, which results in using multicollinear measurements as explanatory variables during regression analysis to identify model coefficients. It is well known that using collinear explanatory variables during regression analysis will result in sub-optimal model coefficients that will degrade the prediction accuracy. In this study, a principle component regression (PCR) approach is applied to tackle the potential issue with collinear measurement data in model coefficient estimation. The experimental results show the resulting PCR models provides 25% improvement on average in root-mean-squared predication accuracy over separate ordinary least square regression models.

Introduction

A color electrophotographic (EP) printing system typically uses four primary colors – cyan, magenta, yellow, and black. Calibrations are performed to maintain consistent color reproduction under different operating conditions. During a calibration, multiple patches of different halftone levels of the same primary color are printed on an intermediate media and are measured with on-board sensors. Calibration models are used to predict the primary color tone values on output media from these on-board sensor measurements. In this study, our aim is to improve the prediction accuracy of the calibration models through a principal component regression (PCR) approach for color EP systems.

The calibration models are developed with life test data. In typical life test, additional color patches are printed on output media immediately following calibration. Their tone values, output tone values, are measured off-line with measurement devices, such as photometers. Calibration models are developed to map the on-board sensor measurements to the output tone values [1]. During calibration, output tone values are predicted based on the on-board sensor measurements using the calibration model. Appropriate tone correction is performed to regulate the color reproduction.

Halftoned color images composed of arrays of closely spaced micro-dots. Changes in operating conditions or different EP parameter settings will impact the sizes of the micro-dots. Assuming operating conditions and EP parameter settings have consistent impacts on the sizes of the micro-dots for a given print, the effects of micro-dot size fluctuation can be detected by the on-board sensor from different halftone patches of the same color. This results in increased correlation among the on-board sensor measurements, i.e., multicollinearities. It is well known that using the collinear measurements as explanatory variables to identify model coefficients directly through ordinary least-square regression (OLSR) will result in sub-optimal model coefficients that will degrade prediction accuracy [2]. Hence, existing calibration models are developed using a single-response regression approach, i.e., the output tone value at a particular halftone level is regressed only with the on-board sensor measurement at the same halftone level.

Recent researches in regression analysis have shown improved prediction accuracy of regression models using multiple explanatory variables as compared to single-response regression models [3, 4]. To address the multicollinearities associated with multiple on-board sensor measurements, in this study, a principal component regression (PCR) approach is proposed [5]. PCR avoids the numerical issues associated with the OLSR by transforming multicollinear sensor measurements into a set of orthogonal principal components (PC) basis. In addition, it achieves biased regression by determining an optimal subset of PC's to be retained while discarding PC's that are less statistically significant. To illustrate the utility of the proposed approach, a first-order linear calibration model for an off-the-shelf in-line color EP printer is developed using existing life test data. Cross-validation results demonstrate 25% improvement in prediction accuracy compared with the existing calibration model.

Method

Calibration Model

Since each primary color is reproduced independently for a single-pass color EP process, a calibration model is developed for each primary color. A calibration model can be written as $\boldsymbol{y} = f(\boldsymbol{z},\boldsymbol{d})$, where $\boldsymbol{y}$ is tone values on paper, $\boldsymbol{z}$ is on-board sensor measurements, and $\boldsymbol{d}$ is uncontrollable but measurable factors/disturbances, such as temperature and humidity. The tone values, $\boldsymbol{y}$, are the measured intensities of the reproduced color patches printed at designated halftone levels. In this study, a tone value is defined as the Euclidian distance (ΔE) in CIE L*a*b* space between the color point of a primary color printed at a particular halftone level and the substrate appearance color. A static linear calibration model is used in this study.

Problem Formulation

Life test data are used to identify the calibration models. For one observation, a set of on-board sensor measurements, measurable disturbances, and the corresponding tone values measured on paper are collected. Denote $z_{ij} \in \Re$ as the j^{th} on-board sensor measurement in the i^{th} observation, $y_{ij} \in \Re$ as the j^{th} tone value measurement in the i^{th} observation, and $d_{ij} \in \Re$ as the j^{th} measurable disturbances in the i^{th} observation. Denote $\boldsymbol{G}$ as the calibration model. The calibration model, $\boldsymbol{G}$, is a linear transformation relating the tone value measurements, $\boldsymbol{y}$, to the sensor measurements, $\boldsymbol{z}$, and the disturbances, $\boldsymbol{d}$.

Consider p on-board sensor measurements, q measurable disturbances, and w tone value measurements are made in one observation, and n observations are made in the life test. Denote $\boldsymbol{Z} = [z_{ij}] \in \Re^{n \times p}$ as the sensor measurement matrix and $\boldsymbol{D} = [\boldsymbol{d}_{ij}] \in \Re^{n \times q}$ as the measurable disturbance matrix. Let $\boldsymbol{X} \in \Re^{n \times r}$ denote the explanatory variable matrix, which is the augmented matrix consisting of matrices $\boldsymbol{Z}$ and $\boldsymbol{D}$, i.e., $\boldsymbol{X} = [\boldsymbol{Z}|\boldsymbol{D}]$ and $r = p + q$. Denote $\boldsymbol{Y} = [\boldsymbol{y}_{ij}] \in \Re^{n \times w}$ as the response variable matrix containing the tone value measurements. Denote $\boldsymbol{G} \in \Re^{r \times w}$ as the calibration model that can be written as $\boldsymbol{Y} = \boldsymbol{XG}$. Note that the matrices are assumed to be centered column-wise. Hence no constant or intercept term is required in the regression model development.

Ordinary Least Square Regression

Consider a standard multivariate regression model:

$$\boldsymbol{Y}=\boldsymbol{XG}+\boldsymbol{E} ,$$

where the error matrix, $\boldsymbol{E}$, satisfies the usual assumption of being independent and identically-distributed. The number of observations is usually larger than the number of calibration color patches printed in a calibration, i.e., $n >> r$. The OLSR solution to the over-determined problem stated above is to minimizes the loss function, i.e.,

$$\boldsymbol{G} = \arg\min \left\| \boldsymbol{Y} - \boldsymbol{XG} \right\|^2 = \arg\min \left\| \boldsymbol{Y} - [\boldsymbol{Z} \mid \boldsymbol{D}] \begin{bmatrix} \boldsymbol{G_Z} \\ \boldsymbol{G_D} \end{bmatrix} \right\|^2 , \quad (1)$$

where the calibration model, $\boldsymbol{G}$, can be split into two matrices $\boldsymbol{G_Z}$ and $\boldsymbol{G_D}$ with proper dimensions corresponding to the sensor measurement matrix, $\boldsymbol{Z}$, and the measurable disturbance matrix, $\boldsymbol{D}$, respectively. The OLSR solution to Eq. (1) is given by

$$\boldsymbol{G}=(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \quad (2)$$

The explanatory variable matrix, $\boldsymbol{X}$, is not of full rank since the column vectors are collinear. The calculation of the matrix $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ is computational challenging especially with matrices that with lower conditioning number. This yields larger variance in model coefficient estimation.

Variance Inflation Factor

Variance inflation factor [6] (VIF) is commonly used to detect multicollinearity among explanatory variables. It is defined as

$$(VIF)_j = \frac{1}{1 - R_j^2} , \quad (3)$$

where R_j^2 stands for the unadjusted coefficient of determination of the j^{th} explanatory variable when it is predicted by the other explanatory variables included in the model. Suppose the j^{th} explanatory variable is linearly correlated to any the other explanatory in the model, R_j^2 is large and, consequently, the VIF value is large. Values of VIF that exceed 10 are often regarded as indicating strong multicollinearity [7] among the explanatory variables and OLSR may not be a good regression approach.

Principal Component Regression

The key idea of PCR is to linearly transform the multicollinear sensor measurement matrix, $\boldsymbol{Z}$, to a principle component (PC) matrix that consists of a set of orthogonal vectors. The model coefficient estimation can be directly carried out following Eq. (2). A singular value decomposition (SVD) on the sensor measurement matrix, $\boldsymbol{Z}$, is performed as the first step to calculate the PC matrix, i.e.,

$$\boldsymbol{Z}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T=\sum_{i=1}^{p} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T ,$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1, \ \sigma_2, \ \ldots, \ \sigma_p) \in \Re^{n \times p}$ is a diagonal matrix of singular values σ_i associated with the principal component PC_i, and $\boldsymbol{U} \in \Re^{n \times n}$ and $\boldsymbol{V} \in \Re^{p \times p}$ are left and right unitary matrices of corresponding singular vectors, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, respectively. The PC matrix, $\boldsymbol{\Psi} \in \Re^{n \times p}$, can be obtained by multiplying the sensor measurement matrix, $\boldsymbol{Z}$, with the right unitary matrix, V, i.e., $\boldsymbol{\Psi} = \boldsymbol{ZV}$. The principal component PC_i is a linear combination of the raw sensor measurements with the coefficients in the associated row vector, $\boldsymbol{v}_i$.

Next, the PC matrix is augmented with the disturbance matrix as the explanatory variable matrix, i.e., $\boldsymbol{X} = [\boldsymbol{\Psi}|\boldsymbol{D}]$, in the subsequent multivariate regression. The loss function is

$$\boldsymbol{\Gamma} = \arg\min \left\| \boldsymbol{Y} - [\boldsymbol{\Psi} \mid \boldsymbol{D}] \begin{bmatrix} \boldsymbol{\Gamma_\Psi} \\ \boldsymbol{\Gamma_D} \end{bmatrix} \right\|^2 , \quad (4)$$

where $\boldsymbol{\Gamma} \in \Re^{r \times w}$ is the coefficient matrix to be determined. The coefficient matrix, $\boldsymbol{\Gamma}$, can be split into two matrices $\boldsymbol{\Gamma_Z}$ and $\boldsymbol{\Gamma_D}$ with proper dimensions corresponding to the PC matrix, $\boldsymbol{\Psi}$, and the disturbance measurement matrix, $\boldsymbol{D}$, respectively. Since the PC matrix, $\boldsymbol{\Psi}$, is of full rank, the solution of the coefficient matrix, $\boldsymbol{\Gamma}$, in Eq. (4) can be carried out directly following Eq. (2). Matching the response variable matrix, $\boldsymbol{Y}$, in Eq. (1) and (4), one can obtain

$$\boldsymbol{Y} = \boldsymbol{\Psi}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Gamma_\Psi} + \boldsymbol{D}\boldsymbol{\Gamma_D} = \boldsymbol{Z}\boldsymbol{G_Z} + \boldsymbol{D}\boldsymbol{G_D} . \quad (5)$$

Therefore the calibration model is

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G_Z} \\ \boldsymbol{G_D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{V}\boldsymbol{\Gamma_\Psi} \\ \boldsymbol{\Gamma_D} \end{bmatrix} .$$

Biased Principal Component Regression

The noise in the sensor measurements can result in bias in regression analysis and, consequently, can increase the uncertainty in model coefficient estimation. Biased PCR excludes the PC's regarded as noise from being used in the regression. Assume the noise in the sensor measurement matrix, $\boldsymbol{Z}$, is additive. It can be decomposed into two matrices – an exact signal matrix $\boldsymbol{S}$, and a noise perturbation matrix $\boldsymbol{N}$ – so that

$$\boldsymbol{Z} = \boldsymbol{S} + \boldsymbol{N} = \boldsymbol{U}_S \boldsymbol{\Sigma}_S \boldsymbol{V}_S^T + \boldsymbol{U}_N \boldsymbol{\Sigma}_N \boldsymbol{V}_N^T,$$

where $\boldsymbol{\Sigma}_S$, $\boldsymbol{U}_S$, and $\boldsymbol{V}_S$, and $\boldsymbol{\Sigma}_N$, $\boldsymbol{U}_N$, and $\boldsymbol{V}_N$ are the singular value matrix, left matrix, and right matrix associated with the SVD of the signal matrix, $\boldsymbol{S}$, and the noise perturbation matrix, $\boldsymbol{N}$, respectively. If PC_i is regarded as noise, the corresponding singular value, σ_i, left singular vector, $\boldsymbol{u}_i$, and right singular vectors, $\boldsymbol{v}_i$, are put to the noise perturbation matrix, $\boldsymbol{N}$. Then the biased PC matrix, $\boldsymbol{\Psi}_S = \boldsymbol{Z}\boldsymbol{V}_S$, is used in the subsequent regression.

Principal Component Selection

A forward selection algorithm is used to determine the PC's to be included in the regression. This study utilizes forward selection proposed by Xie and Kalivas [8]. The forward selection tries out the PC's one by one and includes one PC in the model if it is statistically significant to the response variables. Bayesian information criterion (BIC) [6]

$$n \cdot \ln\left(\frac{RSS}{n}\right) + k \cdot \ln(n), \tag{6}$$

is used as the selection criterion, where *RSS* is residual sum of squares from the estimated model and k is number of the PC's to be included in the forward selection. BIC is known to be more parsimonious compared to other information criterion. Hence the chance of over-fitting can be reduced with using BIC as the selection criterion. The PC selection procedure can be summarized in the following four steps:

Step 1: Compute all the PC's through SVD.
Step 2: Determine the first PC producing the minimum selection criterion following Eq. (6). Call this the first PC subset.
Step 3: Identify the second PC subset as the subset of PC's providing the minimum selection criterion from all possible combinations containing the first PC subset and one more PC that has not been included in the first PC subset. Compute the selection criterion of the second PC subset following Eq. (6).
Step 4: The process stops when the selection criterion of the second subset is larger than that of the first subset or when all PC's are included in the regression. Otherwise, let the second subset be the first subset and continue from step 3.

Preferably, the PC selection can be performed for an individual response variable to preserve the freedom of PC retention. Then the individual response variable is regressed with the selected PC's. The calibration model is the integration of the coefficients from the regression analyses. The signal matrix consisting of selected PC's for the m^{th} response variable can be expressed as

$$\boldsymbol{S}^{(\mathrm{m})} = \boldsymbol{U}_{\mathrm{S}}^{(\mathrm{m})} \boldsymbol{\Sigma}_{\mathrm{S}}^{(\mathrm{m})} (\boldsymbol{V}_{\mathrm{S}}^{(\mathrm{m})})^{\mathrm{T}}. \tag{7}$$

The biased PC matrix of the m^{th} response variable, $\boldsymbol{\Psi}_S^{(m)}$, can be obtained by multiplying the sensor measurements matrix, $\boldsymbol{Z}$, with the right unitary matrix, $\boldsymbol{V}_S^{(m)}$ from Eq. (7), i.e., $\boldsymbol{\Psi}_S^{(m)} = \boldsymbol{Z}\boldsymbol{V}_S^{(m)}$. Let $\boldsymbol{y}^{(m)} \in \Re^n$ denotes the m^{th} column vector in the response variable matrix, $\boldsymbol{Y}$. The loss function to be minimized for the m^{th} response variable is

$$\boldsymbol{\gamma}^{(m)} = \arg\min \left\| \boldsymbol{y}^{(m)} - \left[\boldsymbol{\Psi}_S^{(m)} \mid \boldsymbol{D}\right] \begin{bmatrix} \boldsymbol{\gamma}_S^{(m)} \\ \boldsymbol{\gamma}_D^{(m)} \end{bmatrix} \right\|^2, \tag{8}$$

where $\boldsymbol{\gamma}^{(m)} \in \Re^r$ is the coefficient vector corresponding to $\boldsymbol{y}^{(m)}$ to be determined in the regression. The coefficient vector, $\boldsymbol{\gamma}^{(m)}$, can be split into two vectors, $\boldsymbol{\gamma}_S^{(m)}$ and $\boldsymbol{\gamma}_D^{(m)}$, with proper dimensions corresponding to the biased PC matrix, $\boldsymbol{\Psi}_S^{(m)}$, and the disturbance measurement matrix, $\boldsymbol{D}$, respectively. The calibration model can be obtained by combining the product vectors of multiplication of the coefficient vectors, $\boldsymbol{\gamma}_S^{(m)}$, from Eq. (8) and the associated right unitary matrix, $\boldsymbol{V}_S^{(m)}$, from Eq. (7) with the coefficient vectors, $\boldsymbol{\gamma}_D^{(m)}$, i.e.,

$$\boldsymbol{G} = \left[\frac{[\boldsymbol{V}_S^{(1)}\boldsymbol{\gamma}_S^{(1)} \mid \cdots \mid \boldsymbol{V}_S^{(w)}\boldsymbol{\gamma}_S^{(w)}]}{[\boldsymbol{\gamma}_D^{(1)} \mid \cdots \mid \boldsymbol{\gamma}_D^{(w)}]} \right].$$

Experiment

The proposed PCR procedure is performed on an off-the-shelf one-pass color EP printer. The printer prints and measures nine calibration patches at different halftone levels for each primary color during a calibration, i.e., p=9. These halftone levels are labeled as HL_j, j=1…9, from light to dark. Calibration patches identical to those printed in calibration are printed on output media for each primary color immediately following a calibration. Their tone value measurements are made with spectrophotometers (X-Rite® DTP-70). A commercial white paper (Xerox®) is used as the output media.

The experiment is performed on twenty printers with several consumable sets across a wide range of environmental conditions. The temperature ranges from 15 to 30°C, and the relative humidity ranges from 10 to 80%. Totally more than four hundred observations, i.e., $n > 400$, are made. The models are identified following the proposed PCR procedure using Matlab®. Humidity ratio is chosen to be the measurable disturbance.

VIF of the Sensor Measurements

Table 1 lists the variance inflation factor (VIF) of the experimental sensor measurements based on Eq. (3). The VIF values indicate a high degree of multicollinearity among the sensor measurements, especially for those sensor measurements of the color patches printed with halftone levels in the mid-tone and shadow range. Particularly, the VIF value can be as large as fifty for yellow. The large VIF values demonstrate the necessity to conduct the model development through the PCR.

TABLE I: Variance inflation factor values of the sensor measurements at each halftone level (HL)

	Cyan	Magenta	Yellow	Black
HL_1	1.3	1.2	1.3	1.3
HL_2	3.9	2.6	4.6	2.2
HL_3	7.4	5.3	7.6	4.7
HL_4	8.8	7.2	22.5	8.6
HL_5	11.6	13.8	26.3	10.7
HL_6	17.9	14.6	27.0	14.9
HL_7	19.9	16.4	53.0	23.7
HL_8	18.9	16.2	31.9	15.8
HL_9	11.2	7.6	14.1	12.9

Model Comparison

PCR models are developed through the proposed methods and are compared with separate OLSR models in which a tone value is regressed with single sensor measurement and the humidity ratio at each halftone level. A statistical F-test is conducted to compare the two types of models. The results show the PCR models are significantly better than the OLSR models at all halftone levels for all colors at 99% confidence level.

A 10-fold cross-validation (CV) without replacement is performed with the experimental data (see Fig. 1). Overall the PCR models provide 25% improvement on average over the OLSR models.

Conclusion

A PCR method to improve tone prediction accuracy of calibration models for color EP systems is proposed in this work. A high degree of multicollinearity among calibration color patch measurements is verified through experiments and statistical analyses. This motivates using PCR for calibration model identification. The proposed method includes a forward selection algorithm to determine the optimal subset of PC's to be retained in biased PCR. The effectiveness of the proposed PCR method is verified with experimental data collected under different environmental conditions and consumable usage levels. Statistical tests demonstrate the proposed PCR models outperform separate OLSR models. The PCR models provide 25% improvement on average in root-mean-squared predication accuracy over OLSR models based on cross-validation.

Acknowledgements

We gratefully acknowledge the support from the Hewlett-Packard Company. We would like to specially thank Dennis Abramsohn and Jeff Trask for their valuable guidance in this research.

References

[1] D. A. Johnson, "Calibration of printing devices," U. S. Patent 6,982,812 (2006).

[2] D. C. Montgomery and D. J. Friedman, "Prediction using regression models with multicollinear predictor variables," IIE Transactions, 25: 73 (1993).

[3] L. Breiman and J. H. Friedman. "Predicting multivariate responses in multiple linear regression," J. Roy. Statist. Soc. B, 59: 3 (1997).

[4] R. D. De Veaux and L. H. Ungar. "Multicollinearity: a tale of two nonparametric regressions," in Selecting Models from Data: AI and Statistics, (Springer-Verlag, New York, NY, 1994), pg. 293-302.

[5] W. F. Massy. "Principal components regression in exploratory statistical research," J. of the American Statistical Association, 60: 234 (1965).

[6] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, Applied linear regression models, 4th Ed., (McGraw-Hill Irwin, New York, 2004).

[7] P. J. Curran, S. G. West, and J. F. Finch, "Model-dependent variance inflation factor cutoff values," Quality Engineering, 14: 391 (1996).

[8] Y. Xie and J. Kalivas, "Evaluation of principal component selection methods to form a global prediction model by principal component regression," Analytica Chimica Acta, 348: 19 (1997).

Author Biography

Yan-Fu Kuo is currently a Ph.D. student at School of Mechanical Engineering, Purdue University. He worked as an intern engineer at Hewlett-Packard Company in 2007 and 2008, developing calibration tools to improve color consistency for electrophotography printers.

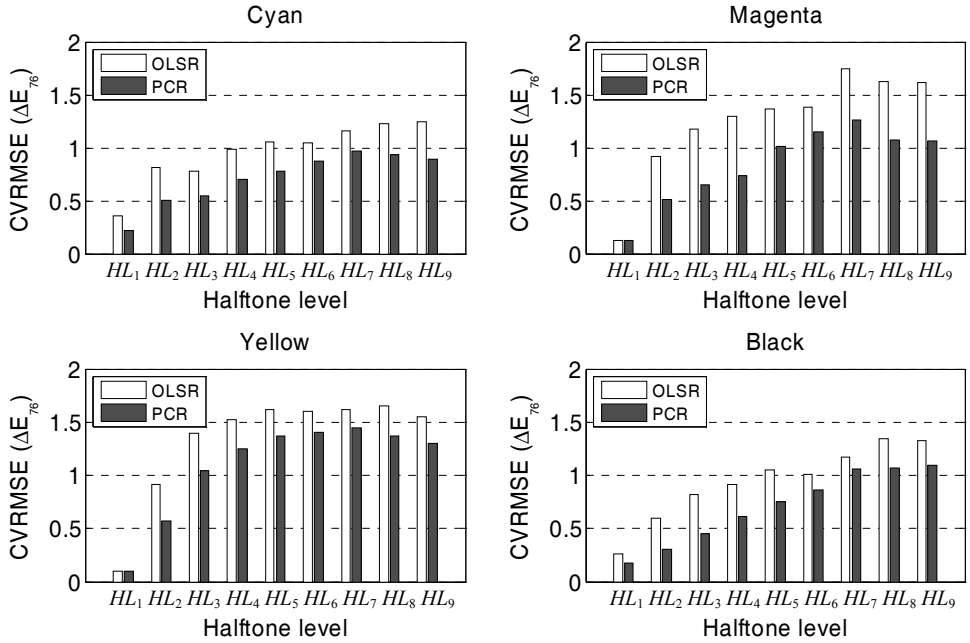

Figure 1. *Cross-validation root-mean-squared errors (CVRMSE) of the ordinary least square regression (OLSR) models and the principal component regression (PCR) models at each halftone level (HL).*