Journal of Imaging Science and Technology® 54(5): 050302–050302-9, 2010.

Improving Tone Prediction in Calibration of Electrophotographic Printers by Linear Regression: Using Principal Components to Account for Co-Linearity of Sensor Measurements

Yan-Fu Kuo▲ **and George T.-C. Chiu**▲
School of Mechanical Engineering, Purdue University, West Lafayette, Indiana 47907
E-mail: ykuo@purdue.edu

Chao-Lung Yang and Yuehwern Yih
School of Industrial Engineering, Purdue University, West Lafayette, Indiana 47907

Jan P. Allebach▲
School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47907

Abstract. *This work employs principal component regression (PCR) to improve tone prediction accuracy for color electrophotography (EP). During calibration, primary color patches at different half-tone levels are printed on a belt and measured using on-board sensors. Regression models are developed to predict primary color tone values on output media from these on-board sensor measurements. The prediction accuracy of the regression models directly impacts the quality and consistency of color reproduction. Analyses have revealed a high degree of correlation among the on-board sensor measurements of the calibration patches from the same primary color. This indicates that multiple on-board sensor measurements are linearly correlated and using multiple on-board sensor measurements to predict the tone value may improve prediction accuracy if the collinearity of the measurements is taken into consideration. In this study, a PCR-based approach is applied to handle the multicollinear measurements in estimating the regression model coefficients. Experimental results show the proposed PCR models reduce root-mean-squared error by 24.7% over ordinary least-squares regression models. © 2010 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2010.54.5.050302]

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 throughputs and operating conditions. During a calibration, multiple patches of different half-tone levels of the same primary color are printed on an intermediate media, and are measured with on-board sensors, such as densitometers (see Figure 1). Calibration models are used to predict the primary color tone values on the output media from these on-board sensor measurements. The prediction accuracy of the calibration models directly impacts the performance of the calibration. 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 data collected in printer life tests. In a typical life test, various tasks are performed under specified operating conditions. The results are recorded and analyzed to ensure that the design specification is met and sufficiently reliable performance is attained. During the life test, additional color patches are printed on output media immediately following a calibration. Their tone values, also referred to in this study as output tone values, are measured offline with devices such as a spectrophotometer. Calibration models are then developed as a mapping of the on-board sensor measurements to the output tone values.[1] When a calibration is performed while the product is in use, on-board sensor measurements are taken to predict tone values with the calibration models. Appropriate tone correction is then performed by adjusting bias voltages or modifying the tone correction mapping. Since tone value measurements on the output media are not available to typical customers, it is crucial to ensure the prediction accuracy of the calibration model under different

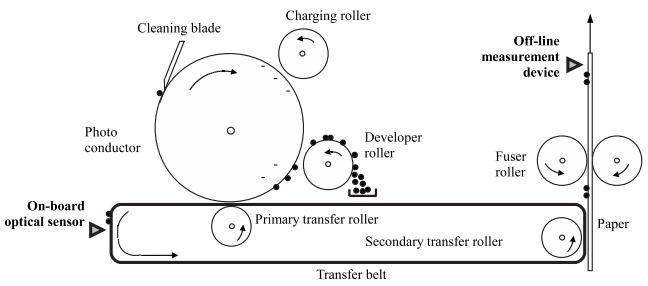

Figure 1. A typical electrophotographic process.

▲IS&T Member.

Received Dec. 23, 2009; accepted for publication Jul. 3, 2010; published online Aug. 16, 2010.
1062-3701/2010/54(5)/050302/9/$20.00.

temperature, humidity and/or other environmental conditions.[2]

Half-toned color images are composed of arrays of closely spaced microdots. Changes in operating conditions or different EP parameter settings will impact the sizes of the microdots. Assuming the impact on the sizes of the microdots is consistent for a given print, the on-board sensor measurements from different half-tone patches of the same color should be consistently higher or lower. This results in increased correlation among the on-board sensor measurements, i.e., multicollinearities. The multicollinearity indicates that multiple on-board sensor measurements are linearly correlated with a tone value and using multiple on-board sensor measurements for tone value prediction can potentially improve the prediction accuracy. However, it is well known that using collinear measurements as explanatory variables to identify model coefficients directly through ordinary least-squares regression (OLSR) will result in suboptimal model coefficients that will degrade prediction accuracy.[3] Hence, existing calibration models are developed using a single-response regression approach, i.e., the output tone value at a particular half-tone level is regressed only with the on-board sensor measurement at the same half-tone level.

Recent research in regression analysis has shown improved prediction accuracy of regression models using multiple explanatory variables as compared to single-response regression models.[4–6] In this study, a principal component regression (PCR) approach[7] is proposed to address the multicollinearities associated with multiple on-board sensor measurements. PCR avoids the numerical issues associated with 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 PCs to be retained while discarding PCs that have less statistical significance. By properly selecting explanatory variables and the associated PCs, a more accurate calibration model can be developed. 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 a 24.7% improvement in prediction accuracy compared with the existing OLSR calibration models for a particular target color EP laser printer model.

The organization of this article is outlined as follows. In the next section, problem formulation and PCR methodology are described. Then a case study with the proposed method and its experimental validation through statistical analyses is illustrated. Concluding remarks are given in the last section.

METHOD

Calibration Model

Since each primary color is printed independently for an in-line color EP process, a calibration model is developed for each primary color. A calibration model $\boldsymbol{G}$ can be written as $\boldsymbol{y}=\boldsymbol{G}(\boldsymbol{w},\boldsymbol{d})$, where $\boldsymbol{y}$ is tone values on paper, $\boldsymbol{w}$ is sensor measurements from on-board densitometers, and $\boldsymbol{d}$ is uncontrollable but measurable factors/disturbances collected in life test, such as temperature, humidity, and throughput.[8] The tone values $\boldsymbol{y}$ are the measured reflectances of the reproduced color patches printed at the designated half-tone levels. In this study, a tone value is defined as the Euclidian distance (ΔE) in CIE L*a*b* space[9] between the color point of a primary color printed at a particular half-tone level and the substrate color. A static linear calibration model is assumed.

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 $w_{ij}\in\Re$, $y_{ij}\in\Re$, and $d_{ij}\in\Re$ as the jth on-board sensor measurement, the jth tone value measurement, and the jth measurable disturbances, respectively, in the ith observation. In this work, the calibration model $\boldsymbol{G}$ is formulated as a linear transformation relating the tone value measurements $\boldsymbol{y}_i=[y_{i1}y_{i2}\ldots y_{ij}]$ to the sensor measurements $\boldsymbol{w}_i=[w_{i1}w_{i2}\ldots w_{ij}]$ and the disturbances $\boldsymbol{d}_i=[d_{i1}d_{i2}\ldots d_{ij}]$.

Consider $p\in\boldsymbol{N}$ on-board sensor measurements, $q\in\boldsymbol{N}$ measurable disturbances, and $l\in\boldsymbol{N}$ tone value measurements are made in one observation, and $n\in\boldsymbol{N}$ observations are gathered. Denote $\boldsymbol{W}=[w_{ij}]\in\Re^{n\times p}$ as the sensor measurement matrix and $\boldsymbol{D}=[d_{ij}]\in\Re^{n\times q}$ as the measurable disturbance matrix. Let $\boldsymbol{X}\in\Re^{n\times r}$ denote the explanatory variable matrix, which is a concatenation of matrices $\boldsymbol{W}$ and $\boldsymbol{D}$, i.e., $\boldsymbol{X}=[\boldsymbol{W}|\boldsymbol{D}]$ and $r=p+q$. Denote $\boldsymbol{Y}=[y_{ij}]\in\Re^{n\times l}$ as the response variable matrix containing the tone value measurements. Note that here the upper case letters represent the concatenation of measurements from n observations, e.g.,

$$\boldsymbol{D}=\begin{bmatrix}\boldsymbol{d}_1\\ \boldsymbol{d}_2\\ \vdots\\ \boldsymbol{d}_n\end{bmatrix}. \tag{1}$$

The calibration model $\boldsymbol{G}\in\Re^{r\times l}$ can be written as $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{G}$. Note that the matrices are assumed to be centered and standardized columnwise.[10] Hence no intercept term is required in the regression model development.

Ordinary Least-Squares Regression

Consider a standard multivariate regression model,

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{G}+\boldsymbol{E}, \tag{2}$$

where the error matrix $\boldsymbol{E}$ satisfies the usual assumption of being independent and identically distributed. The number of observations typically is much more than the number of calibration color patches printed in a calibration, i.e., $n\gg r$. The OLSR solution to the overdetermined problem stated above minimizes the squared error, i.e.,

$$\boldsymbol{G} = \arg\min\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{G}\|^2 = \arg\min\|\boldsymbol{Y} - [\boldsymbol{W}|\boldsymbol{D}]\boldsymbol{G}\|^2 = \arg\min\left\|\boldsymbol{Y} - [\boldsymbol{W}|\boldsymbol{D}]\begin{bmatrix}\boldsymbol{G}_W \\ \boldsymbol{G}_D\end{bmatrix}\right\|^2, \tag{3}$$

where the calibration model $\boldsymbol{G}$ can be split into two matrices $\boldsymbol{G}_W$ and $\boldsymbol{G}_D$ with proper dimensions corresponding to the sensor measurement matrix $\boldsymbol{W}$ and the disturbance measurement matrix $\boldsymbol{D}$, respectively. The OLSR solution to Eq. (3) is given by

$$\boldsymbol{G} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{4}$$

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 computationally challenging especially for matrices with lower conditioning number. This yields larger variance in the model coefficient estimation.

Principal Component Regression (PCR)

The key idea of PCR is to linearly transform the multicollinear sensor measurement matrix $\boldsymbol{W}$ to a principal component (PC) matrix that consists of a set of orthogonal vectors. Then the model coefficient estimation can be directly carried out following Eq. (4). Note that the disturbance measurement matrix $\boldsymbol{D}$ does not need to be included in the transformation since the disturbances should adequately span the entire dynamic range for a complete experimental design. Hence, the sensor measurement matrix $\boldsymbol{D}$ should be associated with minimum multicollinearity.

A singular value decomposition (SVD) on the sensor measurement matrix $\boldsymbol{W}$ is performed as the first step to calculate the PC matrix, i.e.,

$$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \sum_{i=1}^{p}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T, \tag{5}$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \Re^{n\times p}$ is a diagonal matrix of singular values σ_i associated with the ith 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 the 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 $\boldsymbol{W}$ with the right unitary matrix $\boldsymbol{V}$, i.e., $\boldsymbol{\Psi} = \boldsymbol{W}\boldsymbol{V}$. Hence the principal components PC_i are linear combination of the raw sensor measurements with the coefficients in the associated row vector $\boldsymbol{v}_i$. Note that the singular values are usually arranged in descending order, i.e., $\sigma_1 > \sigma_2 > \ldots > \sigma_p$. The magnitude of a singular value represents the variance along the direction of the associated PC. The fraction of the total variance accounted for by PC_i can be calculated by dividing the associated singular value σ_i by the sum of singular values, i.e., $\sigma_i/\Sigma\sigma_i$.

Next, the PC matrix is augmented with the disturbance measurement matrix as the explanatory variable matrix in the subsequent multivariate regression, i.e., $\boldsymbol{X} = [\boldsymbol{\Psi}|\boldsymbol{D}]$. The resulting coefficient matrix can be obtained by solving a standard least-squares optimization problem,

$$\boldsymbol{\Gamma} = \arg\min\|\boldsymbol{Y} - [\boldsymbol{\Psi}|\boldsymbol{D}]\boldsymbol{\Gamma}\|^2 = \arg\min\left\|\boldsymbol{Y} - [\boldsymbol{\Psi}|\boldsymbol{D}]\begin{bmatrix}\boldsymbol{\Gamma}_\Psi \\ \boldsymbol{\Gamma}_D\end{bmatrix}\right\|^2, \tag{6}$$

where $\boldsymbol{\Gamma} \in \Re^{r\times l}$ is the coefficient matrix to be determined. The coefficient matrix $\boldsymbol{\Gamma}$ can be split into two matrices $\boldsymbol{\Gamma}_\Psi$ 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. (6) can be carried out directly following Eq. (4). Matching the response variable matrix $\boldsymbol{Y}$ in Eqs. (3) and (6), one can obtain

$$\boldsymbol{Y} = [\boldsymbol{\Psi}|\boldsymbol{D}]\begin{bmatrix}\boldsymbol{\Gamma}_\Psi \\ \boldsymbol{\Gamma}_D\end{bmatrix} = \boldsymbol{\Psi}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Gamma}_\Psi + \boldsymbol{D}\boldsymbol{\Gamma}_D = \boldsymbol{W}\boldsymbol{G}_W + \boldsymbol{D}\boldsymbol{G}_D = [\boldsymbol{W}|\boldsymbol{D}]\begin{bmatrix}\boldsymbol{G}_W \\ \boldsymbol{G}_D\end{bmatrix} = \boldsymbol{Y}. \tag{7}$$

The calibration model is written as

$$\boldsymbol{G} = \begin{bmatrix}\boldsymbol{G}_W \\ \boldsymbol{G}_D\end{bmatrix} = \begin{bmatrix}\boldsymbol{V}\boldsymbol{\Gamma}_\Psi \\ \boldsymbol{\Gamma}_D\end{bmatrix}. \tag{8}$$

Biased Principal Component Regression

Noise in sensor measurements can result in bias in regression analysis and increase the uncertainty in model coefficient estimation. Biased PCR identifies PCs that do not improve prediction accuracy and excludes them from being used in the regression. Assume the noise in the sensor measurement matrix $\boldsymbol{W}$ is additive. The sensor measurement matrix can be then decomposed into two matrices—an exact signal matrix $\boldsymbol{S}$ and a noise perturbation matrix $\boldsymbol{N}$—so that

$$\boldsymbol{W} = \boldsymbol{S} + \boldsymbol{N} = \boldsymbol{U}_S\boldsymbol{\Sigma}_S\boldsymbol{V}_S^T + \boldsymbol{U}_N\boldsymbol{\Sigma}_N\boldsymbol{V}_N^T, \tag{9}$$

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 unitary matrix, and right unitary matrix from the SVD of the signal matrix $\boldsymbol{S}$ and the noise perturbation matrix $\boldsymbol{N}$, respectively. If a principal component PC_i does not improve model performance based on a set of predetermined criteria, the corresponding singular value σ_i, left singular vectors $\boldsymbol{u}_i$, and right singular vectors $\boldsymbol{v}_i$ are put to the noise perturbation matrix $\boldsymbol{N}$. The biased PC matrix, $\boldsymbol{\Psi}_S = \boldsymbol{W}\boldsymbol{V}_S$, is used in the subsequent regression. Note that vectors in the PC matrix $\boldsymbol{\Psi}$ are orthogonal. Partial regression coefficients and the rank of marginal statistics remain stable when adding or removing PCs in the regression.[11]

Forward Selection

A forward selection algorithm is used to determine the PCs to be included in the regression. PC selection in the PCR is addressed by several studies in the literature.[12–15] Some studies have pointed out that PCs associated with small singular values may be well correlated with the response variables.[16–18] Instead of using traditional top-down selection methods, this study utilizes the forward selection method proposed by Xie and Kalivas.[19] The forward selec-

tion tries out the PCs one by one and includes one PC in the model if it is statistically significant to the response variables. The Bayesian information criterion (BIC),[20]

$$\mathrm{BIC} = n \ln\left(\frac{\mathrm{RSS}}{n}\right) + k \ln(n), \tag{10}$$

is used as the selection criterion, where n is the number of observations, RSS is the residual sum of squares from the estimated model, and k is the number of PCs to be included in the forward selection. The BIC is a tradeoff between model accuracy, i.e., the residual sum of squares (RSS), and model complexity (k), i.e., the number of PCs to be included. Ideally a model with a low BIC value is preferred. BIC is known to be more conservative compared to other information criteria.[21] Hence, the chance of overfitting can be reduced by using BIC as the selection criterion. The PC selection procedure can be summarized in the following four steps:

Step 1: Compute all of the PCs through SVD.

Step 2: Determine the first PC producing the minimum selection criterion by following Eq. (10). Call this the first PC subset.

Step 3: Identify the second PC subset as the subset of PCs 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. (10).

Step 4: The process stops when the selection criterion of the second subset is larger than that of the first subset or when all PCs are included in the regression. Otherwise, replace the contents of the first subset by the contents of the second subset and continue from step 3.

The PC selection should be performed separately for each response variable. Each response variable is regressed with its own set of selected PCs to generate a set of model coefficients. The calibration model is the concatenation of the model coefficients for each response variable. The signal matrix of the selected PCs for the mth response variable can be expressed as

$$\boldsymbol{S}^{(m)} = \boldsymbol{U}_S^{(m)} \boldsymbol{\Sigma}_S^{(m)} (\boldsymbol{V}_S^{(m)})^T. \tag{11}$$

The biased PC matrix of the mth response variable $\boldsymbol{\Psi}_S^{(m)}$ can be obtained by multiplying the sensor measurement matrix $\boldsymbol{W}$ with the right unitary matrix $\boldsymbol{V}_S^{(m)}$ from Eq. (11), i.e., $\boldsymbol{\Psi}_S^{(m)} = \boldsymbol{W}\boldsymbol{V}_S^{(m)}$. Let $\boldsymbol{y}^{(m)} \in \Re^n$ denote the mth column vector in the response variable matrix $\boldsymbol{Y}$. The coefficient vector that minimizes a least-squares loss function for the mth response variable can be obtained as

$$\begin{aligned}\boldsymbol{\gamma}^{(m)} &= \arg\min \|\boldsymbol{y}^{(m)} - [\boldsymbol{\Psi}_S^{(m)} | \boldsymbol{D}] \boldsymbol{\gamma}^{(m)}\|^2 \\ &= \arg\min \left\| \boldsymbol{y}^{(m)} - [\boldsymbol{\Psi}_S^{(m)} | \boldsymbol{D}] \begin{bmatrix} \boldsymbol{\gamma}_S^{(m)} \\ \boldsymbol{\gamma}_D^{(m)} \end{bmatrix} \right\|^2, \end{aligned} \tag{12}$$

where $\boldsymbol{\gamma}^{(m)} \in \Re^{r \times 1}$ 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 solution of the coefficient vector $\boldsymbol{\gamma}^{(m)}$ can be carried out directly following Eq. (4). The calibration model can be obtained by concatenating the product vectors obtained by multiplying the coefficient vectors $\boldsymbol{\gamma}_S^{(m)}$ from Eq. (12) and the associated right unitary matrix $\boldsymbol{V}_S^{(m)}$ from Eq. (11), with the coefficient vectors $\boldsymbol{\gamma}_D^{(m)}$, i.e.,

$$\boldsymbol{G} = \begin{bmatrix} [\boldsymbol{V}_S^{(1)} \boldsymbol{\gamma}_S^{(1)} | \cdots | \boldsymbol{V}_S^{(l)} \boldsymbol{\gamma}_S^{(l)}] \\ [\boldsymbol{\gamma}_D^{(1)} | \cdots | \boldsymbol{\gamma}_D^{(l)}] \end{bmatrix}. \tag{13}$$

Remark. The proposed forward selection algorithm can be applied to determine the disturbances to be included in the regression. Once the optimal PC subset for a tone value is obtained, the algorithm can be used to check whether including any of the disturbances can improve the prediction accuracy.

EXPERIMENT

Experiment Setup

An off-the-shelf one-pass color EP laser printer model is used in the experiment. The printer generates nine calibration patches at different half-tone levels for each primary color during a calibration, i.e., $p=9$. These half-tone levels are labeled as h_j, where $j=1\ldots9$, corresponding to gray values from light to dark. Patches identical to those printed in the calibration are printed on 75 $\mathrm{g/m^2}$ paper (Xerox® 4200 Business) for each primary color immediately following a calibration. Their tone value measurements are made with a set of spectrophotometers (X-Rite® DTP-70) with D65 illuminant and 2° observer. Note that the D65 illumination and ΔE_{76} metrics are adopted in this work to meet the sponsor's requirements and specification. Changing the metrics is not likely to impact the validity of the improvement introduced by the work.

Experiment

The experiment is performed on 20 printers across a wide range of environmental conditions. The temperature ranges from 15 to 30°C, and the relative humidity ranges from 10% to 80%. Several cartridge sets with various lives remaining are used. A total of 419 observations are made for each primary color. Temperature, humidity ratio, and cartridge life remaining are measured and treated as measurable disturbances. The models are identified following the proposed PCR procedure using MATLAB®.

Multicollinearity of the Sensor Measurements

Variance inflation factor[20] (VIF) is commonly used to measure the severity of multicollinearity among explanatory variables. It is defined as

$$(VIF)_j = \frac{1}{1 - R_j^2}, \tag{14}$$

where R_j^2 is the unadjusted coefficient of determination of the jth explanatory variable when it is regressed with the

Table I. Variance inflation factor values of the sensor measurements at each half-tone level (h_j).

	Cyan	Magenta	Yellow	Black
h_1	1.3	1.2	1.3	1.3
h_2	3.9	2.6	4.6	2.2
h_3	7.4	5.3	7.6	4.7
h_4	8.8	7.2	22.5	8.6
h_5	11.6	13.8	26.3	10.7
h_6	17.9	14.6	27.0	14.9
h_7	19.9	16.4	53.0	23.7
h_8	18.9	16.2	31.9	15.8
h_9	11.2	7.6	14.1	12.9

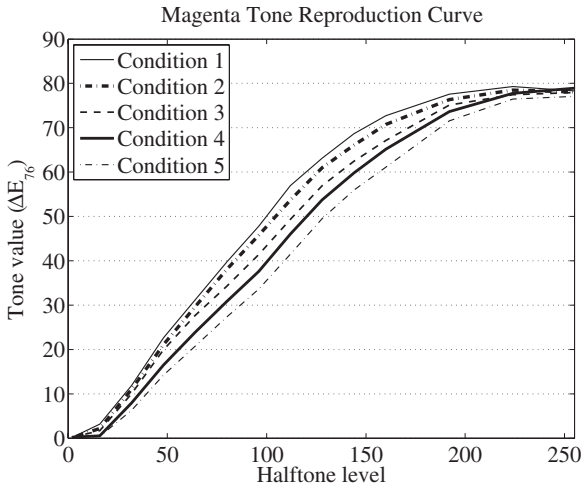

Figure 2. Magenta tone reproduction curves.

other explanatory variables. If the jth explanatory variable is linearly correlated with any of the other explanatory variables in the model, the corresponding R_j^2 and VIF value will be large. VIF values that exceed 10 are often regarded as indicating strong multicollinearity[22] among the explanatory variables, implying that ordinary least-squares regression may not be a good approach.

Table I lists the VIF values of the experimental on-board sensor measurements. It is shown that more than 50% of the sensor measurements are associated with a high degree of multicollinearity, particularly those sensor measurements associated with half-tone levels in the midtone range (half-tone levels h_5 to h_8). The large VIF values also confirm the consistently higher or lower tone values across half-tone levels due to changes in operating conditions (see magenta TRCs illustrated in Figure 2).

Singular Value Decomposition on Sensor Measurements

Singular value decomposition is performed on the sensor measurement matrix $\boldsymbol{W}$. Table II shows the contribution toward total variation in percentage by each PC. PC_1 alone accounts for at least 85% of the total variation for all primary colors. The considerable contribution of PC_1 corresponds with the high degree of multicollinearity shown by the large VIF values.

Table II. Contribution to total variation in percentage by each principal component (PC).

	Cyan (%)	Magenta (%)	Yellow (%)	Black (%)
PC_1	87.0	85.6	93.6	86.8
PC_2	4.5	4.0	2.2	3.9
PC_3	3.0	3.8	1.4	2.7
PC_4	2.0	2.5	0.8	2.4
PC_5	1.2	1.2	0.8	1.5
PC_6	1.0	1.1	0.5	1.0
PC_7	0.6	0.8	0.3	0.7
PC_8	0.5	0.6	0.3	0.5
PC_9	0.4	0.4	0.2	0.4

Remark. Each principal component is a linear combination of different sensor measurements. The large contribution associated with PC_1 indicates that all the sensor measurements from different half-tone levels of the same primary color contain a large share of information in common, which is most likely to be the degree of size fluctuation of the half-tone microdots.

Principal Component Selection

The proposed forward PC selection is performed using the experimental data. Table III lists the selected PCs at each half-tone level h_j. As expected, PC_1 is always selected and is always the most significant PC. However, PC_2, which accounts for the second largest variance, is not always the second significant PC (see magenta and black). In addition, including PC_4, PC_5, or PC_6 improves the model prediction accuracy at certain half-tone levels, in spite of the fact that their contributions to the total variance are small. The inclusion of PC_4, PC_5, or PC_6 suggests that they may contain important information regarding the local tone value variation. These facts indicate that using a conventional top-down selection procedure to determine the optimal set of PCs may not be appropriate for this particular application.

Disturbance Selection

The proposed forward selection algorithm is also applied to determine the disturbances to be included in the regression. The results show that only the humidity ratio is of statistical significance.

Remark. The exclusion of the temperature or the cartridge life remaining suggests that either their impact on tone value variation is adequately captured by the on-board sensor measurements, or they are not well correlated with the variation of the tone values. Hence including them as explanatory variables in the calibration model does not improve the prediction accuracy.

Model Comparison

The calibration models generated by the proposed PCR methods, also referred to in this study as PCR models, are

Table III. Forward selection chosen principal components (PC) at each half-tone level (h_j).

	Cyan		Magenta
h_1	PC_1, PC_2, PC_6	h_1	PC_1
h_2	PC_1, PC_2	h_2	PC_1, PC_3, PC_5, PC_6
h_3	PC_1, PC_2, PC_3	h_3	PC_1, PC_3
h_4	PC_1, PC_2, PC_3	h_4	PC_1, PC_3
h_5	PC_1, PC_2, PC_3	h_5	PC_1, PC_3
h_6	PC_1, PC_2, PC_3	h_6	PC_1, PC_3, PC_2
h_7	PC_1, PC_2	h_7	PC_1, PC_3
h_8	PC_1, PC_2	h_8	PC_1, PC_3
h_9	PC_1, PC_2	h_9	PC_1, PC_3
	Yellow		**Black**
h_1	PC_1	h_1	PC_1, PC_4, PC_6
h_2	PC_1, PC_2, PC_6, PC_5, PC_4	h_2	PC_1, PC_4, PC_6
h_3	PC_1, PC_2, PC_6, PC_5, PC_4	h_3	PC_1, PC_4, PC_3
h_4	PC_1, PC_2	h_4	PC_1, PC_4, PC_3
h_5	PC_1, PC_2	h_5	PC_1, PC_4, PC_3
h_6	PC_1, PC_2	h_6	PC_1, PC_4
h_7	PC_1, PC_2, PC_6	h_7	PC_1
h_8	PC_1, PC_2	h_8	PC_1
h_9	PC_1, PC_2	h_9	PC_1

compared with the models generated by ordinary least-squares regression methods, referred to as OLSR models, proposed by Yang et al.[8] In each of the OLSR models, a single sensor measurement w_j and selected disturbances d_{ij} are used as explanatory variables to predict the tone value y_j at each half-tone level, i.e.,

$$y_j = \alpha_j w_j + \sum_i \beta_{ij} d_{ij}, \tag{15}$$

where α_j and β_{ij} are model coefficients to be determined in the OLSR. Note that a major difference between the two types of models is the number of on-board sensor measurements included as explanatory variables. A PCR model includes multiple on-board sensor measurements as explanatory variables. In contrast, an OLSR model includes only one. Performance indices are used to compare the two models in the following sections.

Cross-validation

A tenfold cross-validation without replacement is performed on the PCR model with the proposed forward selection, the PCR model with conventional top-down selection,[23] and the OLSR models. In the conventional top-down selection, only the first two PCs are included as explanatory variables due to the small contribution to total variance of the remaining PCs (see Table II). The results of the cross-validation (CV) are summarized by comparing the root-mean-squared errors (RMSE) for the three models (see Figure 3). It is shown that overall the PCR model with forward selection gives the least CV RMSEs. This indicates that the forward selection is superior in this application, particularly for magenta and

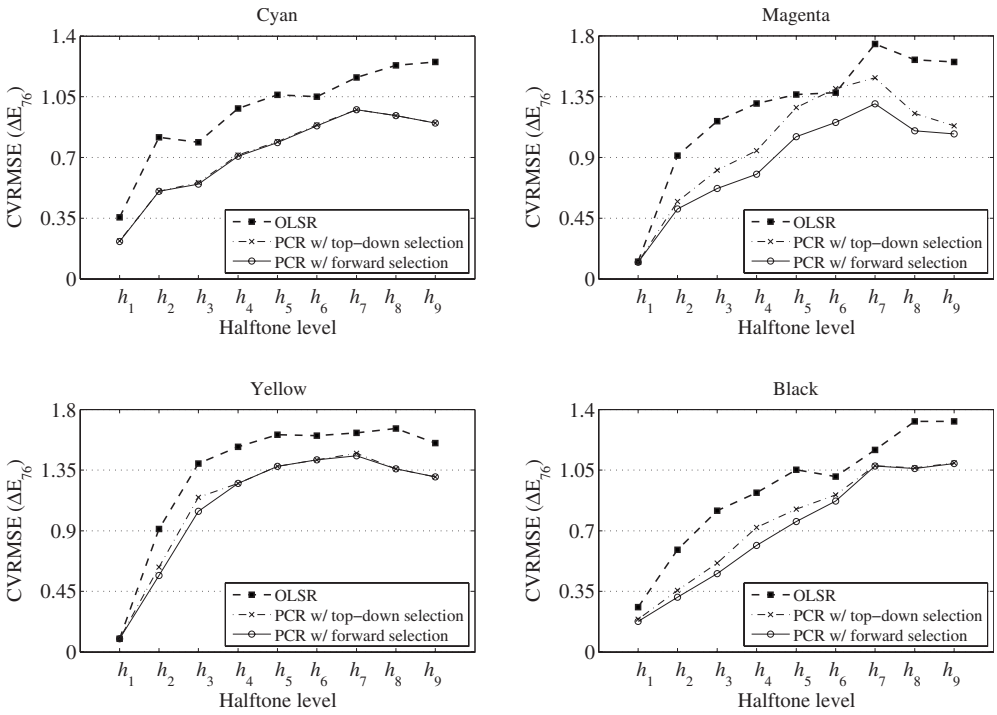

Figure 3. Cross-validation root-mean-squared error (CVRMSE) of the ordinary least-squares regression (OLSR) models, the principal component regression (PCR) models with top-down selection, and the PCR models with forward selection at each half-tone level h_j.

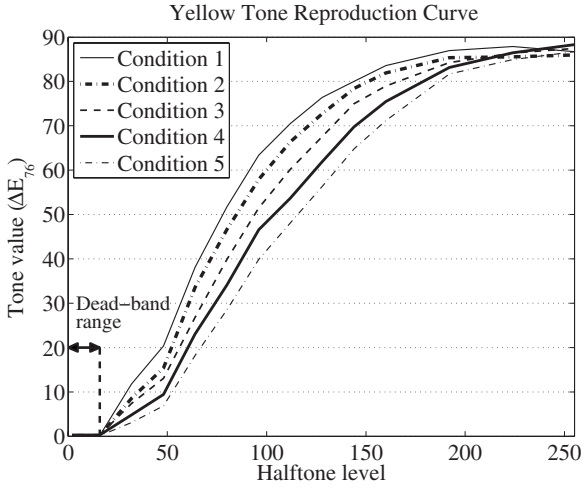

Figure 4. Yellow tone reproduction curves.

black. This is because PC_3 and PC_4 are significant to magenta and black, respectively, in predicting their tone values (see Table II). Conventional top-down selection can ignore these significant PCs and results in suboptimal prediction accuracy.

The average percentage improvement of the PCR models with forward selection over the OLSR models is calculated by

$$Improvement = \frac{e_{OLSR}^{Total} - e_{PCR}^{Total}}{e_{OLSR}^{Total}} \times 100\%, \qquad (16)$$

where e_{OLSR}^{Total} and e_{PCR}^{Total} are the total CV RMSE of the OLSR and the PCR models with forward selection, respectively, for all colorants and at all half-tone levels. They are 40.37 and 30.38 ΔE_{76} units, respectively. The average CV RMSE of the PCR and the OLSR models are 1.12 and 0.84 ΔE_{76} units, respectively. The average improvement of the PCR models is 24.7%.

Note that the PCR models for magenta and yellow at the first half-tone level h_1 do not show significant improvement. This is due that the fact that the first half-tone level h_1 is in the dead-band range of the half-tone level where almost no tone reproduction occurs. The dead-band range for yellow is particularly large (see Figure 4). The mean tone values for yellow and magenta at the first half-tone level h_1 are 0.2 ΔE_{76} units. These tone values are beyond the dynamic range of the on-board sensors.

Statistical partial F-test

The cross-validation study shows that PCR models can provide better prediction accuracy compared to the OLSR models. However, the PCR models include more on-board sensor measurements as explanatory variables. Complex models could provide a better fit without necessarily bearing any interpretable relationship to the underlying process.

Statistical partial F-tests[20] are conducted to determine if the improvements yielded by the PCR models are not due to higher model complexity. The resulting p-values (see Table IV) show that the improvements yielded by the PCR models are significant at a 99% confidence level for almost all colors at all half-tone levels. Here the 99% confidence level corresponds to a threshold of 0.01 for the significance test of improvement. The only exception is the model of yellow at half-tone level h_1.

Table IV. p-values of partial F-test to determine the significance of improvement of the PCR models.

	Cyan	Magenta	Yellow	Black
h_1	$<10^{-9}$	0.0004	0.8978	$<10^{-9}$
h_2	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$
h_3	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$
h_4	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$
h_5	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$
h_6	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$
h_7	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$
h_8	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$
h_9	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$	$<10^{-9}$

Model selection index

Two information-theoretic model selection indices are used to compare the two models: the Bayesian information criterion (BIC), see Eq. (10), and the Akaike information criterion (AIC),[20]

$$\mathrm{AIC} = n \ln\left(\frac{\mathrm{RSS}}{n}\right) + 2\ln(n), \qquad (17)$$

where n is the number of observations and RSS is the residual sum of squares from the regression model. These model selection indices intend to identify the best model as a tradeoff between model accuracy and model complexity. Typically, a model is preferred if it is associated with a smaller index value.

Figure 5 shows the AIC and the BIC values of the two models at each half-tone level. Overall the PCR models are associated with smaller AIC and BIC values, despite the fact that the PCR models are associated with a higher level of complexity. The AIC or the BIC values of the PCR models at half-tone level h_1 for yellow or magenta are slightly higher than those of the OLSR models due to the fact that the half-tone level h_1 is in the dead-band of the tone reproduction for these two colors, as can be seen from Figs. 2 and 4.

Remark. The numerical uncertainty of the model coefficient estimation in regression caused by the multicollinear on-board sensor measurements may be solved with improved computational accuracy. Indeed, with double precision, most calibration models may be developed directly through OLSR without considering multicollinearity. However, the proposed PCR method provides two advantages that are not achievable by using OLSR. First, biased PCR can reduce the chance that the calibration model is overfitted by excluding insignificant PCs from being used in the regression. In addition, the principal components can provide in-

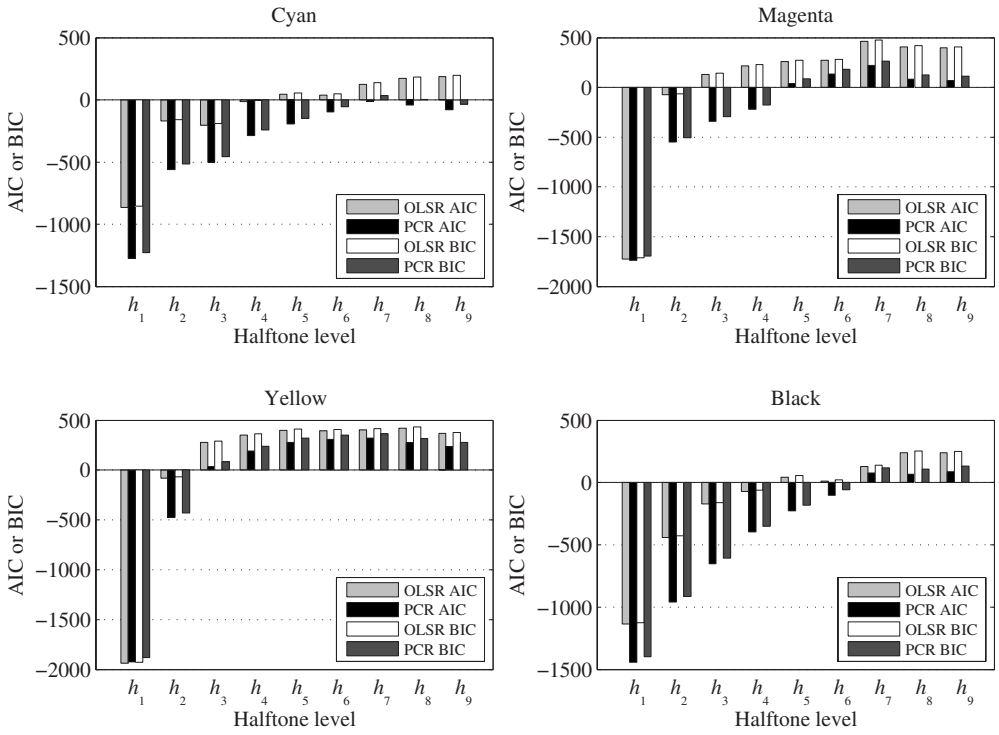

Figure 5. The Akaike information criterion (AIC) values and the Bayesian information criterion (BIC) values of the ordinary least-squares regression (OLSR) models and the principal component regression (PCR) models at each half-tone level h_i.

sight into the importance of the on-board sensor measurements. For example, PC_1 of cyan is

$$\begin{aligned} PC_1 = {} & 0.0323w_1 - 0.0002w_2 - 0.1562w_3 + 0.2999w_4 \\ & + 0.6528w_5 - 0.3603w_6 + 0.2437w_7 - 0.5180w_8 \\ & - 0.0308w_9 \end{aligned} \tag{18}$$

The coefficients of the equation indicate that the normalized on-board sensor measurements associated with half-tone levels in the midtone range (w_4 through w_8) are more strongly correlated with PC_1, compared to those associated with half-tone levels in the highlight or shadow region. PC_1 is the most significant PC to all models at different half-tone levels (see Table III). This suggests that using an on-board sensor measurement associated with a half-tone level in the midtone range, e.g., h_5, can more accurately predict a tone value associated with a half-tone level in the highlight range, e.g., h_1 (see correlation coefficients between the on-board sensor measurements and the tone values shown in Table V for verification). Note that principal components have been previously used to interpret underlying physical processes.[24,25] Following the PCR procedure, the current work may also be extended in further research to investigate the correlation between the PCs and the physical characteristics, such as developability or transfer efficiency, of an EP system.

Table V. Correlation coefficients between the on-board sensor measurements and the tone values associated with half-tone level h_1 in highlight range and h_5 in the midtone range.

		On-board sensor measurement (w_i)	
Correlation coefficient		h_1	h_5
Tone value (y_i)	h_1	0.319	0.758
	h_5	0.297	0.886

CONCLUSION

A PCR method is proposed to improve tone prediction accuracy of calibration models for color EP systems. 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 PCs 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 and model selection indexes demonstrate that the proposed PCR models outperform existing OLSR models. The PCR models provide 24.7% improvement on average in root-mean-squared predication accuracy over existing models based on cross-validation.

ACKNOWLEDGMENTS

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

REFERENCES

[1] D. A. Johnson, U.S. Patent No. 6,982,812 (2006).
[2] A. S. Diamond, in *The Handbook of Imaging Materials*, 2nd ed. (Marcel Dekker, New York, NY, 2002).
[3] D. C. Montgomery and D. J. Friedman, "Prediction using regression models with multicollinear predictor variables", IIE Trans. **25**, 73 (1993).
[4] L. Breiman and J. H. Friedman, "Predicting multivariate responses in multiple linear regression", J. R. Stat. Soc. Ser. B (Methodol.) **59**, 3 (1997).
[5] R. D. De Veaux and L. H. Ungar, "Multicollinearity: a tale of two nonparametric regressions", *Selecting Models from Data: AI and Statistics* (Springer-Verlag, New York, NY, 1994), pp. 293–302.
[6] R. Bro, "Multivariate calibration: what is in chemometrics for the analytical chemist", Anal. Chim. Acta **500**, 185 (2003).
[7] W. F. Massy, "Principal components regression in exploratory statistical research", J. Am. Stat. Assoc. **60**, 234 (1965).
[8] C. Yang, Y. Kuo, Y. Yih, G. T. Chiu, D. A. Abramsohn, G. R. Ashton, and J. P. Allebach, "Improving tone prediction accuracy in calibration for color electrophotography part I—environmental and consumable factors", J. Imaging Sci. Technol. **54** 050301 (2010)
[9] CIE Recommendations on Uniform Color Spaces, Color Difference Equations, and Psychometric Color Terms, Supplement No. 2 to CIE Publication No.15, Colorimetry (E.-1.3.1) 1971, (Bureau Central de la CIE, Paris, 1978).
[10] K. Varmuza and P. Filzmoser, in *Introduction to Multivariate Statistical Analysis in Chemometrics*, 1st ed. (CRC Press, Boca Raton, FL, 2009).
[11] M. H. Graham, "Confronting multicollinearity in ecological multiple regression", Ecology **84**, 2809 (2003).
[12] K. Konstantinides, B. Natarajan, and G. Yovanof, "Noise estimation and filtering using block-based singular value decomposition", IEEE Trans. Image Process. **6**, 479 (1997).
[13] I. T. Jolliffe, "Discarding variables in a principal component analysis, I: artificial data", Appl. Stat. **21**, 160 (1972).
[14] E. R. Mansfield, J. T. Webster, and R. F. Gunst, "An analytic variable selection technique for principal component regression", Appl. Stat. **26**, 34 (1977).
[15] S. Boneh and G. R. Mendieta, "Variable selection in regression models using principal components", Commun. Stat: Theory Meth. **23**, 197 (1994).
[16] A. S. Hadi and R. F. Ling, "Some cautionary notes on the use of principal components regression", Am. Stat. **52**, 15 (1998).
[17] J. Sun, "A correlation principal component regression analysis of NIR data", J. Chemom. **9**, 21 (1995).
[18] J. M. Sutter and J. H. Kalivas, "Which principal components to utilize for principal component regression", J. Chemom. **6**, 217 (1992).
[19] Y. Xie and J. Kalivas, "Evaluation of principal component selection methods to form a global prediction model by principal component regression", Anal. Chim. Acta **348**, 19 (1997).
[20] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li, in *Applied Linear Regression Models*, 4th ed. (Irwin McGraw-Hill, New York, NY, 2004).
[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (Springer-Verlag, New York, NY, 2001).
[22] P. J. Curran, S. G. West, and J. F. Finch, "Model-dependent variance inflation factor cutoff values", Qual. Eng. **14**, 391 (2002).
[23] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. (Springer-Verlag, New York, NY, 2002).
[24] K. Sakaguchi, T. Tachibana, S. Furukawa, T. Katsura, K. Yamazaki, H. Kawaguchi, A. Maki, and E. Okada, "Experimental prediction of the wavelength-dependent path-length factor for optical intrinsic signal analysis", Appl. Opt. **46**, 2769 (2007).
[25] Y. Liu, "Principal component analysis of physical, color, and sensory characteristics of chicken breasts deboned at two, four, six, and twenty-four hours post-mortem", Poult Sci. **83**, 101 (2004).