25

A Data Mining Algorithm for Monitoring PCB Assembly Quality

Feng Zhang
Fairchild Semiconductor South Portland, Maine 04106, USA

1. Introduction

When surface mount technology (SMT) evolves as driven by the continuing miniaturization of electronic components and ever-growing board complexity, in-line defect inspection has become common for ensuring reliable production. For example, as an in-line measurement technique, visual defect metrology is now widely utilized in assessing process capability (Cunninggham & MacKinnon 1998; Rao et al. 1996; Barajas et al. 2003). In discrete printed circuit board (PCB) assembly, the boards within each shift are visually inspected to monitor the variation on operational conditions. Often the visual inspections are performed by automated machines, which utilize sophisticated optical and image processing techniques to detect the defects that lead to the process yield loss.

Literature study on semiconductor industry shows that over 60% of end-of-the-line defects can be traced back to solder paste printing process (Breed 1998; Venkateswaran et al. 1997). Improving the printing process performance is expected to produce reduced rework and lower cost in the downstream stages of PCB assembly by preventing small shifts and twists of components from being defects. Moreover, when components have a large number of pins such as ball grid array (BGA), it is crucial to reduce the variation between the deposits of electronic components after printing so that all joints will be soldered properly (Dempster et al. 1977).

Therefore, inspection systems built in paste printing process should not only detect the defects, but also help the operators identify the underlying root causes of poor yield resulting from inappropriate printing operations, and then develop corrective measures to avoid defective boards (Barajas et al. 2001; Litman 2004). A proper understanding of the patterns of variability among the measured solder paste profile is thus required to facilitate operators adjust the influential stencil printing parameters before a significant damage has occurred. To accommodate such quality control and yield improvement motivations, this paper proposes an effective identification method on root causes of solder paste defects by integrating statistical analysis of solder paste measurements and engineering knowledge of stencil printing process.

Very often, in semiconductor fabrication, the outputs of visual defect inspection constitute a list of binary values. That is, when hundreds of integrated circuits are assembled on a printed circuit board, the inspection machine will indicate each solder joint either good or defective. Classical statistical process control (SPC) techniques have been applied to monitor the process disturbances by charting the percentage of defects per PCB. If the total number

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

of visual defects exceeds a predefined control limit, the identified offending equipment should be tuned up and returned to in-control operational conditions. In PCB assembly, the optimal operational parameters for running the process in control are usually designated by operator's experience, or based on a small sample of measurements in that the cost of replicating massive quantities of PCB for inspection prevents the application of experimental design approaches (Bartholomew & Knott 1999; Gopladrishnan & Srihari 1999). A new diagnosing scheme for identifying the fault pattern present in binary inspection data is addressed in this paper, which is shown to serve as a tool to extract clustered patterns from inspected pastes on PCB and thereby identify corresponding root causes for each cluster of defects. Note that this method does not assume any prior knowledge about the nature of stencil printing faults, or any particular distribution on the size, shape or location of solder joint patterns. In short, this chapter introduces a method for routinely monitoring binary visual inspection data to detect the presence of clustered defects caused by certain assignable causes in stencil printing. As a key aspect of quality control and diagnosing, this root cause identification involves searching for systematic faults that explain the observed variability behavior by incorporating process knowledge.

2. The solder past printing process

A substantial proportion of the defects in PCB assembly occur in solder paste printing. For instance, insufficient solder paste volume may result in solder opens while excess solder paste volume increases the chances of bridging (O'Hara & Lee 1996). Maximizing the uniformity of solder paste profile to reduce subsequent assembly defects is then expected to improve the overall quality of PCB fabrication. On the other hand, the detection and reduction of defects in the earlier stage of SMT manufacturing such as stencil printing also diminishes the cost for other downstream stages (Pan et al. 1999). Therefore, a proper control of stencil printing has become significantly important over the years in yield management. As a major step in SMT manufacturing, stencil printing involves the allocation of adequate amount of solder paste on each component pad. In practice, various potential process factors (e.g., printer alignment, squeegee pressure, printing speed, and separation speed) may impact solder paste printing in achieving high quality. Fig. 1 schematically illustrates the stencil printing operation, where metallic stencil is first placed over a PCB and solder paste is kneaded on one side of the stencil. As shown in Fig. 1(a) and 1(b), the squeegee is pushed over the stencil under predefined pressure and moved to the other side of stencil with specific speed. This procedure makes the solder paste roll to fill the apertures in the stencil and the squeegee blade removes the excess of material, followed by the separation of the stencil from PCB at a slow snap-off speed.

In stencil printing, operation parameters should be adjusted as controllable variables by process engineers. However, such parameter adjustment relies heavily on ad-hoc algorithms or expert knowledge, because the direct printing performance evaluation given visual inspection data is not readily achievable. The lack of an analytical process monitoring mechanism comes from the difficulties in deriving a direct mathematical function between the paste defects and process parameters. Thus, a challenging problem arises on how to utilize the binary inspection information to identify the influential process factors (or systematic causes) that affect solder paste quality. When the sample of inspection data becomes available, as discussed below, a logistic regression model will characterize the

correlation of binary solder paste defects and measured physical profile (e.g., solder paste volume, height, area, etc.), the results of which are then incorporated into a latent variable framework for clustering the systematic causes to explain the variation on solder paste and consequent binary defects.

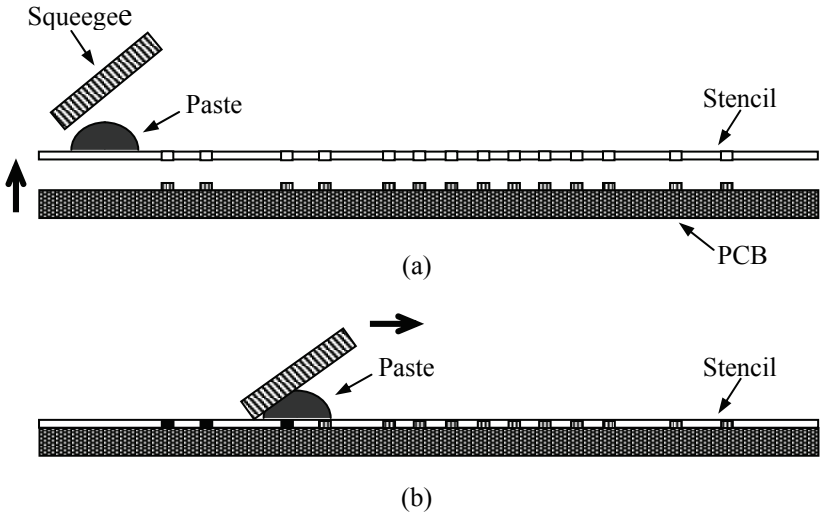

Fig. 1. The schematic illustration of stencil printing process.

3. Logistic regression model

As a common statistical approach for analyzing binary data, logistic regression model has been applied to various data mining and machine learning disciplines such as data classification and predicting the certainty of binary outcome (Bartholomew & Knott 1999; Jaakkola & Jordan 1997; McCulloch 1997). Under the present problem setting, for each solder paste in PCB assembly, let y denote the binary inspection such that 1 for good paste and −1 for failure, and x be a d-dimensional vector representing a set of physical characteristics (called solder paste profile). The logistic regression analysis usually assumes the following quantitative relationship between y and x:

$$p\{y|\boldsymbol{x},\boldsymbol{\beta}\} = \sigma(y\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}) \equiv \frac{1}{1+\exp(-y\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}, \tag{1}$$

where $\boldsymbol{\beta} = [\beta_1\ \beta_2\ \ldots\ \beta_d]^T$ is regression coefficient.
For a set of m measurement couples $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, the log-likelihood of vector $\boldsymbol{\beta}$ in Equation (1) is:

$$L(\boldsymbol{\beta}) = -\sum_{j=1}^{m} \log(1+\exp(-y_j\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j)).$$

It is straightforward to obtain the gradient of log-likelihood function $L(\boldsymbol{\beta})$

$$g = \nabla L(\boldsymbol{\beta}) = \sum_{j=1}^{m} (1 - \sigma(y_j \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_j)) y_j \boldsymbol{x}_j{}'$$

and the second-order Hessian matrix

$$\mathrm{H} = \frac{d^2 L(\boldsymbol{\beta})}{d\boldsymbol{\beta} d\boldsymbol{\beta}^{\mathrm{T}}} = -\sum_{j=1}^{m} \sigma(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_j)(1 - \sigma(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_j)) \boldsymbol{x}_j \boldsymbol{x}_j{}^{\mathrm{T}} . \quad (2)$$

For notation simplicity, the Hessian matrix **H** in Equation (2) is often written in matrix form, i.e., $\mathbf{H} = -\mathbf{X}\mathbf{A}\mathbf{X}^T$, where the non-zero element of diagonal matrix **A** is

$$a_{jj} = \sigma(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_j)(1 - \sigma(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_j)), \; j = 1, 2, \ldots, m,$$

and $\boldsymbol{x}_j$ is the jth column of d×m sample matrix $\mathbf{X} = [\boldsymbol{x}_1 \, \boldsymbol{x}_2 \ldots \boldsymbol{x}_m]$.
Newton optimization algorithm works as an efficient way to estimate the d×1 regression coefficient vector by maximizing $L(\boldsymbol{\beta})$ through the second derivatives (2), which provides the following iterative calculation (3) to estimate $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}_{new} = \boldsymbol{\beta}_{old} + (\mathbf{X}\mathbf{A}\mathbf{X}^{\mathrm{T}})^{-1} \sum_{j=1}^{m} (1 - \sigma(y_j \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_j)) y_j \boldsymbol{x}_j$$

$$= (\mathrm{X}\mathrm{A}\mathrm{X}^{\mathrm{T}})^{-1} \mathrm{X}\mathrm{A}(\mathrm{X}^{\mathrm{T}} \boldsymbol{\beta}_{old} + \left[\frac{y_1}{\sigma(y_1 \boldsymbol{\beta}_{old}^{\mathrm{T}} \boldsymbol{x}_1)} \quad \ldots \quad \frac{y_m}{\sigma(y_m \boldsymbol{\beta}_{old}^{\mathrm{T}} \boldsymbol{x}_m)} \right]^{\mathrm{T}}) . \quad (3)$$

The maximum-likelihood (ML) estimation of $\boldsymbol{\beta}$ is also called iterative re-weighted least squares (IRLS) algorithm, where the computation complexity within each iteration is $O(md^2)$.
As discussed in previous research work, the logistic regression model (1) has been used mostly to understand the role of input variables x in predicting the binary response variable y. In manufacturing practice, however, many of the measurement variables in x are correlated due to some common physical phenomena, which encourage us to seek a parsimonious form of the input variables to summarize their effects on binary outcomes. In other words, the effects on the measured physical profile can be explained by a reduced set of latent variables without loss of statistical information, as described in the next section. Thus, we would refit the regression model (1) with fewer latent variables to provide an interpretation of their influence on binary outputs as observed in defect inspection. This statistical interpretation, equipped with proper pattern clustering and visualization, is shown to enhance the diagnosing of solder paste quality.

4. MLPCA based pattern clustering algorithm

4.1 Latent variable model and MLPCA

When correlations are present among the measured variables x for a product, this implies the existence of common systematic causes that govern such interrelated manners. Therefore, multivariate statistical techniques such as PCA have been proposed to investigate the correlations when multiple variables are involved (Crida et al. 1997). A latent variable model is introduced to relate d characteristics of solder pastes to p unknown systematic causes v, by assuming that v affects the solder paste profile x through a linear model, i.e.,

$$\mathrm{x} = \mathbf{C}\boldsymbol{v} + \mathrm{w}\,, \tag{4}$$

where $\mathbf{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_p]$ is a $d \times p$ constant matrix with full rank, and $\boldsymbol{v} = [v_1, v_2, \ldots, v_p]^{\mathrm{T}}$ is a $p \times 1$ zero-mean random vector with independent components, each scaled without loss of generality to have unit variance.

As assumed in PCA, the latent variables are of smaller dimension (i.e., $p < d$) so that the dependencies among observed data x can be described by a reduced set of variables $\boldsymbol{v}$. Noise w denotes the aggregated effects that are not due to any systematic causes, which is assumed to be white noise, i.e., $\mathrm{w} \sim N(0, \sigma_w^2\mathbf{I})$, and independent of v. It is reasonable to assume that each root cause is associated with distinct physical dynamics so that the latent variables $\boldsymbol{v}$ can be represented by normalized independent Gaussians, that is, $v \sim N(0, \mathbf{I})$. As such, the impacts on measured solder joint profile $\boldsymbol{x}$ from v are quantified by the magnitude of corresponding rows in matrix **C**. Equipped with prior distributions over $\boldsymbol{v}$ and $\boldsymbol{w}$, model (4) now provides a parsimonious probabilistic description for multivariate measurement data $\boldsymbol{x}$ (Hamada & Nelder 1997; Tipping & Bishop 1999). Moreover, the probabilistic assumptions enable an ML estimate for C (denoted by CML) that is shown to span the principal subspaces of $\boldsymbol{x}$ (Tipping & Bishop 1999).

For isotropic Gaussian noise w, model (4) yields the conditional probability of $\boldsymbol{x}$ as:

$$p(\boldsymbol{x}|\boldsymbol{v}) = (2\pi\sigma_w^2)^{-d/2}\exp\left\{-\frac{1}{2\sigma_w^2}\left\|\boldsymbol{x} - \mathbf{C}\boldsymbol{v}\right\|^2\right\}. \tag{5}$$

The Gaussian assumption on v implies that the marginal density of data x can be readily obtained by integrating out v so that $\boldsymbol{x} \sim N(0, \boldsymbol{\Sigma})$, and covariance $\boldsymbol{\Sigma} = \sigma_w^2\mathbf{I} + \mathbf{C}\mathbf{C}^{\mathrm{T}}$.

For a sample of $\{\boldsymbol{x}j: j = 1, 2, \ldots, m\}$ from model (4) and (5), the log-likelihood is

$$L = \sum_{j=1}^{m}\log(p(\boldsymbol{x}_j)) = -\frac{m}{2}\left\{d\log(2\pi) + \ln\left|\Sigma\right| + \mathrm{tr}\{\Sigma^{-1}\mathrm{S}\}\right\}, \tag{6}$$

where **S** is the sample covariance matrix. The estimate of **C** that maximizes the log-likelihood (6) is shown to satisfy (Tipping and Bishop 1999):

$$\mathbf{C}_{ML} = \mathbf{U}_p(\Lambda_p - \sigma_w^2\mathrm{I})^{1/2}\mathrm{R}\,. \tag{7}$$

The interpretation of Equation (7) is that the maximum of log-likelihood is achieved when the column vectors of $d \times p$ matrix $\mathbf{U}_p$ are eigenvectors of **S** corresponding to the p largest eigenvalues. The eigenvalues λ_i are stored in descending order within matrix $\boldsymbol{\Lambda}_p = \mathrm{Diag}\{\lambda_k\}$ ($k = 1, 2, \ldots, p$). The column vectors in $\mathbf{U}_p$ are also called principal eigenvectors due to their relationship with respect to the eigenvectors, and **R** is a $p \times p$ orthogonal matrix. Furthermore, the ML estimate of σ_w^2 is given by

$$\sigma_{ML}^2 = \frac{1}{d-p}\sum_{k=p+1}^{d}\lambda_k\,,$$

in which noise variance is viewed as the average of the d−p smallest eigenvalues.

The maximum-likelihood estimate of **C** in Equation (7) can be calculated by an iterative expectation-maximization (EM) algorithm between the following equations (Booth & Hobert 1999; Dempster et al. 1977):

$$\mathrm{C}_{new} = \mathrm{SC}_{old}(\sigma^2_{w,old}\mathrm{I} + \mathrm{M}^{-1} + \mathrm{C}^{\mathrm{T}}_{old}\mathrm{SC}_{old})^{-1}, \tag{8}$$

$$\sigma^2_{w,new} = \frac{1}{d}\mathrm{tr}(\mathrm{S} - \mathrm{SC}_{new}\mathrm{M}^{-1}\mathrm{C}^{\mathrm{T}}_{new}), \tag{9}$$

where $\mathrm{M} = (\sigma_w^2\mathbf{I} + \mathbf{C}^{\mathrm{T}}\mathbf{C})$. Thus, the optimal $\mathbf{C}$ and noise variance $\sigma_w^2\mathbf{I}$ are obtained when Equations (8) and (9) converge. Note that the rotation matrix R brings somewhat ambiguity in the ML estimation for matrix $\mathbf{C}$. In the proposed method, this ambiguity can be resolved by determining the rotation matrix from $\mathbf{C}^{\mathrm{T}}_{ML}\mathbf{C}_{ML} = \mathbf{R}^{\mathrm{T}}(\mathbf{\Lambda}_p - \sigma^2_w\mathbf{I})\mathbf{R}$, i.e., $\mathbf{R}$ is the eigenmatrix of $\mathbf{C}^{\mathrm{T}}_{ML}\mathbf{C}_{ML}$. As implied in Equation (7), latent variable model (4) effects a mapping from the latent space into the principal subspace of multivariate data x. In this sense the ML estimate $\mathbf{C}_{ML}$ for model (4) is indeed a form of principal component analysis. Therefore, we choose to term the proposed method as maximum-likelihood PCA.

One major advantage of latent variable model and corresponding MLPCA estimate is to offer an effective way to link the variability analysis on solder paste profile and subsequent binary inspections to a candidate set of process faults. Suppose that multivariate measurements $\boldsymbol{x}$ on solder pastes are correlated due to common unobservable process factors $\boldsymbol{v}$, this paper tries to provide an analytical tool for diagnosing product quality by relating variation pattern on physical characteristics to these hypothesized systematic causes. As demonstrated in later case study, this method is developed on a process-oriented basis, which applies MLPCA to determine the latent space of systematic root causes and then project logistic regression coefficients onto this reduced space for pattern clustering and interpretation. The visualization of clustered variation pattern, combined with appropriate engineering knowledge, will help identify the underlying process faults. On the other hand, classical PCA is a data-oriented approach that tries to explain the variance of $\boldsymbol{x}$ by seeking the principal eigenvectors. PCA works well for situations when a single process fault occurs (i.e., $p = 1$), but can not produce interpretable results for process diagnosing when $p > 1$ (Apley & Shi 2001). The limitations of PCA on root cause recognition or fault interpretation thus hamper its diagnostic capabilities in complicated multivariate process control.

The latent variable model (4) also considers the effects from measurement noise on solder paste, which has been a non-neglectable factor when accurate process modeling and diagnosing are required. The probabilistic formulation enables the introduction of likelihood measure for obtaining ML estimate CML. It is worth noting that CML is built on the assumption that p is known. However, the probabilistic model itself does not provide a mechanism to determine p. For practical implementation, we need to address how to define the dimension of latent variable v prior to parameter estimation. For $p = d–1$, the model is equivalent to a full covariant Gaussian distribution, while in case of $p < d–1$ it implies that the remaining $d–p$ directions is caused by noise variance σ_w^2. As a possible approach, cross-validation may compare all potential values of p, however, it becomes expensive in computation when d increases. Simulation results over numerous examples with varying p and d, suggest the following practical rule to determine p and substitute it into the iterative EM algorithm:

$$p = \inf_l \{ \frac{\frac{1}{l+1}\sum_{k=1}^{l+1}\lambda_k - \frac{1}{l}\sum_{k=1}^{l}\lambda_k}{\frac{1}{l}\sum_{k=1}^{l}\lambda_k - \frac{1}{l-1}\sum_{k=1}^{l-1}\lambda_k} >> 1 \}. \tag{10}$$

4.2 The regression coefficient clustering algorithm

The dramatic advances in in-process sensor and data collection technologies enable vast quantities of physical features to be measured about the manufacturing system. For instance, in PCB assembly, laser-optical measurement machines are commonly installed to record detailed dimensional characteristics of wet solder paste after it is deposited onto the board in stencil printing. When electronic components are positioned and the solder is cured in the re-flow oven, dimensional characteristics are obtained via X-ray laminography (Crida et al. 1997; Litman 2004; Neubauer 1997). As in any quality control applications, one fundamental objective considered in this paper is to explain as precisely as possible the nature of variation on solder paste and identify the root causes of binary defects by utilizing the earlier measured physical information.

Although the aforementioned logistic regression method can estimate coefficients for each measurement variable, the high dimensionality of solder profile makes it not efficient for engineers to explore the nature of how the underlying process factors cause the defective outputs. On the other hand, as shown in Fig. 2, the latent variable model helps recognize the patterns of solder paste variation and thereby identify the corresponding systematic causes during stencil printing operation. By integrating the logistic regression method with latent variable model (4), the proposed methodology will quantify the effects on solder profile $\boldsymbol{x}$ and defects y from process faults v, which is performed entirely on the collected sample data with no a priori knowledge about the patterns of variation. Therefore, a core component of this approach includes the proper clustering over regression coefficients with respect to variables $\boldsymbol{v}$, which provides more intuitive insight into the interdependencies among multiple measurement variables.

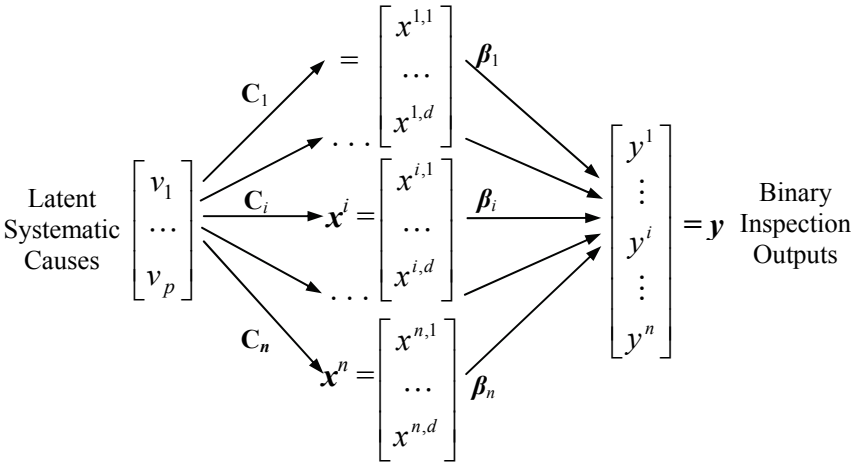

Fig. 2. Illustration of latent variable model that explains the relationship between systematic cause $\boldsymbol{v}$, solder paste profile $\boldsymbol{x}$, and final defect inspection output $\boldsymbol{y}$.

Following the assumptions on model (4), let $\boldsymbol{x} = [x_1, x_2 \ldots x_d]^T$ represent the measurable characteristics of a solder paste, and $\{\boldsymbol{x}_i; i = 1, 2, \ldots, n\}$ be a set of n solder pastes in the board. Fig. 2 implies that p independent causes v_j apply their joint effects on the variation of physical profile $\boldsymbol{x}_i$ through a constant matrix $\mathbf{C}_i$, and produce the consequent binary outputs

y_i through logistic regression coefficient $\boldsymbol{\beta}_i$. In particular, the effect from cause $\boldsymbol{v}_j$ is represented by the j^{th} column vector $\boldsymbol{c}_{i,j}$ in $\mathbf{C}_i$. Since each $\boldsymbol{v}_j$ is scaled to have unit variance, $\boldsymbol{c}_{i,j}$ indicates the magnitude or severity of variation caused by $\boldsymbol{v}_j$. After clustering a sample of solder pastes based on the distribution of their regression coefficients in terms of $\boldsymbol{v}_j$, quality diagnosing of stencil printing becomes possible by assigning a process fault to the solder pastes within the same group.

Prior to the clustering analysis over regression coefficients, Equation (4) is substituted into the logistic regression model, yielding new coefficients $\boldsymbol{\beta}_{v,i} = \mathbf{C}_i^{\mathrm{T}} \boldsymbol{\beta}_i = [\beta_{v,i,1}\ \beta_{v,i,2}\ \ldots\ \beta_{v,i,p}]^{\mathrm{T}}$ for latent variable $\boldsymbol{v}$. Now binary data y_i can be explained by systematic causes $\boldsymbol{v}_j$, which takes the form of a logit function, that is,

$$\text{logit}\{y_i = 1 \mid \boldsymbol{v}, \boldsymbol{\beta}_{v,i}\} = \sum_{j=1}^{p} \beta_{v,i,j} v_j + \boldsymbol{\beta}_i^{\mathrm{T}} \boldsymbol{w} \equiv \boldsymbol{v}^{\mathrm{T}} \boldsymbol{\beta}_{v,i} + \varepsilon_w,$$

where ε_w denotes the transformed noise effect. The new coefficients $\beta_{v,i,j}$ correspond to the change in the log odds per unit change in v_j when v_j does not interact with other sources (this is reasonable given the latent variable model assumptions). Or, the effect of increasing v_j by 1 is to increase the odds that $y_i = 1$ by a factor $\exp(\beta_{v,i,j})$.

Since $\boldsymbol{\beta}_{v,i}$ depends on the systematic causes $\boldsymbol{v}$, the regression coefficients can be classified so that each cluster describes the similar pattern of solder paste variation. In other words, the proposed clustering method is used to separate the impacts from cause v. Once all inspected solder pastes on a PCB are clustered in terms of $\boldsymbol{\beta}_{v,i}$, process diagnosis for variation reduction can be performed since each cluster is mapped to a specific process fault or assignable cause.

4.3 MLPCA based clustering algorithm for quality diagnosing

As a statistical tool for diagnosing the quality of solder pastes, the proposed MLPCA based regression coefficients clustering algorithm is now summarized as follows:

Step 1. Apply logistic regression model (1) to binary inspection data collected from m PCBs, yielding the estimates of coefficients $\boldsymbol{\beta}_i$ for the i^{th} solder paste through sample $\{ y_j^i : i = 1, 2, \ldots, n; j = 1, 2, \ldots, m \}$.

Step 2. Given the set of measured solder paste profile $\boldsymbol{x}_j^i$, determine the dimension p via rule (10) and estimate the matrix $\mathbf{C}_i$ in model (4) by MLPCA method.

Step 3. Calculate new regression coefficients $\boldsymbol{\beta}_{v,i} = \mathbf{C}_i^{\mathrm{T}} \boldsymbol{\beta}_i$, followed by a k-means clustering algorithm (Hastie et al. 2001) over $\boldsymbol{\beta}_{v,i}$ to recognize the coefficient clusters.

Step 4. Present the geometrical clustering results on the board to process operators to identify the process faults by utilizing their engineering knowledge. The diagnosing results of solder paste quality will then lead to appropriate stencil printing operation adjustments.

By taking advantage of the diagnostic information from latent variable model and logistic regression coefficients, the MLPCA based clustering algorithm provides a visual way to relate stencil printing process problems to the variation on solder paste profile and consequent binary defects. Case study in the following section shows that the proposed method is favorable in improving process quality by developing a more interpretable relationship between variation pattern and physical faults.

5. Application in PCB assembly

In stencil printing process, each solder paste is deposited on the board automatically by printing machines, then registered with the screen and printed. Stencil printing is known to be an established technology, however, there are some uncontrolled factors that influence the quality of solder pastes (Lathrop 1997; Liu et al. 2001), and hence cause component failures in PCB assembly. In order to produce pastes with minimal variation on physical profile, the controllable parameters for printing operation should be monitored and adjusted by appropriate diagnosing of solder paste quality. In the present study, solder paste printability was denoted by a physical profile collected from laser triangulation and X-ray based measurement machine. The purpose of the present experimental research is to identify the systematic factors in solder deposition process by quantifying their impacts on paste quality. The set of process factors include printer steel squeegee angle, printing direction, and squeegee speed, etc.

The variation on measured solder paste profile that leads to binary inspection results stems from improper parameter settings of stencil printing, called systematic factors. Their effects on solder paste (such as solder paste volume, area, and height, etc.) are present in multivariate profile $\boldsymbol{x}$. Due to the common factors $\boldsymbol{v}$, variables in $\boldsymbol{x}$ are always highly correlated, as shown in the scatter plots in Fig. 3. For a specific solder paste, the plots were drawn over pairs of distinct physical solder paste features from a sample of $m = 30$ inspected boards. In semiconductor fabrication, the solder paste profile often includes paste thickness, paste volume, shape of heel fillet, shape of toe or center fillet, alignment between pad and lead, pad average width, pad average height, and pad volume, etc. For purpose of illustration, we chose ten physical features as element variables of vector $\boldsymbol{x}$, that is, $d = 10$.

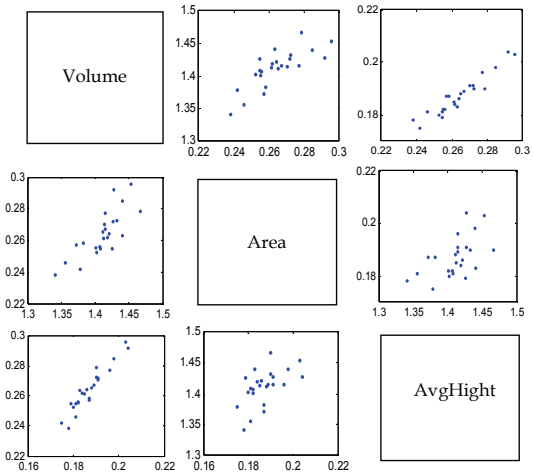

Fig. 3. Scatter plot of selected pairs of measured variables in solder paste profile (e.g., paste volume, area, and average height).

Next, the coefficients clustering algorithm proposed in Section 4 was applied to map the pattern of solder paste defects on PCB to the latent systematic causes, given the assumption that variation of solder paste profile was not completely random due to the measurement noise. To accommodate pattern clustering and visualization, the present experimental study

was undertaken over a region of PCB that consists of more than 3000 solder joints (e.g., n = 3012), as shown in Fig. 4. Given the sample of binary inspection yi and corresponding physical profile $\boldsymbol{x}_i$ (i = 1, 2, …, n), we first calculated the estimation of coefficients $\boldsymbol{\beta}_i$ by logistic regression model (1). MLPCA was then applied to estimate variation pattern matrix $\mathbf{C}_i$, in which the dimension of systematic cause v_i was always determined as two by the rule (7) (i.e., p = 2 for all i). As indicated in the algorithm summary, after projecting original $\boldsymbol{\beta}_i$ onto the latent space spanned by $\mathbf{C}_i$, the new coefficients $\boldsymbol{\beta}_{v,i}$ became available for k-clustering algorithm (Hastie et al. 2001), which classify them into two clusters.

The graphical illustration of clustered $\boldsymbol{\beta}_{v,i}$ in Fig. 4 also validate the presence of two clusters as identified by the standard k-means algorithm. Since each solder paste is positioned on PCB by the unique X-Y coordinates, we can visualize the clustered coefficients on a printed circuit board such that the pastes in each cluster are denoted by the same symbol (e.g., "+" for cluster 1 and "×" for cluster 2).

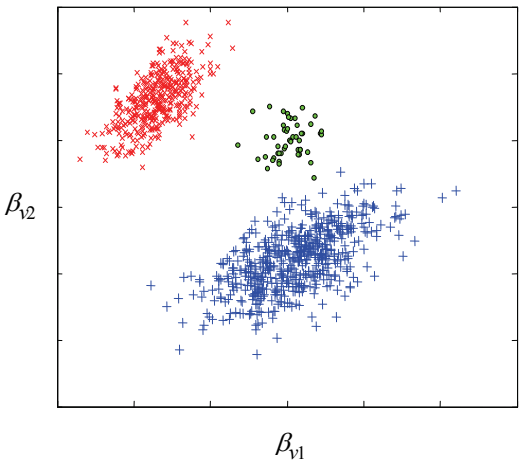

Fig. 4. Scatter plot of their logistic regression coefficients $\boldsymbol{\beta}_v = [\beta_{v1}\ \beta_{v2}]^{\mathrm{T}}$.

MLPCA implied that there were 2 systematic causes that governed the variation over the measured solder paste profile. The pastes denoted by '+' in Fig. 4, for example, were dominantly affected by the first systematic cause v_1, which almost lie on the horizontal direction of the board, while the pastes denoted by '×' were distributed along the vertical direction and influenced mainly by the second cause v_2. The graphical demonstration of clustered coefficient results in Fig. 4 thus helped process engineers to adopt their expert knowledge and experiences in diagnosing the solder paste defects. For instance, the solder pastes denoted by "+" in cluster 1 had relatively large coefficients $\beta_{v,i,1}$ and were mostly distributed along the length of PCB. That is, the systematic cause corresponding to this cluster should influence the solder pastes along the horizontal direction to a greater extent during stencil printing. Intensive discussions with process engineers have provided a reasonable explanation for the causes to be inappropriate parameter settings in controlling the stencil printing speed and printing pressure. These process factors are expected to generate large variation of solder paste profile along the horizontal and vertical direction, respectively, and correspondingly more inspected defects.

Further investigation on other potential process faults implied that the above identified systematic causes are most likely to produce the consistent results. The diagnostic results also agreed with the natural speculation on stencil printing diagram in Fig. 1, where printing speed usually influences the solder pasts along the length of PCB, while printing pressure has greater impacts on solder paste quality than other process factors (e.g., separation distance, printer alignment) along the width of PCB. In addition, detailed inspections revealed that a substantial portion of the quality deficiencies (such as slumping, bridging and bleeding of paste underneath the stencil) along the width of PCB was caused by abnormal high printing pressure during stencil printing.

The case study shows that the systematic pattern on PCB assembly defects is often owing to specific process faults such as inappropriate process operations, rather than completely random due to the environmental or measurement noise. The proposed coefficients clustering algorithm provides an effective process-oriented diagnosis tool for identifying such production irregularities. By assuming the potential systematic causes are mapped to the clusters of solder pastes with similar coefficients, the variation on solder paste profile and corresponding binary defects can be mapped to improper parametric control that deviates from optimal conditions, which will suggest informative corrections to adjust the stencil printing to improve process quality.

6. Conclusion

The distillation of massive quantities of solder paste inspection data into relevant quality information allows rapid understanding of the low production yield in PCB assembly. The statistical diagnosis method proposed in this paper provides more meaningful insights into the defect mechanisms than traditional yield analysis methods, which can identify the assignable causes of defects and their effects on yield by integrating MLPCA and logistic regression model. This offers a systematic representation on the impacts of process condition changes to the variation of solder paste profile. The probabilistic latent variable model allows ML estimation to determine the latent space by iteratively maximizing the likelihood function. In contrast to standard PCA, this approach is also efficient for multivariate process analysis when some sample data are missing. The clustering algorithm over the projected regression coefficients onto the latent space is relatively easy to implement with affordable computational effort. Experimental study demonstrates that the statistical interpretation of solder defect distributions can be enhanced by intuitive pattern visualization for process fault identification and variation reduction.

7. References

Apley, D. & Shi, J. (2001) A Factor-Analysis Method for Diagnosing Variability in Multivariate Manufacturing Processes. *Technometrics*, Vol. 43 (1), pp. 84-95.

Barajas, L. G. ; .Kamen, E.W. & Goldstein, A. (2001) On-line Enhancement of the Stencil Printing Process. *Circuits Assembly*, March, pp. 32-36.

Barajas, L.G. et al. (2003) Process Control in a High-noise Environment with Limited Number of Measurements. *Proceedings of American Control Conference*, Vol. 1, pp. 597-602, Denver, June 2003.

Bartholomew, D.J. & Knott, M. (1998). *Latent Variable Models and Factor Analysis*. Oxford University Press, London.

Booth, J. G. & Hobert, J. P. (1999). Maximizing Generalized Linear Mixed Model Likelihoods with An Automated Monte Carlo EM Algorithm. *Journal of the Royal Statistical Society, Series B*, Vol. 61, pp. 265-285.

Breed, S. (1998). Advances in Intelligent Stencil Printing. *Proceedings of NEPCON West 98*, Anaheim, California, pp. 253-256, 1998.

Crida, R. C. ; Stoddart, A. J. & Illingworth, J. (1997). Using PCA to Model Shape for Process Control. *International Conference on Recent Advances in 3-D Digital Imaging and Modeling*, pp. 318-325, Ottawa, Canada, 1997.

Cunningham, S. P. & MacKinnon, S. (1998). Statistical Methods for Visual Defect Metrology. *IEEE Transactions on Semiconductor Manufacturing*, Vol. 11(1), pp. 48 –53.

Dempster, A. P. ; Laird, N. M. & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1-38.

Gopaladrishnan, L. & Srihari, K. (1999). Process Development for Ball Grid Array Assembly using a Design of Experiments Approach. *Journal of Advanced Manufacturing Technology*, Vol. 15, pp. 587-596.

Hamada, M. & Nelder, J. A. (1997). Generalized Linear Models for Quality-improvement Experiences. *Journal of Quality Technology*, Vol. 29, pp. 292-304.

O'Hara, W. & Lee, N.C. (1996). How Voids Develop in BGA Solder Joints. *Surface Mount Technology*, pp. 44-47, January 1996.

Hastie, T. ; Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Jaakkola, T. S. & Jordan, M. I. (1997). A Variational Approach to Bayesian Logistic Regression Models and Their Extensions. *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, Fort Lauderdale, Florida, 1997.

Lathrop, R. R. (1997). Solder Paste Print Qualification using Laser Triangulation. *IEEE Transactions on Components and Packaging Manufacturing Technologies*, Vol. 20, pp. 174-182.

Litman, E. (2001). Solder Paste Printing: An Inside Look. *Surface Mount Technology*, pp. 30-34, January 2004.

Liu, S. et al. (2001). A Novel Approach for Flip Chip Solder Joint Quality Inspection: Laser Ultrasound and Interferometric System. *IEEE Transactions on Components and Packaging Technologies*, Vol. 24(4), pp. 616-624.

McCulloch, C. E. (1997). Maximum Likelihood Algorithm for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, Vol. 92, pp. 162-170.

Neubauer, C. (1997). Intelligent X-Ray Inspection for Quality Control of Solder Joints. *IEEE Transactions on Components, Packaging and Manufacturing Technology-Part C*, Vol. 20(2), pp. 111-120.

Pan, J. et al. (1999). Critical Variables of Solder Paste Stencil Printing for Micro-BGA and Fine Pitch QFP. *IEEE/CPMT International Electronics Manufacturing Technology Symposium*, Austin, October 1999.

Rao, S. et al. (1996). Monitoring Multistage Integrated Circuit Fabrication Processes. *IEEE Transactions on Semiconductor Manufacturing*. Vol. 9(4), pp.495-505.

Tipping, M. E. & Bishop, C.M. (1999). Mixtures of Probabilistic Principal Component Analyzers. *Neural Computation*, Vol. 11(2), pp. 443-482.

Venkateswaran, S. et al. (1997). A Realtime Process Control System for Solder Paste Stencil Printing. *Proceedings of International Electronics Manufacturing Technology Symposium*, pp. 62-67, Austin, October 1997.

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Feng Zhang (2009). A Data Mining Algorithm for Monitoring PCB Assembly Quality, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/a_data_mining_algorithm_for_monitoring_pcb_assembly_quality

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821