A random effect autologistic regression model with application to the characterization of multiple microstructure samples

Nailong Zhang and Qingyu Yang
Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI, USA

ABSTRACT
The microstructure of a material can strongly influence its properties such as strength, hardness, wear resistance, etc., which in turn play an important role in the quality of products produced from these materials. Existing studies on a material’s microstructure have mainly focused on the characteristics of a single microstructure sample and the variation between different microstructure samples is ignored. In this article, we propose a novel random effect autologistic regression model that can be used to characterize the variation in microstructures between different samples for two-phase materials that consist of two distinct parts with different chemical structures. The proposed model differs from the classic autologistic regression model in that we consider the unit-to-unit variability among the microstructure samples, which is characterized by the random effect parameters. To estimate the model parameters given a set of microstructure samples, we first derive a likelihood function, based on which a maximum likelihood estimation method is developed. However, maximizing the likelihood function of the proposed model is generally difficult as it has a complex form. To overcome this challenge, we further develop a stochastic approximation expectation maximization algorithm to estimate the model parameters. A simulation study is conducted to verify the proposed methodology. A real-world example of a dual-phase high strength steel is used to illustrate the developed methods.

1. Introduction
Microstructure generally refers to the structure of a prepared surface of a material that is examined using a range of microscopy techniques. The microstructure of many materials consists of multiple phases, where a phase is taken to be any part of a material with a distinct chemical or physical structure. For example, a dual-phase steel, a type of two-phase material, has a microstructure that consists of martensite and ferrite. Figure 1 shows an image of the microstructure of a high-strength dual-phase steel sample obtained using an optical microscope.

The microstructure of the material can strongly influence the properties of a material, such as strength, hardness, wear resistance, etc. The effect of the microstructure on material properties has been extensively studied in materials science (see, for example, Lewandowski et al. (1989), Alpas and Zhang (1994), and Zaefferer et al. (2004)) in order to understand the mechanical properties of a material and/or to design new materials. Recently, with the development of computer simulation techniques, the statistical characterization of microstructures has attracted increasing attention (Torquato, 2002; Grigoriu, 2003; Feng et al., 2014). These methods focus on a single microstructure sample.

In this article, we model the variation of multiple microstructure samples, focusing on two-phase materials that consist of two distinct parts with different chemical or physical structures.

The motivating problem of the proposed methodology is the quality control of two-phase steels. In the steel manufacturing processes, due to the constant variation of process parameters—e.g., cooling speed, heating energy, spatial temperature distribution (Li et al., 2006; Irani and Taheri, 2008)—significant variation in the properties of steels can occur between steels produced by different manufacturers, different batches of the same manufacturer, or different products of the same batch. This variation, including both mechanical properties and microstructures, is widely acknowledged in the steel industry. As one of the critical goals of quality control is to reduce variation across products, it is important to model and capture the variations in different microstructure samples of two-phase steels in order to improve quality control. In addition, the proposed microstructure models can be used to create new microstructure samples whose properties can be analyzed using finite element analysis without having to conduct expensive physical experiments.

The existing research on the modeling of microstructures mainly focuses on a single microstructure sample. To describe the single microstructure sample shown in Fig. 1, the basic model assumes that one phase is spatially distributed in the other phase in a uniform manner (Paul, 2013). However, the assumption of a spatially uniform distribution does not allow the complex characteristics of most microstructures to be captured. Another commonly used model utilizes non-parametric two-point correlation functions to model the microstructure. In
this approach the two-point correlation functions are defined as the probability that two pixels in the image share the same phase given the relative displacement of the two pixels (Torquato, 2002). Intensive studies have been conducted on the relation between the properties of a microstructure and the physical properties using this approach (Jiao et al., 2007; Saheli et al., 2008). Another widely studied method is the Gaussian random field approach (Feng et al., 2014), which assumes that a binary image can be modeled by translating a Gaussian random field into a binary field using a threshold value. This method assumes that the binary images are fully characterized by the first two orders of the statistics of the spatial sample; i.e., the mean and the autocorrelation matrix.

Another approach used in the study of a material’s microstructure, the autologistic regression model (e.g., Huffer and Wu (1998); Moon and Russell (2008)), has been widely used to model spatial 0–1 type binary data in various disciplines including ecology, agriculture, and image analysis. For example, Cross and Jain (1983) showed that the autologistic regression model can be used for binary image synthesis/texture generation, which is related to the microstructure modeling problem considered in this article. The autologistic model, which directly imposes a joint Markov random field, has been proved to be a very useful tool for modeling spatial 0–1 type of data. However, it can only be applied to a single sample.

Although significant success has been achieved using the methods available in the literature in terms of understanding material properties and the design of new materials, existing research lacks the capability to model and capture the unit-to-unit variability among multiple microstructure samples. In this article, we propose a novel parametric random effect autologistic regression model to deal with the multi-sample situation in microstructure modelling. Our model is an extension of the classic autologistic regression model that includes the effects of random effect parameters. We also developed two methods to estimate the model parameters given a set of microstructure samples. We first derived the likelihood function of the newly proposed model, based on which the maximum likelihood estimation method is developed. However, due to the complex form of the likelihood function in the proposed model, maximizing the likelihood function is generally difficult. To overcome this challenge, we further developed the stochastic approximation expectation maximization (SAEM) algorithm, called autologistic SAEM (ALSAEM), to estimate the model parameters in an efficient manner. The developed methodology was verified in a simulation study and illustrated using a real-world example. Since the developed model provides a way to capture the unit-to-unit variability in the microstructures of multiple samples, an obvious potential application is in quality control in the manufacture of steel.

The rest of this article is organized as follows. After the Introduction, in Section 2, we introduce the proposed random effect autologistic regression model. In Section 3, the parameter estimation methods are developed. In Section 4, a simulation study is conducted to verify the proposed methodology, and in Section 5 a real-world example of a dual-phase high-strength steel is used to illustrate the developed methods. Finally, a summary is given in Section 6.

2. Statistical modeling

First, we introduce the data description and notations used in the proposed model. The microstructure of two-phase materials can be represented by binary images. Consider a set of $n$ binary microstructure images denoted as $\mathcal{M} = \{X_1, \ldots, X_n\}$; where all image samples $X_1, \ldots, X_n$ are assumed to have the same dimension $d_1 \times d_2$. The $i$th pixel on the $i$th sample is denoted as $X_{i,s}$, where $i \in \{1, \ldots, n\}$ and $s \in \{1, \ldots, d_1 \times d_2\}$. As the image is binary, $X_{i,s} \in \{0, 1\}$, where the 0 and 1 represent two distinguishable phases. A single microstructure image, $X_i$, can also be treated as a lattice, as illustrated in Fig. 2, where each site $s$ corresponds to a pixel of the microstructure image $X_i$. For convenience, we also use $X_i$ to denote the corresponding lattice of the microstructure sample $X_i$.

2.1. The autologistic regression model

The classic autologistic regression model was introduced by Gumpertz et al. (1997) and used to address the problem of spatial autocorrelation in the distribution data of species. It subsequently found application in other disciplines including ecology, agriculture, epidemiology, and image analysis. The autologistic

![Figure 1](image-url). The microscope (left) used to obtain the microstructure image (right) of a dual-phase high-strength steel sample with a 100 × magnification, where the black pixels and white pixels refer to the two phases; i.e., martensite and ferrite, respectively.

![Figure 2](image-url). A lattice corresponding to a sample image of the microstructure.
regression model can also be treated as a Markov random field model, which has been widely used in the modeling of binary spatial data (Cross and Jain, 1983; Hughes et al., 2011). The classic autologistic regression model for a single lattice, denoted by \( \mathbf{X} \) is defined as follows:

\[
\Pr (X_i = 1 | X_i = x_i; t \in N(s)) = \frac{\exp (\eta_i)}{1 + \exp (\eta_i)}, \tag{1}
\]

where \( \eta_i = \beta_0 + \sum_{t \in N(s)} \beta_i x_{it} \), and \( N(s) \) is defined as a set of neighbors of site \( s \). For example, the four sites \( \{l_u, t_d, t_l, t_r\} \) that are adjacent to \( s \) can be chosen as neighbors as illustrated in Fig. 2.

In practice, model (1) is also commonly written in the following form:

\[
\Pr (X_i = 1 | X_i = x_i; t \in N(s)) = \frac{\exp (\eta_i \times x_i)}{1 + \exp (\eta_i)}. \tag{2}
\]

In the autologistic regression model, it is assumed that the probability distribution of a single site on the lattice solely depends on its local neighboring sites, which is called the Markov property. According to the Markov property, the distribution of \( X_i \) conditional on all the other sites on the lattice is equivalent to the distribution of \( X_i \) conditional on \( N(s) \).

2.2. A novel random effect autologistic regression model

In this article, we propose a novel random effect autologistic regression model to capture the unit-to-unit variability among different microstructure samples. Formally, the proposed model is defined as follows:

\[
\begin{align*}
\Pr (X_{i,t} = x_{i,t} | X_{i,t} = x_{i,t}; t \in N_i(s)) &= \frac{\exp (\eta_{i,t} \times x_{i,t})}{1 + \exp (\eta_{i,t})}, \\
\eta_{i,t} &= \alpha_0 + \sum_{q=1}^{q} \alpha_q x_{i,t} + \sum_{t=1}^{q} \beta_{i,t} x_{i,t} \quad t \in N_i(s), \\
\beta_i &= (\beta_{i,0}, \beta_{i,1}, \ldots, \beta_{i,q})^T \sim \mathcal{N}(\mathbf{0}, \Sigma),
\end{align*}
\tag{3}
\]

where \( \alpha = (\alpha_0, \alpha_1, \ldots, \alpha_q)^T \) is a constant vector with dimension \( q + 1 \) and \( \beta_i = (\beta_{i,0}, \beta_{i,1}, \ldots, \beta_{i,q})^T \) is a random effect vector, which is assumed to have a multivariate normal distribution with zero mean and covariance matrix \( \Sigma \). In this article, we focus on the multivariate normal distribution of the model parameters \( \beta_i \); however, other distributions can be used.

In the proposed random effect autologistic regression model (3), the first equation represents the Markov property, similar to the classic autologistic regression model; i.e., the probability mass function (pmf) of \( X_{i,t} \) solely depends on a small set of neighbors of site \( s \) on \( \mathbf{X} \). The constant vector \( \alpha \) captures the fixed effects shared by all samples. The realization of random effect \( \beta_i \) describes the specific properties for sample \( \mathbf{X}_i \). Thus, the unit-to-unit variability among different samples is explained by \( \beta_i \). If we define \( \omega_i = \alpha + \beta_i \), the model (3) can be further simplified to

\[
\begin{align*}
\Pr (X_{i,t} = x_{i,t} | X_{i,t} = x_{i,t}; t \in N_i(s)) &= \frac{\exp (\eta_{i,t} \times x_{i,t})}{1 + \exp (\eta_{i,t})}, \\
\eta_{i,t} &= \omega_{i,0} + \sum_{i=1}^{q} \omega_{i,i} x_{i,t} \quad t \in N_i(s), \\
\omega_i &= (\omega_{i,0}, \omega_{i,1}, \ldots, \omega_{i,q})^T \sim \mathcal{N}(\mu, \Sigma)
\end{align*}
\tag{4}
\]

where \( \mu = \alpha \).

To deal with the sites on the boundaries, we follow the periodic assumption that is generally used in the literature (Ostoja-Starzewski, 2002; Gommès et al., 2012). Specifically, we assume that the leftmost column is connected with the rightmost column, and the first row is connected with the last row. In addition, the top-left corner is connected with the bottom-right corner and the top-right corner is connected with the bottom-left corner. By making this assumption, all sites have the same number of neighbors.

It is worthwhile comparing the proposed model with the classic random effect logistic regression model (Khuri, 2010). Although the proposed random effect autologistic regression has a similar form compared with the random effect logistic regression, they are essentially different. The major difference between these two models is that the conditional pmf of \( X_{i,t} \) in the proposed model also depends on its neighbor sites, which makes the model a type of auto-regression model.

3. Parameter estimation

Given a set of \( n \) binary microstructure sample images \( \mathcal{M} = \{X_1, \ldots, X_n\} \), we develop parameter estimation methods to estimate the model parameters \( \theta = (\mu, \Sigma) \) of the proposed random effect autologistic model. First, we derive the log-likelihood function of the proposed model in Section 3.1, based on which the commonly used Maximum Likelihood Estimate (MLE) is given. However, the MLE method is difficult to implement due to the complex form of the log-likelihood function. To overcome this difficulty, a parameter estimation method called ALSAEM is developed in Section 3.2.

3.1. Log-likelihood function and MLE

Given \( \omega_i \) the conditional pmf of a specific sample lattice \( \mathbf{X}_i \) falls into the exponential family as follows (Cressie, 1993):

\[
Pr(\mathbf{X}_i = x_i | \omega_i) = \frac{\exp(Q(x_i | \omega_i))}{c(\omega_i)},
\tag{5}
\]

where \( c(\omega_i) = \sum_{\mathbf{X}_i} \exp(Q(x_i | \omega_i)) \) is a normalization constant that denotes the ensemble average of \( \exp(Q(x_i | \omega_i)) \) with respect to all possible realizations of \( \mathbf{X}_i \); and \( Q(x_i | \omega_i) \) is called the negentropy function, which is defined as follows:

\[
Q(x_i | \omega_i) = \sum_{s=1}^{d_1 \times d_2} \left[ \omega_{i,0} x_{i,s} + \frac{1}{2} \sum_{t=1}^{q} \omega_{i,t} x_{i,s} x_{i,t} \right], \quad t \in N_i(s).
\tag{6}
\]

In the proposed random effect autologistic regression model, \( \omega_i \) is a random vector and is generally unobserved. Thus, to obtain the unconditional pmf for sample \( \mathbf{X}_i \), we need to calculate
the expectation of Equation (5) with respect to \( \omega_i \) as follows:

\[
\Pr(X_i = x_i; \theta) = \int_{\omega_{i}} \frac{\exp(Q(x_i|\omega_{i}))}{c(\omega_{i})} f(\omega_{i}; \theta) d\omega_{i}, \tag{7}
\]

where \( f(\omega_{i}; \theta) \) is the probability density function (pdf) of the multivariate normal distribution with parameter \( \theta = \{\mu, \Sigma\} \). As the microstructure samples are independent conditional on the unknown parameter \( \theta \), the overall likelihood function can be obtained as the product of the likelihood function for each single sample; that is,

\[
\mathcal{L}(\theta|M) = \prod_{i=1}^{n} \int_{\omega_{i}} \frac{\exp(Q(x_i|\omega_{i}))}{c(\omega_{i})} f(\omega_{i}; \theta) d\omega_{i}. \tag{8}
\]

The MLE of \( \theta \) can be obtained by maximizing Equation (8). However, two challenges exist to maximizing the likelihood function (8), which makes the MLE method difficult to implement in practice. First, as each site on \( X_i \) has two possible values and there are \( d_1 \times d_2 \) sites on \( X_i \), the lattice \( X_i \) has \( 2^{d_1 \times d_2} \) possible realizations. Thus, to calculate the normalization constant \( c(\omega_{i}) \), we need to enumerate all the \( 2^{d_1 \times d_2} \) possible realizations of sample \( X_i \), which is almost impossible when the size of the lattice is large. Second, the dimension of the random vector \( \omega_i \) is large if a large neighbor set is selected. As a result, a high-dimensional integration is needed to evaluate Equation (8), which is not trivial.

### 3.2. The ALSAEM algorithm

To overcome the challenges of the MLE method, we developed the ALSAEM algorithm to estimate the model parameters. In the literature, the EM algorithm (Dempster et al., 1977; Wu, 1983; Bilmes, 1998) is commonly used to find the MLE of parameters for the situations where some variables are missing or unobserved. In the context of the proposed model, if we treat the random vector \( \omega_i \) as an unobserved variable, the EM algorithm is a good fit for this problem.

In the EM method, there is an expectation (E) step and a maximization (M) step, which alternate in an iterative scheme. To apply the EM method to our problem, in the kth iteration of the E step, the expectation of the log-likelihood function, denoted by \( E_{\Theta_{k}}[\log(\mathcal{L}(\theta|\Omega, M))] \), needs to be calculated with respect to the unobserved variables \( \Omega = \{\omega_{1}, \ldots, \omega_{n}\} \) given the current parameter estimates \( \tilde{\Theta}_{k} = \{\tilde{\mu}_{k}, \tilde{\Sigma}_{k}\} \), and \( E_{\Theta_{k}}[\log(\mathcal{L}(\theta|\Omega, M))] \) is maximized with respect to the unknown parameter \( \theta \) in the following M step. We use \( \tilde{\Theta}_{k} = \{\tilde{\mu}_{k}, \tilde{\Sigma}_{k}\} \) to denote the parameter estimate in the kth iteration.

The calculation of \( E_{\Theta_{k}}[\log(\mathcal{L}(\theta|\Omega, M))] \) in the E step of the traditional EM method is difficult, as it involves the integration of log(\( \mathcal{L}(\theta|\Omega, M) \)) with respect to \( \Omega \). To overcome this challenge, we use the SAEM (Delyon et al., 1999) method instead of the traditional EM method. The major difference between SAEM and traditional EM is that in the E step of SAEM the expectation of the log-likelihood function is calculated based on Monte Carlo Markov Chain (MCMC) simulation without integration, such that the computation in the E step is significantly simplified. In addition, we adapt the SAEM framework with a pseudo pmf of the microstructure sample to evaluate the log-likelihood function so that it can overcome the computational difficulty to evaluate the normalization constant \( c(\omega_{i}) \) when calculating log(\( \mathcal{L}(\theta|\Omega, M) \)) in the E step.

The framework of the developed parameter estimation approach is illustrated in Fig. 3. The details of E step and M step are given in the following subsections.

#### 3.2.1. The E step

Under the framework of SAEM, the E step consists of a simulation step and a stochastic approximation step to approximate \( E_{\Theta_{k}}[\log(\mathcal{L}(\theta|\Omega, M))] \). Figure 4 illustrates the internal structure of the E step in the developed algorithm.

Specifically, in the simulation step, we simulate \( \omega_i \) for each sample conditional on both \( \tilde{\theta}_k \) that is the incumbent of the estimate of \( \theta \) and the \( i \)th sample \( X_i \). The set of simulated \( \omega_i \) is denoted by \( \Omega_k = \{\omega_{i,1}, \ldots, \omega_{i,N}\} \) in the \( k \)th iteration. In this article, we apply the Metropolis–Hastings (MH) algorithm (Chib and Greenberg, 1995), which is a type of MCMC method, to construct a Markov chain whose stationary distribution is the conditional distribution of \( \omega_i \) conditional on \( \tilde{\theta}_k \) and \( X_i \).

The MH algorithm also runs iteratively and each of the iterations involves two steps. In its \( l \)th iteration, the first step is to draw a random variable \( \tilde{\lambda}_{i}^{(l)} \) from a distribution \( q(\tilde{\lambda}_{i}^{(l)}; \lambda_{i}^{(l-1)}) \), which is called proposal distribution and the second step is to update \( \lambda_{i}^{(l)} \) according to \( \tau(\tilde{\lambda}_{i}^{(l)}; \lambda_{i}^{(l-1)}) \), which is defined by

\[
\begin{align*}
\tau(\tilde{\lambda}_{i}^{(l)}; \lambda_{i}^{(l-1)}) &= \frac{q(\tilde{\lambda}_{i}^{(l-1)}; \lambda_{i}^{(l-1)}) \prod_{i=1}^{n} \Pr(X_i = x_i; \tilde{\lambda}_{i}^{(l)}) f(\tilde{\lambda}_{i}^{(l)}; \tilde{\theta}_k)}{q(\lambda_{i}^{(l)}; \lambda_{i}^{(l-1)}) \prod_{i=1}^{n} \Pr(X_i = x_i; \lambda_{i}^{(l-1)}) f(\lambda_{i}^{(l-1)}; \tilde{\theta}_k)}. 
\end{align*}
\]

In our developed method, we set \( \lambda_{i}^{(l)} = \tilde{\lambda}_{i}^{(l)} \) with probability \( \tau(\tilde{\lambda}_{i}^{(l)}; \lambda_{i}^{(l-1)}) \) and \( \lambda_{i}^{(l)} = \lambda_{i}^{(l-1)} \) with probability \( 1 - \tau(\tilde{\lambda}_{i}^{(l)}; \lambda_{i}^{(l-1)}) \).

In the first step of MH algorithm, different proposal distributions can be applied in the algorithm. For simplicity, in this

**Figure 3.** Parameter estimation framework for the proposed model.

**Figure 4.** Illustration of the E step.
article, we choose the distribution \(N(\tilde{\mu}_k, \tilde{\Sigma}_k)\) as the proposal distribution \(q(\tilde{\lambda}_i^{(l)}, \eta_i^{(l-1)})\) so that Equation (9) can be simplified to

\[
\tau(\tilde{\lambda}_i^{(l)}; \lambda_i^{(l-1)}) = \frac{\prod_{i=1}^{n} \Pr(X_i = x_i|\tilde{\lambda}_i^{(l)})}{\prod_{i=1}^{n} \Pr(X_i = x_i|\lambda_i^{(l-1)})}.
\]  

(10)

To calculate Equation (10) we have to evaluate \(\Pr(X_i = x_i|\lambda_i^{(l-1)})\) and \(\Pr(X_i = x_i|\tilde{\lambda}_i^{(l)})\) for \(i \in \{1, \ldots, n\}\) based on Equation (7), which involves the normalization constant \(c(l)\) and \(c(l-1)\). To allow us to avoid the use of a normalization constant, we proposed the use of a pseudo pmf instead of the exact pmf of the microstructure sample \(X_i\). The pseudo pmf has been widely used to replace the exact pmf for parameter estimation in the literature on classic autologistic regression modeling, with the corresponding estimate being referred to as the pseudo likelihood estimate (Cross and Jain, 1983). The pseudo pmf of a sample \(X_i\) is the product of the conditional pmf of each site on the lattice conditional on its neighbors:

\[
\Pr_p(X_i = x_i|\omega_i) = \prod_{i=1}^{d_i \times d_i} \frac{\exp(\eta_i \times x_i)}{1 + \exp(\eta_i)}.
\]  

(11)

By using Equation (11) instead of the exact pmf, Equation (10) becomes

\[
\tau(\tilde{\lambda}_i^{(l)}; \lambda_i^{(l-1)}) = \frac{\prod_{i=1}^{n} \Pr_p(X_i = x_i|\tilde{\lambda}_i^{(l)})}{\prod_{i=1}^{n} \Pr_p(X_i = x_i|\lambda_i^{(l-1)})}.
\]  

(12)

In the approximation step, the approximated expectation of the log-likelihood function denoted as \(J_k(\theta)\) is calculated by applying a stochastic approximation method (Delyon et al., 1999) developed in the literature. Specifically, \(J_k(\theta)\) is calculated using the follow formula:

\[
J_k(\theta) = \begin{cases} 
\log(L(\theta|M, \Omega_k)) ; & \text{if } k < K_1 \\
\frac{1}{k-K_1 + 1} \sum_{p=K_1}^{k} \log(L(\theta|M, \Omega_p)) ; & \text{otherwise}
\end{cases}
\]  

(13)

where the constant \(K_1\) is used to improve the convergence rate. When \(k < K_1\), only the log-likelihood function in the \(k\)th iteration is used to calculate \(J_k(\theta)\), whereas when \(k \geq K_1\), the moving average of the log-likelihood functions is used. For the details of the approximation method, the reader can refer to Delyon et al. (1999).

3.2.2. The M step

In the following M step, Equation (13) needs to be maximized with respect to \(\theta\) and the calculated \(\theta_{k+1}\) in the next iteration. The following theorem gives the optimum solution to maximize Equation (13). The detailed proof is listed in the Appendix.

**Theorem 1.** The optimal solution to maximize Equation (13) for the random effect autologistic regression model is given as follows:

\[
\begin{align*}
\mu & = \frac{1}{n} \sum_{i=1}^{n} (\omega_{i,k}) & \text{if } k < K_1 \\
\mu & = \frac{1}{n(k-K_1+1)} \sum_{p=K_1}^{k} \sum_{i=1}^{n} \omega_{i,p} & \text{if } k \geq K_1, \\
\Sigma & = \frac{1}{n(k-K_1+1)} \sum_{p=K_1}^{k} \sum_{i=1}^{n} (\omega_{i,p} - \tilde{\mu})(\omega_{i,p} - \tilde{\mu})^T & \text{if } k < K_1 \\
\Sigma & = \frac{1}{n(k-K_1+1)} \sum_{p=K_1}^{k} \sum_{i=1}^{n} (\omega_{i,p} - \tilde{\mu})(\omega_{i,p} - \tilde{\mu})^T & \text{if } k \geq K_1
\end{align*}
\]

(14)

3.2.3. Summary of the ALSAEM algorithm

We summarize the proposed ALSAEM algorithm as follows.

**Algorithm 1:**

- Initialize the model parameters, i.e., set \(\hat{\theta}_1 = \{\hat{\mu}_1, \hat{\Sigma}_1\}\); for \(k\) in 1 : \(K_2\), where \(K_2\) is the total number of iterations to run:
  - E step:
    - Simulate \(\Omega_k\) by MH algorithm with pseudo pmf:
      - Set \(\tilde{\lambda}_i^{(0)} = \tilde{\mu}_k\);
      - for \(i\) in 1 : \(L\), where \(L\) is the total number of iterations used in the MH algorithm within each simulation step:
        - for \(i\) in 1 : \(n\)
          - (1) Draw a random variable \(\tilde{\lambda}_i^{(l)}\) from \(N(\tilde{\mu}_k, \tilde{\Sigma}_k)\);
          - (2) Calculate Equation (12) with pseudo probability: generate a random number \(u\) that is uniformly distributed with support \([0, 1]\); if \(\tau(\tilde{\lambda}_i^{(l)}; \lambda_i^{(l-1)}) > u\), \(\lambda_i^{(l)} = \tilde{\lambda}_i^{(l)}\), else \(\lambda_i^{(l)} = \lambda_i^{(l-1)}\);
          - Set \(\lambda_i = \{\lambda_{i,1}, \ldots, \lambda_{i,L}\}\);
        - M step:
          - Calculate \(\theta\) by Equation (14) and update \(\hat{\theta}_{k+1} = \theta\).

4. Simulation study

A simulation study was conducted to investigate the performance of the proposed model parameter estimation method developed in Section 3. In the simulation study, we generated a sample set \(M_n\) containing \(n\) microstructure samples based on the proposed random effect autologistic regression model (4), given the simulation model parameters \(\theta_s = \{\mu_s, \Sigma_s\}\). Then, based on the simulated sample set \(M_n\), the ALSAEM method developed in Section 3 was applied to estimate the parameters of the model, whose values were compared to simulation parameters \(\theta_s\) to assess the performance of the developed method.

To simulate microstructure sample set \(M_n\), we first generated a set of \(\omega_s\), denoted by \(\Omega_s = \{\omega_s, \ldots, \omega_{s,n}\}\), from the multivariate normal distribution \(N(\mu_s, \Sigma_s)\). Next, given each \(\omega_s\), we developed an MCMC method that is similar to the procedure proposed by Cross and Jain (1983), to generate samples
of microstructure images. The main idea behind this algorithm is to increase the likelihood function of the lattice's realization. Specifically, in the MCMC simulation method, in each of the iterations, we randomly chose one site and reversed its value. We accepted the reversal with a probability equal to the ratio of probabilities of the lattice that were calculated using Equation (5) before the reversal and that after the reversal. The details of the simulation process are described in the Algorithm 2:

Algorithm 2:
Given the parameter $\theta_i = \{\mu, \Sigma\}$, draw a set of random variables $\omega_i; i = 1, \ldots, n_s$ from $N(\mu, \Sigma)$, and let $\Omega_i = \{\omega_{i1}, \ldots, \omega_{i}n_s\}$;
for $i$ in 1 : $n_s$;
Initialize the lattice $X_i^{(0)}$ with random variables 0 and 1;
for $l$ in 1 : $L_i$, where $L_i$ is the total number of iterations to generate each sample in the MCMC method:
randomly select a site $X_{ij}^{(l)}$; set $\tilde{X}_{ij}^{(l)} = X_{ij}^{(l-1)}$ and $\tilde{X}_{ij}^{(l)} = 1 - X_{ij}^{(l-1)}$
calculate $\Delta = Q(X_{ij}^{(l)}|\omega_{i}), Q(X_{ij}^{(l-1)}|\omega_{i});$
generate a random number $u$ that is uniformly distributed with support [0, 1];
if $\exp(\Delta) > u: X_{ij}^{(l)} = \tilde{X}_{ij}^{(l)}$; else $X_{ij}^{(l)} = X_{ij}^{(l-1)}$
Set $\mathcal{M}_i = \{X_i^{(1)}, \ldots, X_i^{(L_i)}\}$

In the simulation study, the neighbor set $N(s)$ is chosen to contain four nearest sites from site $s$ on the lattice; i.e., $N(s) = \{t_u, t_d, t_l, t_r\}$ (see Fig. 2) for simplicity. The selection of four neighboring sites, which are called direct neighbors, is one of the two neighbor definitions that are widely used in the research fields of image processing and computer graphics (Gonzalez and Woods, 2002). The other widely used neighbor definition is eight-connected neighbors; i.e., $N_i(s) = \{t_u, t_d, t_l, t_r, t_u, t_d, t_l, t_r\}$. In addition, we assumed that the microstructure was homogeneous, so that $t_u$ and $t_d$ shared the same coefficient $\omega_{i1}$, and $t_l$ and $t_r$ shared the same coefficient $\omega_{i2}$. Thus, $\omega_i$ consisted of three variables; i.e., $\omega_i = (\omega_{i0}, \omega_{i1}, \omega_{i2})^T$, which follows a multivariate normal distribution $N(\mu, \Sigma)$. To demonstrate the generality of the proposed method, the following arbitrary parameters $\mu$ and $\Sigma$ were applied in the simulation study:

$$\mu = [3, -2.4, 1]^T, \quad \Sigma = \begin{bmatrix} 0.7 & 0.4 & -0.3 \\ 0.4 & 0.6 & 0.0 \\ -0.3 & 0.0 & 0.5 \end{bmatrix}.$$  

We considered different sample sizes—i.e., $n_s = 30, 100, 300,$ and 600—and different dimensions of microstructure samples; i.e., $d_1 \times d_2 = 32 \times 32, 64 \times 64, 128 \times 128,$ and $192 \times 192$. In addition, we set $L_i = 2 \times 10^6$ in Algorithm 2 so that after $L_i$ iterations the log-likelihood value of the simulated samples reached the maximum values and for the remaining periods it was stable.
Based on the simulated samples, the model parameters were estimated using Algorithm 1. We considered different values of the parameter $K_1$ in Algorithm 1; i.e., $K_1 = 100$ and 500. In addition, we set $K_2 = 1500$, and $L = 6000$. As for the initial parameters $\theta_1$, we randomly chose 20 microstructure samples and applied the traditional autologistic regression model to estimate the parameters of each individual sample. Then we calculated the sample mean and the sample covariance matrix of the estimated parameters, which were used as the initial parameters $\hat{\mu}_1$ and $\hat{\Sigma}_1$, respectively.

Although the proposed parameter estimation method based on the SAEM algorithm can overcome the computational complexity of the MLE method, it still takes a long time. Specifically, a single replication can take several days using a high-performance grid-enabled computing system with 120 CPU cores. Thus, it is difficult, if not impossible, to conduct a simulation study by repeating multiple replications. Due to the time limitation, we only implemented the simulation with a single replication for each simulation scenario. The estimated parameters vs. the iteration numbers for each scenario are plotted in the supplement.

Based on simulation results, we compared the sum of squared errors for the estimated parameters for different image dimensions and different sample sizes. We show plots the sum of squared errors in Fig. 5 and Fig. 6, for $K_1 = 100$ and $K_1 = 500$, respectively.

In Fig. 5 and Fig. 6, it can be seen that there is a trend that the sum of squared errors decreases with an increase in the sample size, although a strictly monotone pattern does not exist. As a smaller sum of squared errors indicates a better estimation performance, a large sample size is preferred for the model parameter estimation. In contrast, the sample dimension does not significantly affect the estimation performance. In addition, it can be seen that the estimation performance of the mean parameters is better than that of the covariance parameters.

The parameter $K_1$ is traditionally used in the SAEM algorithm to control the convergence rate (Lavielle, 2014). In essence, the parameter $K_1$ determines at which iteration the stochastic approximation becomes effective. That is, the stochastic approximation starts to work only after $K_1$ iterations. Thus, if the initial values are far from the real parameters, a large value of $K_1$ is needed so that a long “burn in” period is performed to make the estimated parameters close to the real parameters; and after that period, the stochastic approximated parameters can quickly converge. In Fig. 5 and Fig. 6, it can be seen that the estimation performances do not change significantly for $K_1 = 100$ and 500. This is due to the point that the initial parameters we found are close to the real parameters, so the effect of stochastic approximation is not significant.

5. Case study

The developed method was used to study multiple microstructure samples obtained for a dual-phase high-strength steel. Due to their exceptional properties, such as high-strength values and good formability, dual-phase high-strength steels are being increasingly used in automotive industry.
In the case study, 56 samples were obtained from the same batch of manufactured steel products. The steel samples were prepared using several processing steps including grinding, etching, and polishing. After preparation, the microstructures of the steel samples were obtained using a microscope with a 100x magnification. For each sample, we obtained a microstructure
image with dimensions of $200 \times 200$ pixels. Figure 7 shows two images of microstructure samples from the total of 56 samples.

It can be seen from Fig. 7 that significant unit-to-unit variability exists among the microstructure samples.

It is reasonable to assume that the high-strength steel materials used in the case study are homogeneous. The materials, however, were manufactured using a rolling process, in which the microstructural particles are elongated in the rolling direction. Thus, the microstructure images need to be considered as being anisotropic. It should be noted that not all dual-phase steels are manufactured using a rolling process and for these dual-phase steels it is sufficient to make the isotropic assumption, which can be treated as a special case of anisotropy.

To implement Algorithm 1, we chose the eight-connection neighbor set in Equation (4); i.e., $N_i(s) = \{t_u, t_d, t_l, t_r, t_{lu}, t_{ld}, t_{ru}, t_{rd}\}$. Based on the homogeneous and anisotropic material assumption, $(t_u, t_d), (t_l, t_r), (t_{lu}, t_{rd}),$ and $(t_{ld}, t_{ru})$ were assumed to share the same coefficients $\omega_{i1}, \omega_{i2}, \omega_{i3},$ and $\omega_{i4}$, respectively. Thus, $\omega_i$ consisted of five elements, $\omega_i = (\omega_{i0}, \omega_{i1}, \omega_{i2}, \omega_{i3}, \omega_{i4})^T$.

The parameters we estimated consisted of the mean and covariance of a five-dimensional normal distribution. The parameters in Algorithm 1 were set to $K_1 = 1000$, $K_2 = 2000$, and $L = 10000$. The estimated parameters of each iteration step during the parameter estimation process are plotted in Fig. 8.

From Fig. 8, it can be seen that all the estimated parameters converge after 1000 iterations, the estimated parameters were obtained as follows:

$$\hat{\mu} = \begin{bmatrix} -8.04 & 2.26 & 3.05 & 1.42 & 1.40 \end{bmatrix}^T,$$

$$\hat{\Sigma} = \begin{bmatrix}
0.12 & -0.07 & 0.02 & 0.03 & -0.09 \\
-0.07 & 0.07 & -0.03 & -0.02 & 0.04 \\
0.02 & -0.03 & 0.05 & -0.03 & -0.02 \\
0.03 & -0.02 & -0.03 & 0.13 & -0.10 \\
-0.09 & 0.04 & -0.02 & -0.10 & 0.16 \\
\end{bmatrix}.$$
6. Conclusions

In this article, we proposed a novel random effect autologistic regression model that extends the classic autologistic regression to model the unit-to-unit variability of multiple samples with spatial binary data. We developed an MLE method to estimate the model parameters given a set of microstructure samples. However, the likelihood function of the proposed model involves a high-dimensional integral, which makes the parameter estimation challenging. In addition, the likelihood function is difficult to evaluate due to the complex computation in calculating an unknown normalization constant. To overcome these challenges, we further developed a SAEM algorithm to estimate the model parameters. We conducted a simulation study to verify the proposed methodology. As the estimated parameters are close to the simulation parameters, the proposed parameter estimation performs well. In addition, a real-world example of a dual-phase high-strength steel was used to illustrate the developed methods.

It should be noted that, for the same data set, both the classic autologistic regression model and the proposed random effect model can be applied. If researchers have a strong belief that there is some group factor among samples based on domain knowledge and that group factor is of interest, then the random effect model can be a better choice than the fixed effect model.

The microstructure of a material can strongly influence its properties such as strength, hardness, wear resistance, etc. As the proposed random effect model is able to capture the sample-to-sample variation, it can be applied to the quality control of steel products. In the developed model, the random effects \( \omega \) that capture the spatial distribution of two phases are assumed to have a multivariate normal distribution with parameters \( \theta = \{ \mu, \Sigma \} \). Given the microstructure samples of a collection of qualified steel products, the mean vector \( \hat{\mu} \) and covariance matrix \( \hat{\Sigma} \) can be estimated by using the parameter estimation methods developed in this article, based on which the control charts of random effect \( \omega \) can be established (Montgomery, 2012). The obtained control charts can be used to detect whether a new steel product is defective by analyzing its microstructures. Specifically, the traditional autologistic regression model can be applied to estimate the parameters \( \omega_{\text{new}} \) of the new microstructure sample. If the estimated parameters \( \hat{\omega}_{\text{new}} \) are outside of the control limits, it can be concluded that the new steel product is defective. A detailed procedure for steel quality control is an interesting topic for further research.

Another future topic will be to reconstruct new microstructure samples that can be used to assess a material’s macroscopic properties by applying microstructure-based finite element analysis without the need to conduct expensive physical experiments. In addition, extending the proposed method from two-phase materials to multi-phase materials is also of interest. The proposed random effect model can be extended to materials with more than two phases in a similar way; the binomial logistic regression is extended to multinomial logistic regression (Hastie et al., 2009). However, due to the complexity of the multi-phase random effect autologistic regression model, the parameter estimation method is more difficult than the existing one, and thus further investigation is needed.

Acknowledgements

The authors thank the Department Editor and the referees for their valuable comments that helped to improve this article.

Funding

The presented work was supported by the National Science Foundation under grant CMMI-1404276 to Wayne State University.

Notes on contributors

Nailong Zhang is a Ph.D. candidate in the Department of Industrial & Systems Engineering at Wayne State University. He received his B.Eng. degree in Mechanical Engineering from Harbin Institute of Technology, Harbin, China, in 2009. His research interests include statistical methods in reliability engineering as well as maintenance planning for complex systems.

Qingyu Yang received B.S. and M.S. degrees in Automatic Control and Intelligent System from the University of Science and Technology of China in 2000 and 2003, respectively; an M.S. degree in Statistics and a Ph.D. degree
in Industrial Engineering from the University of Iowa in 2007 and 2008, respectively. Currently, he is an Assistant Professor in the Department of Industrial and Systems Engineering at Wayne State University. His research interests include statistical data analysis, reliability and quality, and complex system modeling. He is a member of INFORMS and IIE.

References

Alpas, A.T. and Zhang, J. (1994) Effect of microstructure (particulate size and volume fraction) and counterface material on the sliding wear resistance of particulate-reinforced aluminum matrix composites. *Metallurgical and Materials Transactions A*, 25, 969–983.

Bilmes, J.A. (1998) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report 97-021. International Computer Science Institute, Berkeley, CA.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis–Hastings algorithm. *The American Statistician*, 49, 327–335.

Cressie, N.A.C. (1993) *Statistics for Spatial Data*, Wiley, New York, NY.

Cross, G.R. and Jain, A.K. (1983) Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5, 25–39.

Delyon, B., Lavielle, M. and Moulines, E. (1999) Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27, 94–128.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.

Feng, J.W., Li, C.F., Cen, S. and Owen, D.R.J. (2014) Statistical reconstruction of two-phase random media. *Computers & Structures*, 137, 78–92.

Gommes, C.J., Jiao, Y. and Torquato, S. (2012) Microstructural degeneracy associated with a two-point correlation function and its information content. *Physical Review E*, 85, 747–761.

Gonzalez, R.C. and Woods, R.E. (2002) *Digital Image Processing*, Prentice Hall, Upper Saddle River, NJ.

Grigoriev, M. (2003) Random field models for two-phase microstructures. *Journal of Applied Physics*, 94, 3762–3770.

Gumpertz, M.L., Graham, J.M. and Ristaino, J.B. (1997) Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 131–156.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition, Springer, New York.

Huffer, F.W. and Wu, H. (1998) Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics*, 54, 509–524.

Hughes, J., Haran, M. and Caragea, P.C. (2011) Autologistic models for binary data on a lattice. *Environmetrics*, 22, 857–871.

Irani, M. and Taberi, A.K. (2008) Effect of forging temperature on homogeneity of microstructure and hardness of precision forged steel spur gear. *Materials Chemistry and Physics*, 112, 1099–1105.

Jiao, Y., Stillinger, F.H. and Torquato, S. (2007) Modeling heterogeneous materials via two-point correlation functions: basic principles. *Physical Review E*, 76, 73–82.

Khuri, A.I. (2010) *Linear Model Methodology*, CRC Press, Boca Raton, FL.

Lavielle, M. (2014) *Mixed Effects Models for the Population Approach: Models, Tasks, Methods and Tools*, CRC Press, Boca Raton, FL.

Lewandowski, J.J., Liu, C., and Hunt Jr, W.H. (1989) Effects of matrix microstructure and particle distribution on fracture of an aluminum metal matrix composite. *Materials Science and Engineering: A*, 107, 241–255.

Li, B.L., Tsuji, N. and Kamikawa, N. (2006) Microstructure homogeneity in various metallic materials heavily deformed by accumulative roll-bonding. *Materials Science and Engineering: A*, 423, 331–342.

Montgomery, D.C. (2012) *Statistical Quality Control*, seventh edition, Wiley, New York.

Moon, S. and Russell, G.J. (2008) Predicting product purchase from inferred customer similarity: an autologistic model approach. *Management Science*, 54, 71–82.

Ostoja-Starzewski, M. (2002) Microstructural randomness versus representational volume element in thermomechanics. *Journal of Applied Mechanics*, 69, 25–35.

Paul, S.K. (2013) Real microstructure based micromechanical model to simulate microstructural level deformation behavior and failure initiation in DP 590 steel. *Materials & Design*, 44, 397–406.

Saheli, G., Garmestani, H. and Gokhale, A. (2008) Homogenization relations for elastic properties of two-phase composites using two-point statistical functions. *Journal of Mechanics of Materials and Structures*, 3, 85–106.

Torquato, S. (2002) *Random Heterogeneous Materials: Microstructure and Macroscopic Properties*, Springer, New York.

Wu, C.F.J. (1983) On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11, 95–103.

Zaefferer, S., Ohlert, J. and Bleck, W. (2004) A study of microstructure, transformation mechanisms and correlation between microstructure and mechanical properties of a low alloyed trip steel. *Acta Materialia*, 52, 2765–2778.