### COSMOS Technical Report 12-08

# Bayesian Linear Mixed-effect Modeling of Optical Profiles in Low-Emittance (Low-E) Glass Manufacturing Processes

Li Zeng Assistant Professor Department of Industrial and Manufacturing Systems Engineering The University of Texas at Arlington Arlington, TX 76019 (lzeng@uta.edu) Nan Chen Assistant Professor Department of Industrial and Systems Engineering National University of Singapore Singapore, 117576 (isecn@nus.edu.sg)

# Abstract

Low-emittance (low-E) glass manufacturing has become an important sector of the glass industry for the great economical and environmental impacts of low-E glasses. However, the quality control scheme in the current practice is rather primitive and advanced statistical quality control methods need to be developed. As an effort for this purpose, this paper considers the modeling of optical profiles which are typical quality measurements in low-E glass manufacturing processes. Linear mixed-effect models are used for the data, and a Bayesian modeling approach is proposed for parameter estimation and model selection. The effectiveness of the proposed approach is validated in a numerical study, and its use in practice is demonstrated in a case study where this approach is applied to a real dataset from a low-E glass producer. The established model using this approach will provide a foundation for quality monitoring and variation reduction in low-E glass manufacturing.

**Keywords:** Bayes factors, Linear mixed-effect (LME) models, Gibbs sampling, Low-E glass, Optical profiles, Polynomial models

## **1. Introduction**

The concerns on energy and environment during the past two decades have led to fundamental improvements in house/building design which are partially made possible by many technological innovations on energy efficiency of windows. One critical innovation is the *low*-

*emittance (low-E) glass*. The low-E glass is manufactured through physical or chemical coating processes, as illustrated in Figure 1, where solid materials, e.g., metal, metal oxide, and metal nitride, are deposited on the surface of flat glass ribbons in a sequence of sealed chambers to enhance the thermal/optical performance of products. With the depositions, such glasses can lower the heat flow through the window by reflecting up to 90% of infrared radiation while allowing visible light to enter (Arasteh *et al.*, 2004; Carmody, Selkowitz and Heschhong, 1996). Thus, they are able to reduce unwanted heat gain in summer and heat loss in winter. Studies suggest that if all windows purchased over the next fifteen years are made by low-E glass and incorporated with other readily available efficiency improvements, the collective annual energy bill of US could be reduced by 25 percent or over 2 billion per year by 2010 (Frost, Arasteh and Eto, 1993). The huge benefits have made low-E glass manufacturing an important sector of the glass industry.



Figure 1. Low-E glass manufacturing process and optical profiles

One main quality concern in low-E glass manufacturing processes is the uniformity of coating on the glass surface. In the current practice of quality control in those processes, automated online inspection is typically conducted where the finished product is scanned by laser to measure optical properties of sampled locations on the glass. One example of the measured optical properties is the spectral reflectance profile, i.e., the percentage of light that reflects from the glass surface over a range of wavelengths, as shown in the lower panel of Figure 1, where  $\lambda$  is the wavelength and r is the reflectance. From such optical profiles, popular

color scales, such as a<sup>\*</sup> (degree of redness/greenness) and b\*(degree of yellowness/blueness) (Hunter Lab, 2008), are calculated and monitored as quality metrics. However, the color scales are simple, univariate summaries of the information in the optical profiles, which are not adequate to reflect the state of the process and thus may not perform well in detecting changes in the process. Obviously, the optical profiles need to be monitored directly for achieving more reliable quality control.

Monitoring of quality profiles has received much attention in recent years due to the increasing popularity of such types of data in manufacturing processes. Woodall (2004, 2007) and Noorossana *et al.* (2011) give excellent reviews of the state of art in this field. The basic idea for profile monitoring is first building a statistical model to characterize the shape of the profiles, and then monitoring the parameter estimates of the model. The modeling step is very critical as it establishes a foundation for the monitoring step as well as other statistical analyses of the data. For this reason, this study will focus on modeling the optical profiles as our starting effort to advance quality control schemes in the low-E manufacturing processes.

Concerning modeling of quality profiles, various linear models (e.g., Kim, Mahmoud and Woodall, 2003) and nonlinear models (e.g., Williams, Woodall and Birch, 2007) have been proposed in the literature, depending on the characteristics of data. The model that fits the optical profiles in Figure 1 is polynomial models. Such models have been used in existing studies. For example, Kazemzadeh, Noorossana and Amiri (2008, 2009) use regular polynomial models and evaluate their performance in Phase I and Phase II profile monitoring through simulations; Kazemzadeh, Noorossana and Amiri (2010) consider correlated errors in polynomial modeling to characterize between-profile autocorrelation; and Amiri, Jensen and Kazemzadeh (2009) conduct a case study in which linear mixed-effect (LME) models initially proposed by Jensen, Birch and Woodall, 2008) for linear profiles are applied to polynomial profiles from automotive industry.

The LME models is very popular in areas such as social and medical research where the variation in individuals is typically modeled as random effects (Gelman and Hill, 2007; Wu, 2010). The motivation for its use in profile modeling is mainly to characterize within-profile

correlations that are commonly present in profile data (Jensen, Birch and Woodall, 2008). Another advantageous feature of the LME models that is not emphasized in the existing studies is its decomposition of variations in the data into two parts, the variation in the coefficients of the model and the variation in the random errors. The former can be viewed as representing the intrinsic variation of the manufacturing process, while the latter represents noises in the process and data collection such as measurement errors. This decomposition can provide more information for process diagnosis and control than regular models. The LME model is especially suitable for cases where the variation in the coefficients is predominant, i.e., smooth profiles with small random errors, which is the exact characteristic of the optical profiles from the low-E glass manufacturing processes as shown in Figure 1.

However, the current methods for fitting LME models for profile data bear two important limitations: (*i*) The restricted maximum likelihood (REML) method, which is commonly used for parameter estimation in LME modeling, has issues in handling some special problems that may arise in practice. For example, the iterative algorithm for finding REML estimates may not converge when some variance components are small, and/or the model is misspecified (Jensen, Birch and Woodall, 2008), and this method may not work when sample size is small. (*ii*) As a typical practice in regression modeling, model selection needs to be conducted, i.e., determine which terms in a polynomial model have random effects, to find the best model for the data. In the existing studies, this is either not considered or done by subjective judgment. A formal model selection procedure is needed, which is, however, a very challenging task using conventional methods due to the complexity of the LME model.

To conquer the above limitations, this paper proposes a Bayesian LME modeling approach for polynomial profiles. Aiming to solve the issues in building LME models in a unified framework, this approach consists of two components: a Bayesian procedure for parameter estimation and a procedure based on the Bayes factor for model selection. It is advantageous in the following aspects: First, Bayesian approaches are powerful in dealing with small sample sizes, and the Bayesian way for parameter estimation is more stable than the REML methods as estimation is conducted through simulation instead of optimization. Second, the model selection procedure is conceptually easy and can compare any two LME models. Finally, in the unified framework of the proposed approach, parameter estimation and model selection will be conducted simultaneously based on one set of posterior samples of parameters, which provide great convenience to practitioners. It is also worth mentioning that although the proposed approach will be demonstrated using polynomial profiles in this paper, they can also be used to model linear profiles as polynomial profiles is essentially a special case of liner profiles.

The remainder of the paper is organized as follows. Section 2 presents the basics of the LME model and problem formulation. The procedure for parameter estimation is given in Section 3, and the procedure for model selection is given in Section 4. Section 5 reports the results of a numerical study to demonstrate the effectiveness of the proposed approach. The results of a case study are given in Section 6 where this approach is applied to a real dataset from a low-E glass producer. Finally, Section 7 concludes the paper and discusses future work.

### 2. The Linear Mixed-effect Model and Problem Formulation

#### 2.1. The linear mixed-effect model

Let x be the wavelength, and y be the corresponding response in an optical profile. Assume a dataset contains m profiles, and each profile contains n sampling points. The general LME model with p-degree polynomials is as follows

$$y_{ij} = \beta_p x_j^p + \dots + \beta_k x_j^k + \dots + \beta_0 + \alpha_{i,p} x_j^p + \dots + \alpha_{i,k} x_j^k + \dots + \alpha_{i,0} + \varepsilon_{ij}$$
(1)

where i=1,...,m is the index of profiles, j=1,...,n is the index of sampling points, and k=0,...,p is the index of the degree of polynomials. Like in a regular linear regression model,  $\varepsilon$  is the random error which is commonly assumed to follow a Normal distribution with variance  $\sigma_{\varepsilon}^2$ 

$$\mathcal{E}_{ij} \sim N(0, \sigma_{\varepsilon}^2) \tag{2}$$

But unlike the regular linear model, there are two sets of coefficients:  $\beta_p,...,\beta_0$  are fixed, unknown values, called *fixed effects*, while  $\alpha_{i,p},...,\alpha_{i,0}$  are random variables, called *random*  *effects*, which vary from profile to profile. The random effects are typically assumed to be independently and normally distributed

$$\alpha_{i,p} \sim N(0, \sigma_p^2), ..., \alpha_{i,0} \sim N(0, \sigma_0^2)$$
 (3)

where  $\sigma_p^2, ..., \sigma_0^2$  are their variances. The variances in (2) and (3), i.e.,  $\sigma_p^2, ..., \sigma_0^2, \sigma_{\varepsilon}^2$  are also called the *variance components* in the model. As mentioned in the Introduction, the terms in the LME model bear an intuitive interpretation: the fixed effects,  $\beta$ s, represent the base-line of the *x*-*y* relationship, the random effects,  $\alpha$ s, represent the intrinsic variation of the manufacturing process, and the random error  $\varepsilon$  represents the noises in the process and data collection.

For convenience, the matrix form of the model  $(1)\sim(3)$  is given below, which will be used throughout the following discussions for convenience:

$$\mathbf{y}_i = \mathbf{X}(\boldsymbol{\beta} + \boldsymbol{\alpha}_i) + \boldsymbol{\varepsilon}_i, \qquad i = 1,...,m$$
(4)

where

$$\mathbf{y}_{i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \dots \\ y_{in} \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} x_{1}^{p} & x_{1}^{p-1} & \dots & x_{1}^{0} \\ x_{2}^{p} & x_{2}^{p-1} & \dots & x_{2}^{0} \\ \dots & \dots & \dots & \dots \\ x_{n}^{p} & x_{n}^{p-1} & \dots & x_{n}^{0} \end{bmatrix}, \ \mathbf{\beta} = \begin{bmatrix} \boldsymbol{\beta}_{p} \\ \boldsymbol{\beta}_{p-1} \\ \dots \\ \boldsymbol{\beta}_{0} \end{bmatrix}$$
$$\mathbf{a}_{i} = \begin{bmatrix} \alpha_{i,p} \\ \alpha_{i,p-1} \\ \dots \\ \alpha_{i,0} \end{bmatrix} \sim N(0, \mathbf{\Sigma}_{\alpha}), \ \mathbf{\Sigma}_{\alpha} = \begin{bmatrix} \sigma_{p}^{2} & 0 & \dots & 0 \\ 0 & \sigma_{p-1}^{2} & \dots & 0 \\ \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & \sigma_{0}^{2} \end{bmatrix}$$
$$\mathbf{\varepsilon}_{i} = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \dots \\ \varepsilon_{in} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma}_{\varepsilon}), \ \mathbf{\Sigma}_{\varepsilon} = \sigma_{\varepsilon}^{2} \mathbf{I}_{n}$$

### 2.2 Problem formulation

There are two tasks involved in building the LME model: (*i*) *Parameter estimation under a* given model: the parameters of interest in most applications include the fixed effects  $\beta_p,...,\beta_0$  and the variance components  $\sigma_p^2,...,\sigma_0^2,\sigma_{\varepsilon}^2$  (the random effects are treated as nuisance parameters). Estimates of these parameters need to be found. (*ii*) *Model selection:* under the

same degree of polynomials, alternative LME models differ in the polynomial terms that have random effects. Equation (1) is actually a full model where all polynomial terms have random effects. In practice, it is possible that some terms may have no or negligible random effects. For example, a 2-degree polynomial model may only have random effect on the quadratic term or the linear term or the intercept. Figure 3 in Section 5 shows simulated data from such models. Thus, the alternative models need to be compared to determine the best model for the data.

### 3. Bayesian Approach for Parameter Estimation

This section presents a Bayesian approach for estimating the parameters of the full model in (4). Results of other alternative models can be obtained similarly by simply setting the random effects of some polynomial terms to be zero.

Following the standard practice of Bayesian statistics, the central task in parameter estimation is to find the posterior distribution of the parameters. In the LME model, which is essentially a hierarchical model, the posterior of the parameters of interest, i.e., the fixed effects and variance components, is not directly available due to the existence of the nuisance parameters, i.e., the random effects, in the model. Consequently, we need to find the joint posterior of all the parameters

$$P(\boldsymbol{\beta}, \{\sigma_p^2, ..., \sigma_0^2\}, \sigma_{\varepsilon}^2, \{\boldsymbol{\alpha}_i, i = 1, ..., m\} | \{\boldsymbol{y}_1, ..., \boldsymbol{y}_m\})$$
(5)

to obtain the (marginal) posterior of the parameters of interest. As the joint posterior is not analytically tractable, Markov chain Monte Carlo (MCMC) algorithms (Robert and Casella, 2004) will be used to generate a set of samples from it. Location estimates of these samples, such as mean and mode, will be used as point estimates of the parameters of interest.

However, sampling from the joint posterior in (5) is very challenging for the high dimension of the parameter space. A popular way to solve this problem is to use the *Gibbs sampler*, which is a powerful tool in sampling hierarchical, high-dimensional posteriors (Gelman, 2004). The basic idea of Gibbs sampling is to divide the parameter set into subgroups and sample from the conditional posterior distribution of each subgroup given all other parameters. As the subgroups have smaller dimensions and their conditional posteriors are usually much simpler, the sampling can be realized using conventional MC or MCMC algorithms. Hobert and Casella (1996) propose a Gibbs sampling procedure for estimating LME models without considering the model selection problem. Our proposed procedure described below modifies their procedure in two aspects to fit our problem: considerations for model selection are integrated into the sampling; and an alternative sampling strategy is provided which can solve numerical problems that may occur in implementing the procedure.

### Proposed Gibbs Sampling Procedure

The parameters of model (4) can be divided into 3 subgroups:

- **G1**: The random effects  $\{\alpha_i, i = 1, ..., m\}$
- **G2**: The variance components of the random effects  $\sigma_p^2,...,\sigma_0^2$
- **G3**: The fixed effects and variance of the random error  $\beta$ ,  $\sigma_{\epsilon}^2$

Samples of them will be generated through the following steps:

Step 1: Sampling parameters in G1 from their conditional posterior distribution

$$P(\{\boldsymbol{\alpha}_i, i=1,...,m\} \mid \boldsymbol{\Sigma}_{\alpha}, \boldsymbol{\beta}, \sigma_{\varepsilon}^2, \{\mathbf{y}_1,...,\mathbf{y}_m\})$$

Step 2. Sampling parameters in G2 from their conditional posterior distribution

$$P(\sigma_p^2,...,\sigma_0^2 \mid \{\mathbf{a}_i, i = 1,...,m\}, \mathbf{\beta}, \sigma_{\varepsilon}^2, \{\mathbf{y}_1,...,\mathbf{y}_m\}) = P(\sigma_p^2,...,\sigma_0^2 \mid \{\mathbf{a}_i, i = 1,...,m\})$$

Step 3: Sampling parameters in G3 from their conditional posterior distribution

$$P(\boldsymbol{\beta}, \sigma_{\varepsilon}^{2} | \{\boldsymbol{\alpha}_{i}, i = 1, ..., m\}, \boldsymbol{\Sigma}_{\alpha}, \{\boldsymbol{y}_{1}, ..., \boldsymbol{y}_{m}\}) = P(\boldsymbol{\beta}, \sigma_{\varepsilon}^{2} | \{\boldsymbol{\alpha}_{i}, i = 1, ..., m\}, \{\boldsymbol{y}_{1}, ..., \boldsymbol{y}_{m}\})$$

By iterating the above steps, a sequence of samples will be obtained, which follow the joint posterior in (5). About the notations of the parameters in G2, note that when they are treated as random variables like in Step 2, we use the individual form, i.e.,  $\sigma_p^2,...,\sigma_0^2$ , since they will be sampled separately; when they are treated as given information like in Step 1 and 3, we use the matrix form  $\Sigma_{\alpha}$  for simplicity. This will be followed throughout the following discussions. It is also worth mentioning that the equality in Step 2 achieves because according to Eq. (3), given the random effects, their variance components are independent of other parameters. The equality in Step 3 achieves for a similar reason. There are two issues in conducting the sampling in each step of the procedure: specifying the priors and finding the conditional posterior distribution. Solutions of them will be given in Section 3.1 and 3.2 respectively. Detailed steps of the Gibbs sampling procedure are summarized in Appendix I.

# **3.1 Specification of priors**

The parameters of the LME model fall into two categories: mean parameters, i.e., the fixed effects and random effects, and variance parameters, i.e., the variance components, priors for these two types of parameters need to be specified. To enable objective inference as well as provide convenience to data analysis in practice, noninformative or weakly informative priors will be used. The priors are denoted as  $\pi(\cdot)$  in the following.

For the fixed effects  $\beta$ , a uniform prior will be used, i.e.,

$$\pi(\boldsymbol{\beta}) \propto 1 \tag{6}$$

The priors of the random effects are just the normal distributions in Eq. (3).

The specification of priors for the variance parameters needs more caution as inappropriate setting of these parameters in Gibbs sampling will lead to noninformative posteriors which makes the inference invalid (Hobert and Casella, 1996). In general, two types of priors have been commonly used for variance parameters in the literature (Gelman, 2006):

$$\pi_1(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{r+1} \qquad \pi_2(\sigma^2) \propto \text{IG}(\omega, \omega)$$

where the first one is a noninformative prior, and *r* is a user specified value. For example, r = -1 leads to the uniform prior on  $\sigma^2$ , i.e.,  $\pi(\sigma^2) \propto 1$ , r = -0.5 leads to the uniform prior on  $\sigma$ , i.e.,  $\pi(\sigma^2) \propto 1/\sigma$ . It is worth mentioning that the popular prior  $\pi(\sigma^2) \propto 1/\sigma^2$  corresponding to r = 0 is not appropriate here as it will yield noninformative posteriors. The second one is a weakly informative prior, where IG(·) represents Inverse-gamma distribution, and  $\omega$  is a small value such as 0.01 or 0.001.

In our approach, the uniform prior on  $\sigma^2$  will be used for the variance of the random error,

$$\pi(\sigma_{\varepsilon}^2) \propto 1 \tag{7}$$

for its simplicity, and the weakly informative priors will be used for the variance components of the random effects

$$\pi(\sigma_k^2) \propto \text{IG}(\omega, \omega) \quad k = p, ..., 0 \tag{8}$$

to facilitate the computation in model selection. More details of this will be given in Section 4.

# 3.2 Conditional posterior distributions

(1)  $P(\{\boldsymbol{\alpha}_i, i=1,...,m\} | \boldsymbol{\Sigma}_{\alpha}, \boldsymbol{\beta}, \sigma_{\varepsilon}^2, \{\mathbf{y}_1,...,\mathbf{y}_m\})$ 

Given other parameters, the random effects of different profiles are independent, so they can be sampled separately, that is, we will sample from  $P(\boldsymbol{\alpha}_i | \boldsymbol{\Sigma}_{\alpha}, \boldsymbol{\beta}, \sigma_{\varepsilon}^2, \mathbf{y}_i)$ , *i*=1,...,*m*. By (3) and (4), this distribution is

$$P(\mathbf{a}_{i} | \mathbf{\Sigma}_{\alpha}, \mathbf{\beta}, \sigma_{\varepsilon}^{2}, \mathbf{y}_{i}) \propto \pi(\mathbf{a}_{i} | \mathbf{\Sigma}_{\alpha}) \cdot P(\mathbf{y}_{i} | \mathbf{a}_{i}, \mathbf{\beta}, \sigma_{\varepsilon}^{2})$$

$$= \prod_{k=0}^{p} \pi(\alpha_{i,k} | \sigma_{k}^{2}) \cdot N(\mathbf{y}_{i} | \mathbf{X}(\mathbf{\beta} + \mathbf{a}_{i}), \sigma_{\varepsilon}^{2} \mathbf{I}_{n})$$

$$= \prod_{k=0}^{p} N(\alpha_{i,k} | 0, \sigma_{k}^{2}) \cdot N(\mathbf{y}_{i} - \mathbf{X}\mathbf{\beta} | \mathbf{X}\mathbf{a}_{i}, \sigma_{\varepsilon}^{2} \mathbf{I}_{n})$$

which is actually a normal distribution (Hobert and Casella, 1996)

$$\boldsymbol{\alpha}_{i} \mid \boldsymbol{\Sigma}_{\alpha}, \boldsymbol{\beta}, \sigma_{\varepsilon}^{2}, \boldsymbol{y}_{i} \sim N\left( (\mathbf{X}'\mathbf{X} + \sigma_{\varepsilon}^{2}\boldsymbol{\Sigma}_{\alpha}^{-1})^{-1} (\mathbf{X}'(\boldsymbol{y}_{i} - \mathbf{X}\boldsymbol{\beta})), \sigma_{\varepsilon}^{2} (\mathbf{X}'\mathbf{X} + \sigma_{\varepsilon}^{2}\boldsymbol{\Sigma}_{\alpha}^{-1})^{-1} \right)$$
(9)

Note that using (9) means that the vector  $\mathbf{a}_i = [\alpha_{i,p},...,\alpha_{i,0}]'$  will be sampled simultaneously. This sampling strategy is convenient, but may be subject to the problem of instability caused by the matrix inversion operations in calculating the mean and variance of the multivariate normal distribution in (9). When some variance components of the random effects are very small, the term  $\Sigma_{\alpha}^{-1}$  will have some very large and small elements, making it close to singular or badly scaled. As a result, the inversion operation of the term  $\mathbf{X}'\mathbf{X} + \sigma_c^2 \Sigma_{\alpha}^{-1}$  will be impossible or unreliable. This is a common numerical problem encountered in matrix computation (Dongarra, Moler and Wilkinson, 1983), and occurred frequently in our simulation studies. It is especially serious at the beginning of the Gibbs sampling if the starting values are far from the true values of the parameters. A natural idea to solve this problem is to avoid manipulating matrices by sampling each element in  $\mathbf{a}_i$ , i.e.,  $\alpha_{i,k}$ , k=0,...,p, separately. That means we need to find the conditional posterior of  $\alpha_{i,k}$  given other elements in  $\mathbf{a}_i$ . This posterior is found to be

$$\alpha_{i,k} \mid \boldsymbol{\alpha}_{i}^{(-k)}, \boldsymbol{\Sigma}_{\alpha}, \boldsymbol{\beta}, \sigma_{\varepsilon}^{2}, \mathbf{y}_{i} \sim N \begin{pmatrix} \sum_{j=1}^{n} (x_{j}^{k})^{2} z_{ij} / \sigma_{\varepsilon}^{2} & 1 \\ \frac{1}{\sigma_{k}^{2}} + \frac{\mathbf{x}^{k} \cdot \mathbf{x}^{k}}{\sigma_{\varepsilon}^{2}} & \frac{1}{\sigma_{k}^{2}} + \frac{\mathbf{x}^{k} \cdot \mathbf{x}^{k}}{\sigma_{\varepsilon}^{2}} \end{pmatrix} \quad k = 0, ..., p$$
(10)

where  $\mathbf{\alpha}_{i}^{(-k)}$  is the vector of  $\mathbf{\alpha}_{i}$  with  $\alpha_{i,k}$  removed,  $\mathbf{x}^{k} = [x_{1}^{k}, x_{2}^{k}, ..., x_{n}^{k}]'$  is the column in **X** corresponding to  $\alpha_{i,k}$ ,  $\mathbf{X}^{(-k)}$  is **X** with this column removed, and  $\mathbf{z}_{i} = \mathbf{y}_{i} - (\mathbf{X}\mathbf{\beta} + \mathbf{X}^{(-k)}\mathbf{\alpha}_{i}^{(-k)})$ . The proof of (10) is given in Appendix II.

(2)  $P(\sigma_p^2,...,\sigma_0^2 | \{ \boldsymbol{\alpha}_i, i = 1,...,m \})$ 

Since the random effects are independent, their variances can be sampled separately, that is, we will sample  $\sigma_k^2 \sim P(\sigma_k^2 | \{\alpha_{i,k}, i = 1, ..., m\})$ , k=p,...,0. Under the weakly informative prior in (8), this distribution is

$$P(\sigma_k^2 | \{\alpha_{i,k}, i = 1, ..., m\}) \propto \pi(\sigma_k^2) \cdot P(\{\alpha_{i,k}, i = 1, ..., m\} | \sigma_k^2)$$
$$= IG(\omega, \omega) \cdot \prod_{i=1}^m N(\alpha_{i,k} | \sigma_k^2)$$

Taking the conjugacy of the prior into account, this distribution is also an Inverse-gamma distribution (Gelman, 2006)

$$\sigma_k^2 | \{ \alpha_{i,k}, i = 1, ..., m \} \rangle \sim \text{IG}\left( \omega + \frac{m}{2}, \ \omega + \frac{1}{2} \sum_{i=1}^m \alpha_{i,k}^2 \right)$$
(11)

(3)  $P(\boldsymbol{\beta}, \sigma_{\varepsilon}^2 | \{\boldsymbol{\alpha}_i, i = 1, ..., m\}, \{\mathbf{y}_1, ..., \mathbf{y}_m\})$ 

If the random effects  $\{\mathbf{\alpha}_i, i = 1,...,m\}$  are known, we can remove them from the model in (4) by subtracting them from the response  $\{\mathbf{y}_1,...,\mathbf{y}_m\}$ . In this way the LME model degenerates to a conventional linear regression model. Consequently, the problem here becomes to find the posterior of a linear model, which is a well studied topic known as *Bayesian linear regression models* in the literature. Typically, decomposition will be made to the posterior,

$$P(\boldsymbol{\beta}, \sigma_{\varepsilon}^{2} | \{\boldsymbol{\alpha}_{i}, i = 1, ..., m\}, \{\mathbf{y}_{1}, ..., \mathbf{y}_{m}\}) = P(\sigma_{\varepsilon}^{2} | \{\boldsymbol{\alpha}_{i}, i = 1, ..., m\}, \{\mathbf{y}_{1}, ..., \mathbf{y}_{m}\})$$
$$\times P(\boldsymbol{\beta} | \sigma_{\varepsilon}^{2}, \{\boldsymbol{\alpha}_{i}, i = 1, ..., m\}, \{\mathbf{y}_{1}, ..., \mathbf{y}_{m}\})$$

that is, the joint posterior can be decomposed as the product of the marginal posterior of  $\sigma_{\varepsilon}^2$  and the conditional posterior of  $\beta$  given  $\sigma_{\varepsilon}^2$ . A well-known result is (Gelman, 2004)

$$\sigma_{\varepsilon}^{2} | \{ \boldsymbol{\alpha}_{i}, i = 1, ..., m \}, \{ \mathbf{y}_{1}, ..., \mathbf{y}_{m} \} \sim \mathrm{IG} \left( \frac{mn-2}{2}, \frac{\left( \mathbf{y} - \Xi(\hat{\boldsymbol{\beta}} + \boldsymbol{\alpha}) \right)' \left( \mathbf{y} - \Xi(\hat{\boldsymbol{\beta}} + \boldsymbol{\alpha}) \right)}{2} \right)$$
(12)  
$$\boldsymbol{\beta} | \sigma_{\varepsilon}^{2}, \{ \boldsymbol{\alpha}_{i}, i = 1, ..., m \}, \{ \mathbf{y}_{1}, ..., \mathbf{y}_{m} \} \sim N \left( \hat{\boldsymbol{\beta}}, \ \sigma_{\varepsilon}^{2} (\Xi' \Xi)^{-1} \right)$$

where  $\hat{\boldsymbol{\beta}} = (\boldsymbol{\Xi}'\boldsymbol{\Xi})^{-1}\boldsymbol{\Xi}'(\mathbf{y} - \boldsymbol{\Xi}\boldsymbol{\alpha})$ , **y** is the vector stack of  $\mathbf{y}_1, \dots, \mathbf{y}_m$ ,  $\boldsymbol{\Xi}$  is the stack of  $\mathbf{X}$ , and  $\boldsymbol{\alpha}$  is the stack of  $\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_m$ .

### 4. Model Selection Using the Bayes Factor

In a full model with *p*-degree polynomials, the pool of candidate models contains  $2^{p+1}$  models, denoted as  $\{M_1, M_2, ..., M_{2^{p+1}}\}$ . Comparison of these models is a challenging problem which is difficult to solve using conventional methods such as likelihood ratio tests (Pauler, Wakefield and Kass, 1999). Bayesian statistics provides a powerful tool for this problem, the *Bayes factor* (BF) (Gelfand and Dey, 1994; Kass and Raftery, 1995), which is able to compare any models regardless of their forms. However, the computation of Bayes factors is not a trivial problem, and many methods have been developed for it. A convenient method proposed by Chib (1995), called Chib's method later, will be adapted to our problem. The definition of the Bayes factor and a brief review of computation methods will be given in Section 4.1, and details of the proposed model selection procedure will be given in Section 4.2.

#### 4.1. The Bayes factor and computation methods

The Bayes factor of two competing models,  $M_i$  and  $M_j$ ,  $i \neq j$ , is defined by (Gelman, 2004)

$$BF_{ij} = \frac{P(\mathbf{y} \mid M_i)}{P(\mathbf{y} \mid M_j)} = \frac{\int \pi(\mathbf{\theta}_i \mid M_i) P(\mathbf{y} \mid \mathbf{\theta}_i, M_i) d\mathbf{\theta}_i}{\int \pi(\mathbf{\theta}_j \mid M_j) P(\mathbf{y} \mid \mathbf{\theta}_j, M_j) d\mathbf{\theta}_j}$$
(13)

where **y** is the data, and  $\boldsymbol{\theta}_i$  is the parameter vector of model  $M_i$ . For example, if  $M_i$  is the full model in (4), the parameter vector is  $\boldsymbol{\theta}_i = [\boldsymbol{\beta}, \sigma_p^2, ..., \sigma_0^2, \sigma_{\varepsilon}^2]'$ . The parameter vector of other candidate models is just a subset of it.  $P(\mathbf{y}|M_i)$  is the likelihood of data under model  $M_i$ , called *marginal likelihood* (ML) in the literature,  $\pi(\boldsymbol{\theta}_i|M_i)$  is the corresponding prior, and  $P(\mathbf{y}|\boldsymbol{\theta}_i, M_i)$  is the corresponding sampling density, or likelihood. Essentially, using the marginal likelihood as a measure of model strength, the Bayes factor is the Bayesian version of likelihood ratio, except that the marginal likelihood is obtained by integrating, instead of maximizing, over the parameter space. Intuitively, when the Bayes factor is larger than 1, it means that the data support  $M_i$  more than  $M_j$ , and thus  $M_i$  should be selected. A set of cutoff values have been suggested in the literature and widely used as guidelines in the applications of this tool (Kass and Raftery, 1995),

BF	2log(BF)	Evidence against $M_i$
1~3	0~2	Barely worth mentioning
3~20	2~6	Positive
20~150	6~10	Strong
>150	>10	Very strong

As can be seen from (13), calculating the Bayes factor involves calculating the two marginal likelihoods, which are integrals over the parameter space. When the parameter space has a high dimension and/or the models are complex, this will be a very difficult task. This is, unfortunately, the case in our problem. Different methods have been developed for solving this problem (Clyde and George, 2004; Han and Carlin, 2001), which can be divided into two categories: *model augmentation (MA) methods* and *marginal likelihood estimation (MLE) methods*. The key difference between them lies in that the former calculate the Bayes factor directly, while the latter obtain the Bayes factor indirectly by calculating the marginal likelihood under each model.

Specifically, the MA methods (e.g., Carlin and Chib, 1995; Sisson, 2005) rest on the fact that

$$BF_{ij} = \frac{P(\mathbf{y} \mid M_i)}{P(\mathbf{y} \mid M_j)} = \frac{P(M_i \mid \mathbf{y})}{P(M_j \mid \mathbf{y})} / \frac{\pi(M_i)}{\pi(M_j)} = \frac{P(M_i \mid \mathbf{y})}{P(M_j \mid \mathbf{y})}$$
(14)

where  $\pi(M_i)$  and  $\pi(M_j)$  are priors of the two competing models, and the last equality achieves when the two priors are set to be equal, which is reasonable in most applications. (14) implies that the Bayes factor can be obtained by finding the ratio of the marginal posterior probabilities  $P(M_i | \mathbf{y})$  and  $P(M_j | \mathbf{y})$ . Based on this idea, the MA methods represent each model in the pool of candidate models by an indicator variable  $M \in \{1, 2, ...,\}$ , that is, M=1 indicates model  $M_1$ . This variable will be treated as an additional parameter, and MCMC algorithms, such as Metropolishasting (M-H) algorithms, are used to find its marginal posterior distribution. The marginal posterior probabilities of each model will then be obtained to fill in (14). Though such methods can find Bayes factors for all the comparisons through an integral sampling, this sampling is normally very complex due to the high dimension of the parameter space and particularly the varying number of parameters under different models which requires some dimension jumping techniques (Green, 1995). In contrast, the MLE methods calculate the marginal likelihood for each candidate model separately, and obtain the Bayes factor by finding the ratio of the marginal likelihood estimates, or, more conveniently, the difference between their logarithms

$$\log(BF_{ii}) = \log(P(\mathbf{y} \mid M_i)) - \log(P(\mathbf{y} \mid M_i))$$
(15)

This method is followed in our study as it is conceptually simpler and technically easier.

Because of the importance of marginal likelihoods in Bayesian inference, various MLE methods have been developed (e.g., Chen, 2005; Chib and Jeliazkov, 2001; Meng and Schilling, 2002). One method that has received much attention is the Chib's method for its simplicity and easiness in implement (Han and Carlin, 2001). This method is especially efficient for hierarchical models like the mixed-effect models considered in our study. More important, it is designed to obtain marginal likelihood estimates from posterior samples from Gibbs sampling, which makes it possible to integrate model selection with the parameter estimation approach described in Section 3.

#### 4.2. The proposed model selection procedure

The proposed model selection procedure is illustrated in Figure 2. Starting from the simplest model, i.e., a regular polynomial model without random effects, a set of posterior samples of the model parameters will be obtained through the Gibbs sampling procedure given in Section 3. These samples will then be used to estimate the marginal likelihood of the model. The Bayes factor of two adjacent models will be found by calculating the difference between the logarithms of their ML estimates by (15). Note that the posterior samples will also be used to calculate the

point estimates of the parameters when the best model is determined. The Chib's method to estimate marginal likelihoods will be described in the following. Appendix III summarizes the implementing steps of this procedure to provide convenience to practitioners.



Figure 2. The model selection procedure based on Bayes factors

To simplify, the model notations will be dropped in this section, e.g.,  $P(\mathbf{y})$  will denote the ML to be estimated, since the same method will be used for each model in the candidate pool. Also, as in Section 3, the full model in (4) will be used to demonstrate this method. Let the set of posterior samples obtained in the Gibbs sampling be

$$\{(\boldsymbol{\theta}^{(1)}, \boldsymbol{\alpha}^{(1)}), \dots, (\boldsymbol{\theta}^{(G)}, \boldsymbol{\alpha}^{(G)})\}$$

where G is the total number of samples,  $\theta$  is the parameters of interest in Model (4), i.e.,

$$\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{\sigma}_p^2, ..., \boldsymbol{\sigma}_0^2, \boldsymbol{\sigma}_{\varepsilon}^2]'$$

and  $\alpha$  is the random effects. The Chib'method is based on a simple identity

$$P(\mathbf{y}) = \frac{\pi(\mathbf{\theta})P(\mathbf{y} \mid \mathbf{\theta})}{P(\mathbf{\theta} \mid \mathbf{y})}$$

which applies for any value of  $\theta$ . For a given  $\theta$  (say  $\theta^*$ ), the ML can be estimated by

$$\log(\hat{P}(\mathbf{y})) = \log(P(\mathbf{y} \mid \boldsymbol{\theta}^*)) + \log(\pi(\boldsymbol{\theta}^*)) - \log(\hat{P}(\boldsymbol{\theta}^* \mid \mathbf{y})$$
(16)

which contains three quantities: the log-likelihood  $\log(P(\mathbf{y} | \boldsymbol{\theta}^*))$ , the log-prior  $\log(\pi(\boldsymbol{\theta}^*))$  and the posterior ordinate  $\log(\hat{P}(\boldsymbol{\theta}^* | \mathbf{y}))$ . They will be calculated individually using the posterior samples to obtain the ML. Details on how to select  $\boldsymbol{\theta}^*$  and calculate each of these quantities will be given as follows. (1) Selection of  $\theta^*$ : The choice of  $\theta^*$  is theoretically not critical as the simple identity holds for any value of  $\theta$ . However, for a given set of posterior samples  $\{\theta^{(1)},...,\theta^{(G)}\}$ , the probabilities in (16) is likely to be more accurately estimated at a high density point on  $P(\theta|\mathbf{y})$  than a point at the tails. Thus, appropriate choices of  $\theta^*$  are sample mode or mean of  $\{\theta^{(1)},...,\theta^{(G)}\}$ . The sample mean will be used for convenience, that is

$$\boldsymbol{\theta}^* = [\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*_{\alpha}, \boldsymbol{\sigma}^{2^*}_{\varepsilon}]' = \frac{1}{G} \sum_{g=1}^G \boldsymbol{\theta}^{(g)}$$
(17)

(2) Calculation of the log-likelihood: Note that the likelihood function  $P(\mathbf{y}|\boldsymbol{\theta})$  in the LME model is not directly available as the data depend on both the parameters of interest  $\boldsymbol{\theta}$  and the nuisance parameters  $\boldsymbol{\alpha}$ . From (4), it is easy to get

$$\mathbf{y}_i \mid \mathbf{\theta} \sim N(\mathbf{X}\mathbf{\beta}, \mathbf{X}\mathbf{\Sigma}_{\alpha}\mathbf{X}' + \sigma_{\varepsilon}^2 \mathbf{I})$$

Considering the independence of profiles, we will have

$$\log(P(\mathbf{y} \mid \boldsymbol{\theta}^*)) = \log\left(\prod_{i=1}^{m} P(\mathbf{y}_i \mid \boldsymbol{\theta}^*)\right) = \sum_{i=1}^{m} \log\left(N(\mathbf{y}_i \mid \mathbf{X}\boldsymbol{\beta}^*, \ \mathbf{X}\boldsymbol{\Sigma}_{\alpha}^*\mathbf{X}' + \sigma_{\varepsilon}^{2*}\mathbf{I})\right)$$
(18)

(3) Calculation of the log-prior: By the independence of parameters, the log-prior

$$\log(\pi(\boldsymbol{\theta}^*)) = \log(\pi(\boldsymbol{\beta}^*)) + \sum_{k=0}^{p} \log(\pi(\sigma_k^{2^*})) + \log(\pi(\sigma_{\varepsilon}^{2^*}))$$

There are three terms at the right of the equation. For the first and the third term, since noninformative priors in (6)~(7) are used for  $\beta$  and  $\sigma_{\varepsilon}^2$ , they are not defined. However, they are the same for all the candidate models as the two parameters are present in all of them, and thus cancel out in calculating the log(*BF*) in (15). So these two terms can be ignored. For the second term, this property does not apply because the candidate models may have different numbers of random effects and consequently the number of variance components. That is the reason for using the weakly informative prior in (8) instead of the noninformative prior. Under this prior, the second term is

$$\sum_{k=0}^{p} \log(\pi(\sigma_k^{2^*})) = \sum_{k=0}^{p} \mathbf{IG}(\sigma_k^{2^*} \mid \omega, \omega)$$
(19)

(4) Calculation of the posterior ordinate: The posterior ordinate can be estimated by

$$\log(\hat{P}(\boldsymbol{\theta}^* | \mathbf{y})) = \frac{1}{G} \sum_{g=1}^{G} \log(P(\boldsymbol{\theta}^* | \mathbf{y}, \boldsymbol{\alpha}^{(g)}))$$
(20)

To obtain this quantity, the conditional posterior distribution  $P(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\alpha})$  needs to be found first. The result is

$$P(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\alpha}) = \mathbf{IG}\left(\sigma_{\varepsilon}^{2} \middle| \frac{mn-2}{2}, \frac{\left(\mathbf{y} - \Xi\boldsymbol{\alpha} - \Xi(\Xi'\Xi)^{-1}\Xi'\mathbf{y}\right)'\left(\mathbf{y}_{i} - \Xi\boldsymbol{\alpha} - \Xi(\Xi'\Xi)^{-1}\Xi'\mathbf{y}\right)}{2}\right)$$

$$\cdot N\left((\Xi'\Xi)^{-1}\Xi'(\mathbf{y} - \Xi\boldsymbol{\alpha}), \ \sigma_{\varepsilon}^{2}(\Xi'\Xi)^{-1}\right) \cdot \prod_{k=0}^{p} \mathbf{IG}\left(\sigma_{k}^{2} \middle| \boldsymbol{\omega} + \frac{m}{2}, \ \boldsymbol{\omega} + \frac{1}{2}\sum_{i=1}^{m} \alpha_{i,k}^{2}\right)$$
(21)

The proof is given in Appendix IV. Consequently, the log posterior ordinate is

$$\log(P(\boldsymbol{\theta}^* | \mathbf{y}, \boldsymbol{\alpha}^{(g)})) = \log \operatorname{IG}\left(\sigma_{\varepsilon}^{2^*} \middle| \frac{mn-2}{2}, \frac{\left(\mathbf{y} - \Xi \boldsymbol{\alpha}^{(g)} - \Xi (\Xi' \Xi)^{-1} \Xi' \mathbf{y}\right)' \left(\mathbf{y}_i - \Xi \boldsymbol{\alpha}^{(g)} - \Xi (\Xi' \Xi)^{-1} \Xi' \mathbf{y}\right)}{2}\right)$$
$$+ \log\left[N(\boldsymbol{\beta} | (\Xi' \Xi)^{-1} \Xi' (\mathbf{y} - \Xi \boldsymbol{\alpha}^{(g)}), \sigma_{\varepsilon}^{2^*} (\Xi' \Xi)^{-1})\right]$$
$$+ \sum_{k=0}^{p} \log\left[\operatorname{IG}\left(\sigma_{k}^{2^*} \middle| \boldsymbol{\omega} + \frac{m}{2}, \boldsymbol{\omega} + \frac{1}{2} \sum_{i=1}^{m} \alpha_{i,k}^{*2}\right)\right]$$
(22)

#### 5. Numerical Study

A numerical study has been done to demonstrate the performance of the proposed approach. In the study, three datasets are generated from different LME models with the same degree of polynomials. The approach described in Section 3 is applied to each dataset to estimate their parameters, and the model selection procedure described in Section 4 is applied to compare candidate models. In the following, the setting of parameters in the simulation is first given, and then the results in parameter estimation and model selection are discussed. We actually have done extensive simulations by varying the parameter settings, and similar results are obtained.

### 5.1 Setting of parameters

Considering LME models with 2-degree polynomials, three datasets D1, D2, and D3 are generated from the following models

**D1:**  $y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \varepsilon_{ij}$  **D2:**  $y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,1} x_j + \varepsilon_{ij}$ **D3:**  $y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,0} + \varepsilon_{ij}$  which have random effect on the quadratic term, linear term and intercept, respectively. They are often referred to as random-quadratic-slope model, random-slope model and random-intercept model in the literature. The parameters of the models are set to be

$$\beta_2 = \beta_1 = \beta_0 = 2, \ \alpha_{i,2} \sim N(0, \ 0.6), \ \alpha_{i,1} \sim N(0, \ 0.6), \ \alpha_{i,0} \sim N(0, \ 0.6), \ \varepsilon_{ij} \sim N(0, \ 0.6)$$

Here equal values are used for the fixed effects and variance components to facilitate comparison of accuracy of their estimates. Under each model, m=60 profiles are generated with n=31sampling points. The sampling points are equally spaced within [-3.1, 3], that is, x=-3.1, -2.9, -2.7,...,2.9. The simulated data are shown in Figure 3. We can see that the three datasets exhibit apparently different characteristics. In practice, it might be possible to decide by eye the correct models for the data in such simple cases, but in more complex cases, e.g., with higher degrees of polynomials and/or multiple random effects present in the model, a formal model selection procedure must be used.



Figure 3. Simulated datasets used in the numerical study

# 5.2 Results of parameter estimation

In the first step to analyze a dataset, it is reasonable to fit a full model for it, i.e.,

$$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,1} x_j + \alpha_{i,0} + \varepsilon_{ij}$$

Under this model, the Gibbs sampling procedure in Appendix I is conducted for each dataset. The results of the three datasets show similar features, so those of D1 will be given here as an example. Figure 4 shows the stream of posterior samples of each parameter. Since casually picked starting values are used in the sampling, the samples are initially far from the truth, especially those of  $\beta_2$  and  $\sigma_{\varepsilon}^2$ , but then gradually converge. Another noticeable phenomenon in the figure is that two variance components  $\sigma_0^2$  and  $\sigma_1^2$  converge to very small values, indicating that the random effects of the intercept and liner term are negligible, and a random-quadraticslope model might be a better fit to the data rather than the full model. The effect of starting values disappear roughly at the 6000 iteration for all parameters, so the samples before that are discarded, and the remaining are used to obtain the posteriors of the parameters.



Figure 4. Posterior samples obtained from Globs sampling

Figure 5 shows the histograms of the samples, which represent the empirical posterior distributions of each parameter. We can see that the centers of these distributions are very close to the true values of the parameters. Consistent with the streams in Figure 4, the posteriors of  $\sigma_0^2$  and  $\sigma_1^2$  exhibit different shapes from other parameters, which concentrate in a small region

around 0, a sign of their insignificance. The means of the posteriors are  $2.00(\beta_2)$ ,  $2.00(\beta_1)$ ,  $1.92(\beta_0), 0.58(\sigma_{\varepsilon}^2), 0.54(\sigma_2^2), 0.0007(\sigma_1^2), 0.001(\sigma_0^2)$ , which can be used as point estimates of the parameters.



# 5.3 Results of model selection

Table 1. Candidate models for the data						
Model	Equation	Terms with random effects				
$M_0$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \varepsilon_{ij}$	no				
$M_1$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \varepsilon_{ij}$	quadratic				
$M_2$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,1} x_j + \varepsilon_{ij}$	linear				
$M_3$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,0} + \varepsilon_{ij}$	intercept				
$M_4$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,1} x_j + \varepsilon_{ij}$	quadratic & linear				
$M_5$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,0} + \varepsilon_{ij}$	quadratic & intercept				
$M_6$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,1} x_j + \alpha_{i,0} + \varepsilon_{ij}$	linear & intercept				
$M_7$	$y_{ij} = \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,1} x_j + \alpha_{i,0} + \varepsilon_{ij}$	all				
$M_2$ $M_3$ $M_4$ $M_5$ $M_6$ $M_7$	$\begin{aligned} y_{ij} &= \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,1} x_j + v_{ij} \\ y_{ij} &= \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,1} x_j + \varepsilon_{ij} \\ y_{ij} &= \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,0} + \varepsilon_{ij} \\ y_{ij} &= \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,1} x_j + \alpha_{i,0} + \varepsilon_{ij} \\ y_{ij} &= \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,1} x_j + \alpha_{i,0} + \varepsilon_{ij} \end{aligned}$	intercept quadratic & linear quadratic & intercej linear & intercept all				

For LME models with 2-degree polynomials, there are  $2^3=8$  candidate models as listed in Table 1. Since the pool of candidate models is not very large, we do not follow the sequential selection strategy illustrated in Figure 2, but just calculate the marginal likelihoods and Bayes factors for all the models following the procedure in Appendix III. The estimated marginal likelihoods and Bayes factors (all in log scale) are displayed in Table 2. Note that the true models for dataset D1, D2 and D3 are  $M_1$ ,  $M_2$ , and  $M_3$  respectively, and the Bayes factors in Table 2 are associated with the comparison of each model verse the true model of the data. According to the cut-off values given in Section 4.1, the true models are strongly supported in all cases, and the evidence in many cases are overwhelming. This means that using the Bayes factors, true models can be correctly identified.

	D1		D2		D3	
	ML	BF	ML	BF	ML	BF
$M_0$	-5023.5	2665.3	-3521.5	1204.2	-2741.7	486.3
$M_1$	-2358.2		-3531.6	1214.3	-2543.6	288.2
$M_2$	-5030.1	2671.9	-2317.3		-2750.5	495.1
$M_3$	-4472.1	2113.9	-3528.5	1211.2	-2255.4	
$M_4$	-2366.5	8.3	-2327.6	10.3	-2552.2	296.8
$M_5$	-2364.6	6.4	-3538.7	1221.4	-2264.1	8.7
$M_6$	-4479.1	2120.9	-2323.2	5.9	-2263.9	8.5
$M_7$	-2373.1	14.9	-2334.1	16.8	-2272.4	17

Table 2. Estimated marginal likelihoods and Bayes factors

### 6. Case Study

The proposed approach has also been applied to a set of optical profile data from a large low-E glass producer in US. For confidentiality reasons, the name of the company and information of their products are not disclosed in this text. The raw data are shown in Figure 6(a), which consist of m=45 profiles with n=40 equally spaced sampling points at  $\lambda$ =705, 710,..., 900nm. A direct observation of the data is that the profiles differ largely in their shapes, especially intercepts, so the LME models should be a good fit.



Figure 6. Optical profiles from a company: raw data (a) and transformed data (b)

Before applying the proposed modeling approach, some preprocessing is done to the raw data. This includes the centering/scaling transformation of  $\lambda$  values, i.e.,  $x=[\lambda-average(\lambda)]/50$ , which can improve the numerical properties of the fitting, and the determination of the degree of polynomials used for the data through fitting regular polynomial models to each profile and checking the residuals. It is found that the appropriate choice is p=3. One transformed profile is shown in Figure 6(b), where the curve is the fitted degree-3 polynomial model for the profile. Such data will be used in the following analysis.

Model selection is first conducted to determine the best LME model for the data. Following the procedure in Appendix III, the selection starts from the simplest model, i.e., regular 3-degree polynomial model without random effects, and then considers models with one random effect, models with two random effects, etc. This process only proceeds when strong evidence is present, i.e., 2log(BF)>6, in supporting a more complex model. Finally, the model with random effects in the quadratic term, linear term and intercept is selected

$$y_{ij} = \beta_3 x_j^3 + \beta_2 x_j^2 + \beta_1 x_j + \beta_0 + \alpha_{i,2} x_j^2 + \alpha_{i,1} x_j + \alpha_{i,0} + \varepsilon_{ij}$$

The log Bayes factor comparing this model and the full model, which yields the second largest marginal likelihood, is 9.9, which indicates strong superiority of the selected model.



Figure 7. Posteriors of the parameters of the selected model

Then the parameters of the selected model are estimated. Figure 7 shows the empirical posterior of each parameter. The point estimates are  $-0.1870(\beta_3)$ ,  $-1.1396(\beta_2)$ ,  $8.5296(\beta_1)$ ,  $30.2240(\beta_0)$ ,  $0.025(\sigma_{\varepsilon}^2)$ ,  $0.0045(\sigma_2^2)$ ,  $0.0111(\sigma_1^2)$ ,  $0.8208(\sigma_0^2)$ . The variance of the random intercept is predominant among the four variance components, which is consistent with our direct observation of the raw data.

## 7. Conclusions and Discussions

This study considers the modeling of optical profiles in low-E glass manufacturing processes, which is the basis of quality monitoring and control in such processes. We propose to use the linear mixed-effect model for the data and develop a unified Bayesian approach for parameter estimation and model selection. The results in the numerical study show that the proposed approach is able to identify the true model and provide accurate estimates of the parameters. The implementation of this approach in practice is demonstrated in the case study using a real dataset.

Quality control in low-E glass manufacturing is an appealing research topic, and this study is just a starting point of our efforts. There are many open issues following the direction of the current study, examples of which are: First, based on the established models using the proposed approach, Phase I and Phase II monitoring methods need to be developed. Second, like many other chemical processes, the low-E glass manufacturing process is featured by random noises that may present in the production. As a result, the quality measurements may contain a considerable amount of outliers. This gives the motivation to develop modeling approaches to quality profiles that are robust to the effect of outliers. Finally, the optical profiles used in this study actually only contain part of the spectrum as shown in Figure 1 and 6. When measurements of the whole spectrum (typically 380~1050nm) are considered, the LME model will not be adequate to characterize the profile, and more complex models, such as piecewise linear mixed-effect models and nonlinear mixed-effect models need to be used. These topics will be investigated in our future work.

### Appendices

## **Appendix I. The Gibbs Sampling Procedure in Section 3**

Starting step: Specify starting values  $\boldsymbol{\beta}^{(0)}, \{\boldsymbol{\sigma}_p^{2(0)}, ..., \boldsymbol{\sigma}_0^{2(0)}\}, \boldsymbol{\sigma}_{\varepsilon}^{2(0)}, \{\boldsymbol{\alpha}_i^{(0)}, i = 1, ..., m\}$ 

A simple method to find starting values is fitting a regular polynomial model to the data. The coefficient estimates will be used as starting values of  $\beta$ , the sample variance of those estimates will be used as starting values of  $\{\sigma_p^2,...,\sigma_0^2\}$ , and the estimate of random error variance will be used as the starting value of  $\sigma_{\varepsilon}^2$ . The random effects can simply be set to 0. *Step 1*: Generate a sample  $\{\alpha_i^{(g)}, i = 1,...,m\}$  from the conditional posterior in (10). *Step 2*: Generate a sample  $\{\sigma_p^{2(g)},...,\sigma_0^{2(g)}\}$  from the conditional posterior in (11). *Step 3*: Generate a sample  $\{\beta_p^{(g)}, \sigma_{\varepsilon}^{2(g)}\}$  from the conditional posterior in (12). Then go back to Step 1 and repeat this process for g=1,...,G.

# **Appendix II. Proof of (10)**

We will first prove a theorem and then prove (10) based on that.

**Theorem:** Let  $U_1$ ,  $U_2$ ,...,  $U_n$  be n independently normally distributed random variables with a common mean and scaled variances, that is

$$U_j \sim N\left(\theta, \frac{\sigma^2}{c_j}\right) \qquad j = 1,...,n$$

where  $c_1, ..., c_n$  are known nonzero constants,  $\sigma^2$  is known, and  $\theta$  is the unknown mean that needs to be estimated. Let  $u_1, u_2, ..., u_n$  be the corresponding observations of these variables. Under a normal prior for  $\theta$ , i.e.,

$$\theta \sim N(\mu_0, \tau_0^2)$$

the posterior is

$$\theta \mid u_1, \dots, u_n \sim N(\mu, \tau^2)$$

where

$$\mu = \frac{\frac{\mu_0}{\tau_0^2} + \sum_{j=1}^n c_j \mu_j / \sigma^2}{\frac{1}{\tau_0^2} + \sum_{j=1}^n c_j / \sigma^2} , \quad \tau^2 = \frac{1}{\frac{1}{\tau_0^2} + \sum_{j=1}^n c_j / \sigma^2}$$

**Proof:** Under the given prior, the posterior of  $\theta$  is

$$P(\theta | u_{1},...,u_{n}) \propto \pi(\theta) \cdot \prod_{j=1}^{n} P(u_{j} | \theta, \sigma^{2} / c_{j}) = N(\theta | \mu_{0}, \tau_{0}^{2}) \cdot \prod_{j=1}^{n} N(u_{j} | \theta, \sigma^{2} / c_{j})$$

$$\propto \exp\left\{-\frac{1}{2\tau_{0}^{2}}(\mu - \mu_{0})^{2}\right] \cdot \prod_{j=1}^{n} \exp\left[-\frac{c_{j}}{2\sigma^{2}}(u_{j} - \theta)^{2}\right]$$

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{(\mu - \mu_{0})^{2}}{\tau_{0}^{2}} + \frac{\sum_{j=1}^{n} c_{j}(u_{j} - \theta)^{2}}{\sigma^{2}}\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\frac{\theta - \frac{\mu_{0}\sigma^{2} + \tau_{0}^{2}\sum_{j=1}^{n} c_{j}u_{j}}{\sigma^{2} + \tau_{0}^{2}\sum_{j=1}^{n} c_{j}}\right]$$

which is a normal distribution with mean and variance

$$\mu = \frac{\mu_0 \sigma^2 + \tau_0^2 \sum_{j=1}^n c_j u_j}{\sigma^2 + \tau_0^2 \sum_{j=1}^n c_j}, \quad \tau^2 = \frac{\sigma^2 \tau_0^2}{\sigma^2 + \tau_0^2 \sum_{j=1}^n c_j}$$

The result in the Theorem can be obtained by simple manipulations of the above formulas. Note that when  $c_n \equiv 1$ , i.e.,  $U_1$ ,  $U_2$ ,...,  $U_n$  are i.i.d. random variables following  $N(\theta, \sigma^2)$ , the posterior mean and variance simplify to

$$\mu = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{\mu}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \ \tau^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

which is a well known result in the literature (Gelman, 2004). ■

Coming back to the LME model, by (4),

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}^{(-k)}\boldsymbol{\alpha}_i^{(-k)} + \boldsymbol{\alpha}_{i,k}\mathbf{x}^k + \boldsymbol{\varepsilon}_i, \qquad k = p,...,0$$

which can also be written as

$$\mathbf{y}_{i} - (\mathbf{X}\boldsymbol{\beta} + \mathbf{X}^{(-k)}\boldsymbol{\alpha}_{i}^{(-k)}) = \boldsymbol{\alpha}_{i,k}\mathbf{x}^{k} + \boldsymbol{\varepsilon}_{i} = \mathbf{z}_{i}$$
(A1)

Let  $\mathbf{z}_i = [z_{i1}, ..., z_{in}]', \mathbf{x}^k = [x_1^k, ..., x_n^k]'$  and  $\mathbf{\varepsilon}_i = [\varepsilon_{i1}, ..., \varepsilon_{in}]'$ , by (A1) and (2)

$$z_{ij} = \alpha_{i,k} x_j^k + \varepsilon_{ij} \sim N(\alpha_{i,k} x_j^k, \sigma_{\varepsilon}^2) \qquad j = 1, ..., n$$

or equivalently

$$\frac{z_{ij}}{x_j^k} \sim N\left(\alpha_{i,k}, \frac{\sigma_{\varepsilon}^2}{\left(x_j^k\right)^2}\right)$$

We can treat  $z_{ij}/x_j^k$ , j=1,...,n, as *n* independently normally distributed variables with a common mean  $\alpha_{i,k}$  and scaled variances  $\sigma_{\varepsilon}^2/(x_j^k)^2$ , which falls into the exact situation assumed in the Theorem. Also, by (3), the prior of the common mean  $\alpha_{i,k}$  is a normal distribution

$$\alpha_{i,k} \sim N(0, \sigma_k^2)$$

According to the Theorem, the corresponding posterior is

$$\alpha_{i,k} \mid \mathbf{z}_i \sim N \left( \frac{\sum_{j=1}^n (x_j^k)^2 z_{ij} / \sigma_{\varepsilon}^2}{\frac{1}{\sigma_k^2} + \sum_{j=1}^n (x_j^k)^2 / \sigma_{\varepsilon}^2}, \frac{1}{\frac{1}{\sigma_k^2} + \sum_{j=1}^n (x_j^k)^2 / \sigma_{\varepsilon}^2} \right)$$

By expressing the above formula in matrix form, (10) will be resulted. ■

### Appendix III. The model selection procedure in Section 4

Starting from the simplest model, for each given model  $M_i$ ,

Step 1: Generate posterior samples  $\{(\boldsymbol{\theta}^{(1)}, \boldsymbol{\alpha}^{(1)}), ..., (\boldsymbol{\theta}^{(G)}, \boldsymbol{\alpha}^{(G)})\}$  by the procedure in Appendix I. Step 2. Estimate the log marginal likelihood  $\log(\hat{P}(\mathbf{y}))$  by (16)

(2.1) Calculate the mean,  $\theta^*$ , of the posterior samples of  $\theta$ .

- (2.2) Calculate the log-prior  $\log(P(\mathbf{y} | \boldsymbol{\theta}^*))$  by (18).
- (2.3) Calculate the log-prior  $\sum_{k=0}^{p} \log(\pi(\sigma_k^{2^*}))$  by (19) and treat it as  $\log(\pi(\theta^*))$ .
- (2.4) Calculate the log posterior ordinate  $\log(\hat{P}(\boldsymbol{\theta}^* | \mathbf{y}))$  by (22).
- (2.5) Calculate  $\log(\hat{P}(\mathbf{y}))$  by (16) using the results in (2.2)~(2.4).

Step 3. Calculate the log Bayes factor of  $M_i$  vs.  $M_{i-1}$  by (15).

Step 4. If the log Bayes factor is larger than 3, conclude that  $M_i$  is better than  $M_{i-1}$ , and then move to model  $M_{i+1}$ ; otherwise conclude that  $M_{i-1}$  is the best model and stop this procedure.

# Appendix IV. Proof of (21)

By the independence of parameters  $(\beta, \sigma_{\varepsilon}^2)$  and  $\{\sigma_p^2, ..., \sigma_0^2\}$ ,

$$P(\boldsymbol{\theta} | \mathbf{y}, \boldsymbol{\alpha}) = P(\boldsymbol{\beta}, \sigma_{\varepsilon}^{2} | \mathbf{y}, \boldsymbol{\alpha}) \cdot P(\sigma_{p}^{2}, ..., \sigma_{0}^{2} | \mathbf{y}, \boldsymbol{\alpha})$$
  
$$= P(\sigma_{\varepsilon}^{2} | \mathbf{y}, \boldsymbol{\alpha}) \cdot P(\boldsymbol{\beta} | \sigma_{\varepsilon}^{2}, \mathbf{y}, \boldsymbol{\alpha}) \cdot P(\sigma_{p}^{2}, ..., \sigma_{0}^{2} | \mathbf{y}, \boldsymbol{\alpha})$$
  
$$= P(\sigma_{\varepsilon}^{2} | \mathbf{y}, \boldsymbol{\alpha}) \cdot P(\boldsymbol{\beta} | \sigma_{\varepsilon}^{2}, \mathbf{y}, \boldsymbol{\alpha}) \cdot P(\sigma_{p}^{2}, ..., \sigma_{0}^{2} | \boldsymbol{\alpha})$$
 (A2)

So to find the joint conditional posterior at the left side, we just need to find the three marginal conditional posteriors at the right. Under the uniform prior in (7),

$$\sigma_{\varepsilon}^{2} | \mathbf{y}, \boldsymbol{\alpha} \sim \mathbf{IG}\left(\frac{mn-2}{2}, \frac{(\mathbf{y} - \Xi\boldsymbol{\alpha} - \Xi(\Xi'\Xi)^{-1}\Xi'\mathbf{y})'(\mathbf{y}_{i} - \Xi\boldsymbol{\alpha} - \Xi(\Xi'\Xi)^{-1}\Xi'\mathbf{y})}{2}\right)$$

under the uniform prior in (4),

$$\boldsymbol{\beta} \mid \sigma_{\varepsilon}^{2}, \mathbf{y}, \boldsymbol{\alpha} \sim N((\Xi'\Xi)^{-1}\Xi'(\mathbf{y}-\Xi\boldsymbol{\alpha}), \ \sigma_{\varepsilon}^{2}(\Xi'\Xi)^{-1})$$

and under the weakly informative prior in (8),

$$\sigma_k^2 \mid \boldsymbol{\alpha} \sim \mathrm{IG}\left(\omega + \frac{m}{2}, \ \omega + \frac{1}{2}\sum_{i=1}^m \alpha_{i,k}^2\right)$$

Plugging in these distributions into (A2) will lead to (21). ■

### References

- Amiri, A., Jensen, W.A., and Kazemzadeh, R.B. (2009) A case study on monitoring polynomial profiles in the automotive industry. *Quality and Reliability Engineering International*, 26, 509-520.
- Arasteh, D., Carmody, J., Lee, E.S., and Selkowitz, S. (2004) Window systems for highperformance buildings, W. W. Norton & Company, New York, NY.
- Carlin, B.P., and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *Journal of Royal Statistical Society B*, **57**(3), 473-484.
- Carmody, J., Selkowitz, S., and Heschhong, L. (1996) *Residential windows: a guide to new technologies and energy performance.* W. W. Norton & Company, New York, NY.
- Chen, M-H (2005) Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica*, **59**(1), 16-29.
- Chib, S. (1995) Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, **90**(432), 1313-1321.

- Chib, S. and Jeliazkov, I. (2001) Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**, 270-281.
- Clyde, M., and George, E.I. (2004) Model uncertainty. *Statistical Science*, **19**(1), 81-94.
- Dongarra, J.J., Moler, C.B., and Wilkinson, J.H. (1983) Improving the accuracy of computed eigenvalues and eigenvectors, *SIAM Journal on Numerical Analysis*, **20**(1), 23-45.
- Frost, K., Arasteh, D., and Eto, J. (1993) Savings for energy efficient windows: current and future savings from new fenestration technologies in the residential market. *Lawrence Berkeley Laboratory Report 33965*, Lawrence Berkeley Laboratory.
- Gelfand, A.E., and Dey, D.K. (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of Royal Statistical Society B*, **56**(3), 501-514.
- Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1**(3), 515-533.
- Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2004) *Bayesian data analysis*, 2nd edition, Chapman & Hall/CRC.
- Gelman, A., and Hill, J. (2007) *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York, NY.
- Green, P.J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711-732.
- Han, C., and Carlin, B.P. (2001) Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, **96**(455), 1122-1132.
- Hobert, J.P., and Casella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models, *Journal of the American Statistical Association*, **91**(436), 1461-1473.
- Hunter Lab (2008) CIE L\*a\*b\* color scale. Insight on Color, 8(7), 1-4.
- Jensen, W.A., Birch, J.B., and Woodall, W.H. (2008) Monitoring correlation within linear profiles using mixed models. *Journal of Quality Technology*, **40**(2), 167-183.
- Kass, R.E., and Raftery, A.E. (1995) Bayes factors. Journal of the American Statistical Association, **90**(430), 773-795.
- Kazemzadeh, R.B., Noorossana, R., and Amiri, A. (2008) Phase I monitoring of polynomial profiles. *Communications in Statistics Theory and Methods*, **37**(10), 1671-1686.
- Kazemzadeh, R.B., Noorossana, R., and Amiri, A. (2009) Monitoring polynomial profiles in quality control applications, *International Journal of Advanced Manufacturing Technology*, 42, 703-712.
- Kazemzadeh, R.B., Noorossana, R., and Amiri, A. (2010) Phase II monitoring of Autocorrelated polynomial profiles in AR(1) processes. *Transaction E: Industrial Engineering*, **17**(1), 12-24.
- Kim, K., Mahmoud, M.A, and Woodall, W.H. (2003) On the monitoring of linear profiles. *Journal of Quality Technology*, **35**, 17-328.

- Meng, X.-L., and Schilling, S. (2002) Warp bridge sampling. *Journal of Computational and Graphical Statistics*, **11**, 552-586.
- Noorossana, R., Saghaei, A., and Amiri, A. (2011) *Statistical analysis of profile monitoring*. John Wiley & Sons, Hoboken, NJ.
- Pauler, D.K., Wakefield, J.C., and Kass, R.E. (1999) Bayes factors and approximations for variance component models, *Journal of the American Statistical Association*, 94(448), 1242-1253.
- Robert, C.P., and Casella, G (2004) *Monte Carlo statistical methods*, 2nd edition, Springer, New York, NY.
- Sisson, S.A. (2005) Transdimensional Markov chains: a decade of progress and future perspectives. *Journal of the American Statistical Association*, **100**(471), 1077-1089.
- Williams, J.D., Woodall, W.H., and Birch, J.B. (2007) Statistical monitoring of nonlinear product and process quality profiles, *Quality and Reliability Engineering International*, **23**, 925-941.
- Woodall, W.H. (2007) Current research on profile monitoring. Produção, 17(3), 420-425.
- Woodall, W.H., Spitzner, D.J., Montgomery, D.C., and Gupta, S. (2004) Using control charts to monitor process and product quality profiles. *Journal of Quality Technology*, **36**(3), 309-320.
- Wu, L. (2010) Mixed effects models for complex data. Taylor & Francis, Boca Raton, FL.