Simulation Data Generation from Extended EGA Model and Optimization of Alignment Strategy for Lithography

Shinichi NAKAJIMA[†‡] and Sumio WATANABE[†]

[†] Tokyo Institute of Technology,
4259 Nagatsuda, Midori-ku, Yokohama-shi,
Kanagawa, 226-8503 Japan.
E-mail: nakajima.s@cs.pi.titech.ac.jp,
swatanab@pi.titech.ac.jp

[‡] Nikon Corporation,
201-9 Miizugahara, Kumagaya-shi,
Saitama, 360-8559 Japan.

Abstract

An exposure tool, which is a key component in the field of optical microlithography for manufacturing semiconductor devices, has to align a silicon wafer to overlay the previously formed patterns accurately. For that purpose, the tool measures several marks on the wafer and then estimates the grid parameters that represent the arrangement of the patterns. In this paper, from consideration of circumstance of exposure tools in device factories, we assume that the measurement data are subject to a hidden Markov model, in which the grid parameter vector of each wafer corresponds to the hidden state vector at each time. We call this model "Extended EGA(Enhanced Global Alignment) model". Based on this assumption, we classify the components of the hidden state vector into several groups and propose an algorithm that treats each group separately. We roughly estimate the model parameters that specify Extended EGA model using a small amount of real data. By considering the estimates to be the true model parameter values, we generate simulation data to compensate lack of real data. They are used for optimization of some algorithm parameters and options, and for evaluation of dependence of the algorithm on some condition parameters, which include the number of samples and some of the model parameters.

1. INTRODUCTION

Semiconductor devices such as memories, processing units and so on are manufactured by the lithography technology. An exposure tool that copies circuit patterns from a mask to photosensitizer, which coats a silicon wafer, is a key component for this technology. Very high alignment accuracy is required so that the patterns of the previous layer, whose typical line width is 100nm, are overlaid with the patterns of the new layer correctly. For that purpose, the exposure tool measures the positions of several marks on the wafer and then estimates the grid parameters that represent the arrangement of the previously formed patterns. In our conventional method called "EGA(Enhanced Global Alignment)", the first order polynomial regression is applied for the estimation [1].

Recently, as higher accuracy is required for manufacturing higher density circuit devices, an exposure tool has to cope with higher order factors. One of the major factors is an outlier problem. For rejecting outliers, we developed an algorithm that utilizes the normal mixture models [2].

In this paper, we consider about another factor, which is higher order distortion of pattern arrangement. From consideration of circumstance of exposure tools in device factories, we assume that the measurement data are subject to a hidden Markov model, in which the grid parameter vector of each wafer corresponds to the hidden state vector at each time. We call this model "Extended EGA model". Based on this

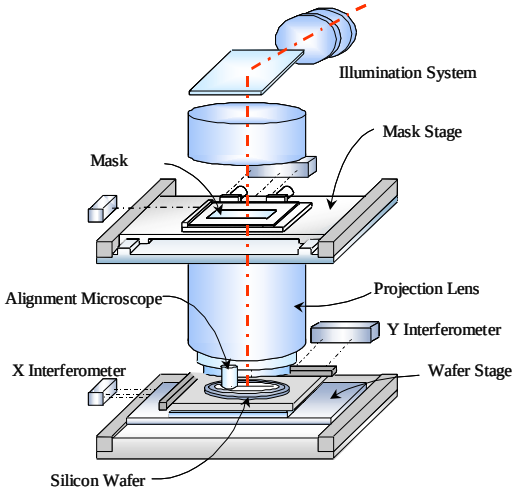

Figure 1: Exposure tool

assumption, we classify the components of the hidden state vector into several groups and propose an algorithm that treats each group separately. Every time measurement of a wafer finishes, the algorithm calculates an estimate of the hidden state vector both on a supposition that a transition of each group has occurred and the opposite supposition that the transition has not occurred. Then, the algorithm infers which supposition is true using a penalized likelihood approach [3].

For optimizing some algorithm parameters and options, we can not use enough amount of real data. That is because experiments interfere with customer's productivity. Therefore, using a limited amount of real data, we roughly estimate the model parameters that specify Extended EGA model. The estimates are considered as the true model parameter values in generating simulation data to compensate lack of real data. The simulation data are used not only for optimization of the algorithm but also for evaluation of dependence of the algorithm on some condition parameters, which include the number of samples and some of the model parameters. As a result of the simulation, we found that, for our problem, a good penalty coefficient value of the penalized likelihood approach is twice as large as the value that makes the approach equivalent to AIC. It is approximately independent of the number of samples, the transition probability of a group of components of the hidden state vector and the transition amplitude of the group.

We explain an exposure tool and the conventional EGA method in Section 2. In Section 3, We describe Extended EGA model and estimate the model parameters, which are used for simulation data generation, from a limited amount of experimental data. An proposed algorithm is described in Section 4. In Section 5, we show simulation results that show the effectiveness of the proposed algorithm. Conclusions are in Section 6.

2. BACKGROUND

2.1. Exposure Tool

The area of an image of a mask is $26 \times 33\text{mm}^2$ on a surface of a wafer, whose radius is 150mm. An exposure tool (Figure 1) exposes dozens of shots covering the whole on a surface of a wafer by stepping the wafer stage. Alignment marks, which are to be measured before exposure of the next layer, are included in the mask patterns and are copied to each shot on the wafer simultaneously with the other circuit patterns. Another exposure tool, which exposes the next layer, measures the positions of several marks on the wafer using the alignment microscope, and then estimates the pattern arrangement. Based on the estimated arrangement, the wafer stage is positioned accurately when each shot is exposed.

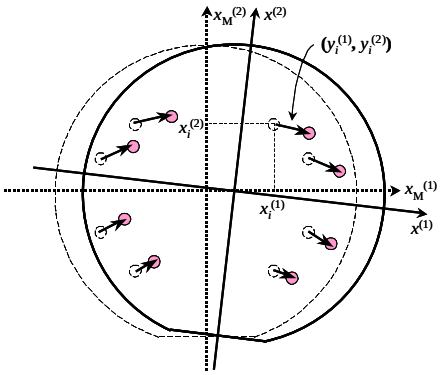

Figure 2: Enhanced Global Alignment

2.2. Enhanced Global Alignment

Though any wafer is pre-aligned before loading on the wafer stage, it is not enough accurate for overlay. Moreover, a wafer can extend or shrink by a little temperature variation. Therefore, at least, the shift, the scaling and the rotation of a wafer have to be estimated.

Let the input vector, which is the designed position vector of a mark, be $\boldsymbol{x} = (x^{(1)}, x^{(2)})^t$, and let the output vector be $\boldsymbol{y} = \boldsymbol{x}_M - \boldsymbol{x}$, where $\boldsymbol{x}_M$ is the measured position vector of the mark (see Figure 2), we assume that the output vector is subject to a polynomial regression model

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\boldsymbol{y} - \boldsymbol{a}^t\boldsymbol{\phi}(\boldsymbol{x})|^2}{2\sigma^2}\right), \qquad (1)$$

where $\boldsymbol{a}$ is the $K \times 2$ grid parameter matrix and $\boldsymbol{\phi}(\boldsymbol{x})$ is the K-dimensional polynomial vector

$$\begin{aligned}\boldsymbol{\phi}(\boldsymbol{x}) &= (\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \cdots, \phi_K(\boldsymbol{x}))^t \\ &= \left(1,\ x^{(1)},\ x^{(2)},\ {x^{(1)}}^2,\ x^{(1)}x^{(2)},\ {x^{(2)}}^2,\ \cdots\right)^t.\end{aligned}$$

In our conventional method, the first order terms are employed, $K = 3$. The grid parameter matrix $\boldsymbol{a}$ is estimated by the maximum likelihood method. Because $\boldsymbol{x}$ can be distributed over the whole of a wafer, the amplitude of $\boldsymbol{y}$ concerned with any mark is much smaller than the extent of $\boldsymbol{x}$ due to the pre-alignment. Therefore, the first order terms are enough to represent the shift, the scaling and the rotation of a wafer approximately [1].

3. EXTENDED EGA MODEL

3.1. A Hidden Markov Model

Higher order distortion of pattern arrangement is considered to be caused mainly from individual grid difference of tools. Usually, each layer is exposed by a different tool because of efficiency of manufacturing. Therefore, matching of pattern arrangement between tools is important. Straightness of the mirrors, which a wafer stage has for measurement of its position by the interferometers, is strictly required(see Figure 1). Therefore, the error from straightness of the mirrors is calibrated and compensated in each tool. However, as required accuracy become higher, residual distortion becomes non-negligible. We assume from the above consideration that the measurement data are subject to the following model.

Because we assumed in (1) that the components of the output vector, $y^{(1)}$ and $y^{(2)}$, are independent each other, we focus on one component of the output vector for simplicity. Therefore, in this section and the latter sections, we regard $\boldsymbol{a}$ as a K-dimensional vector, and let the output scalar be $y = y^{(1)}$. We use an orthonormalized polynomial vector $\boldsymbol{e}(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{\phi}(\boldsymbol{x})$, where $\boldsymbol{Q}$ is a $K \times K$ lower triangular matrix that is generated via the Gram-Schmidt procedure, so that

$$\frac{1}{n}\sum_{i=1}^{n} e_j(\boldsymbol{x}_i)e_k(\boldsymbol{x}_i) = \delta_{jk},$$

where δ_{jk} is the Kronecker delta. The set of the input vectors $\{\boldsymbol{x}_i; i \in [1, n]\}$ is assumed to be common for each wafer. The corresponding grid parameter vector is

$$\boldsymbol{b} = \left(\boldsymbol{Q}^{-1}\right)^t \boldsymbol{a}, \tag{2}$$

which satisfies $\boldsymbol{a}^t\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{b}^t\boldsymbol{e}(\boldsymbol{x})$ for arbitrary $\boldsymbol{x}$.

Then, we classify the grid parameter components into H groups. $\boldsymbol{b}$ and $\boldsymbol{e}(\boldsymbol{x})$ are represented as

$$\boldsymbol{b} = \left(\boldsymbol{b}_1^t, \quad \boldsymbol{b}_2^t, \quad \cdots, \quad \boldsymbol{b}_H^t\right)^t \qquad \text{and}$$
$$\boldsymbol{e}(\boldsymbol{x}) = \left(\boldsymbol{e}_1(\boldsymbol{x})^t, \quad \boldsymbol{e}_2(\boldsymbol{x})^t, \quad \cdots, \quad \boldsymbol{e}_H(\boldsymbol{x})^t\right)^t$$

respectively, where $\boldsymbol{b}_h$ is a d_h-dimensional vector that consists of the components belonging to the h-th group, and $\boldsymbol{e}_h(\boldsymbol{x})$ is a d_h-dimensional vector that consists of the corresponding polynomials.

Because $\boldsymbol{b}$ is not a directly observable, we regard it as the hidden state vector of a hidden Markov model. Assuming that the components belonging to the same group make transition simultaneously, a hidden Markov model, which is called Extended EGA model in this paper, is described as follows. The conditional density of the output scalar y given the input vector $\boldsymbol{x}$ and the hidden state vector $\boldsymbol{b}$ at the time s, which corresponds to the s-th wafer, is

$$p(y|\boldsymbol{x}, \boldsymbol{b}(s)) = \frac{\exp\left(-\frac{1}{2\sigma^2(s)}\left(y - \boldsymbol{b}(s)^t\boldsymbol{e}(\boldsymbol{x})\right)^2\right)}{\sqrt{2\pi\sigma^2(s)}}, \tag{3}$$

and the conditional density of the h-th group hidden state vector $\boldsymbol{b}_h(s)$ given the one at the previous time $(s-1)$ is

$$p\left(\boldsymbol{b}_h(s)|\boldsymbol{b}_h(s-1)\right) = (1-q_h)\cdot\delta_{d_h}\left(\boldsymbol{b}_h(s) - \boldsymbol{b}_h(s-1)\right) + q_h \cdot \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{b}_h(s) - \tilde{\boldsymbol{b}}_h\right)^t \boldsymbol{\Sigma}_h^{-1}\left(\boldsymbol{b}_h(s) - \tilde{\boldsymbol{b}}_h\right)\right)}{(2\pi)^{d_h/2}\left|\boldsymbol{\Sigma}_h\right|^{1/2}}, \tag{4}$$

where $\delta_{d_h}(\boldsymbol{b}_h)$ is the Dirac delta function that satisfies

$$\int_W \delta_{d_h}(\boldsymbol{b}_h)d\boldsymbol{b}_h = \begin{cases} 0 & \text{if } \boldsymbol{0} \notin W \\ 1 & \text{if } \boldsymbol{0} \in W \end{cases},$$

where W is a subset of d_h-dimensional space. The variables and the parameters of this model are shown on Table 1. The upper two are the variables depending on s and the middle four are the parameters. The lower two appears in Section 3.2.

Table 1: The variables and the parameters of Extended EGA model and the parameters for simulation.

V	$\boldsymbol{b}_h(s)$	the h-th group hidden state vector at s
	$\sigma^2(s)$	observed noise variance at s
P	$\tilde{\boldsymbol{b}}_h$	transition average vector of $\boldsymbol{b}_h$
	$\boldsymbol{\Sigma}_h$	transition covariance matrix of $\boldsymbol{b}_h$
	d_h	dimension of $\boldsymbol{b}_h$
	q_h	transition probability of $\boldsymbol{b}_h$
S	σ_h	transition amplitude of $\boldsymbol{b}_h$, $\boldsymbol{\Sigma}_h = \sigma_h^2\mathbf{1}^{(d_h)}$
	$\tilde{\sigma}^2$	constant true value of $\sigma^2(s)$

3.2. Determination of the Model Parameters

We employ a classification as follows. The number of the groups is $H = 2$. The first group consists of the three linear terms, $d_1 = 3$,

$$\boldsymbol{e}_1(\boldsymbol{x}) = \boldsymbol{Q}_{\{1,3\}}\left(1, \quad x^{(1)}, \quad x^{(2)}\right)^t,$$

and the second group consists of the second and the third order terms, $d_2 = 7$,

$$\boldsymbol{e}_2(\boldsymbol{x}) = \boldsymbol{Q}_{\{4,10\}}\left(\cdots, \quad {x^{(1)}}^u {x^{(2)}}^v, \quad \cdots\right)^t \qquad (u, v \geq 0, 2 \leq u+v \leq 3),$$

Table 2: The estimated grid parameter values of one lot that consists of ten wafers.

	$\boldsymbol{b}_1$			$\boldsymbol{b}_2$							$\sqrt{\sigma^2}$
$[u,v]$	$[0,0]$	$[1,0]$	$[0,1]$	$[2,0]$	$[1,1]$	$[0,2]$	$[3,0]$	$[2,1]$	$[1,2]$	$[0,3]$	
Ave.	0.3630	0.2273	-1.9043	0.0008	-0.0016	0.0024	0.0039	0.0015	-0.0010	-0.0008	0.0028
dev.	0.0558	0.0027	2.5530	0.0005	0.0006	0.0004	0.0003	0.0003	0.0004	0.0002	0.0010

where $\boldsymbol{Q}_{\{j,k\}}$ is the extracted submatrix from the l-th rows of the orthonormalizing matrix $\boldsymbol{Q}$, which appears in (2), where $l \in [j,k]$. We also assume that a transition of the first group occurs every wafer, $q_1 = 1$, and a transition of the second group occurs probabilistically. Though the transition probability of the second group q_2 depends on a situation of the factory where the tool is, we assume that the range of the parameter is $q_2 \in [0.001, 0.1]$. We assume $\tilde{\boldsymbol{b}}_h = \boldsymbol{0}^{(d_h)}$ for $h \in [1,H]$, where $\boldsymbol{0}^{(d)}$ is the $d \times d$ zero matrix. This is because valid estimation of $\tilde{\boldsymbol{b}}_h$ is very difficult. This information is used only for simulation data generation but not for algorithm development.

We assume $\boldsymbol{\Sigma}_h = \sigma_h^2 \boldsymbol{1}^{(d_h)}$, where $\boldsymbol{1}^{(d)}$ is the $d \times d$ identity matrix and σ_h is the transition amplitude of the h-th group. We make $\sigma_1 = 0$ in generating simulation data because it does not effect upon evaluation of the proposed algorithm described in Section 4. We assume that the true value of $\sigma^2(s)$ is independent of s and is equal to a constant $\tilde{\sigma}^2$. Now, what we have to determine are σ_2 and $\tilde{\sigma}^2$, which are listed on Table 1 as parameters for simulation.

We estimate these values roughly from the limited amount of real data. We got data of a few lots of wafers. Each lot consists of ten wafers that are exposed by the same tool. Sixty-one shots on each wafer were exposed and the marks of all the shots were measured in experiment. Table 2 shows the averages of the $\boldsymbol{b}(s)$ component estimators over the ten wafers of typical one of the lots, and the deviations. The square root of the average of $\sigma^2(s)$ and the one of the deviation are also shown. A pair of numbers $[u,v]$ in the second row indicates the component corresponding to the term ${x^{(1)}}^u {x^{(2)}}^v$. The values on the table are calculated after a normalization of the input vector so that the radius of a wafer is equal to unity, and they are shown in micrometer. According to Table 2, the norm of the average vector of $\boldsymbol{b}_2$ is equal to $0.002\sqrt{d_2}$ approximately. Therefore, we assume $\sigma_2 = 0.002$. According to the table, we also assume $\sqrt{\tilde{\sigma}^2} = 0.003$.

Table 2 also supports the assumption that $\boldsymbol{b}_1$ makes transitions even if the wafers are exposed by the same tool and $\boldsymbol{b}_2$ does not. Now, we can generate simulation data. Our final goal is development of an algorithm that estimates the hidden state vector $\boldsymbol{b}(s)$, which corresponds to the grid parameter vector of the s-th wafer, after the measurement of the s-th wafer.

4. ALGORITHM

4.1. Basic Concept

An algorithm we propose consists of two steps. The first step is estimation of the hidden vector both on a supposition that a transition of each group has occurred and the opposite supposition that the transition has not occurred. The second step is inference of which supposition is true. We explain them in the following two subsections.

4.2. Hidden State Estimation

On a supposition that the transition of the h-th group has occurred at s, the hidden state $\boldsymbol{b}_h(s)$ is estimated by the maximum likelihood estimation,

$$\hat{\boldsymbol{b}}_h^{\mathrm{TR}}(s) = \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{e}_h(\boldsymbol{x}_i). \tag{5}$$

On the opposite supposition, according to (4), the estimator of $\boldsymbol{b}_h(s)$ should be

$$\hat{\boldsymbol{b}}_h^{\mathrm{NT}}(s) = \boldsymbol{b}_h(s-1). \tag{6}$$

However we do not know the previous true state $\boldsymbol{b}_h(s-1)$. We consider the following two options.

(A) Replace $\boldsymbol{b}_h(s-1)$ with the hidden state estimator for the previous wafer $\hat{\boldsymbol{b}}_h(s-1)$,

$$\hat{\boldsymbol{b}}_h^{\mathrm{NT}}(s) = \hat{\boldsymbol{b}}_h(s-1). \tag{7}$$

(B) Replace $\boldsymbol{b}_h(s-1)$ with the average of $\hat{b}_h^{\mathrm{TR}}(s')$ where $s' \in [s'', s]$ and s'' is the inferred time, when the latest transition did occur, by the second step described in Section 4.3,

$$\begin{aligned}\hat{\boldsymbol{b}}_h^{\mathrm{NT}}(s) &= \frac{1}{s-s''+1}\sum_{s'=s''}^{s} \hat{\boldsymbol{b}}_h^{\mathrm{TR}}(s') \\ &= \frac{(s-s'')\hat{\boldsymbol{b}}_h^{\mathrm{NT}}(s-1) + \hat{\boldsymbol{b}}_h^{\mathrm{TR}}(s)}{s-s''+1}.\end{aligned} \tag{8}$$

This is the maximum likelihood estimation under the condition that all the inferences during $s' \in [s'', s]$ were right.

4.3. Inference of Transition Occurrence

To infer which supposition is true, we use a penalized likelihood approach [3]. On each supposition, the hidden state vector has been estimated by the previous step. We can calculate the log likelihood of each supposition.

A penalized likelihood criterion, which is the sum of the log likelihood and a penalty factor that is proportional to degree of freedom, is

$$PL\left(V^{TR}\right) = L - \beta(n) \cdot \sum_{h \in V^{TR}} d_h, \tag{9}$$

where L is the maximum log likelihood of a supposition, V^{TR} is a set of integers such that

$$V^{TR} = \{h \in [1, H]; \text{the } h\text{-th group makes transition on the supposition}\},$$

and $\beta(n)$ is a penalty coefficient. The supposition that has the maximum PL value is adopted. This approach is equivalent to AIC when $\beta(n) = 1$, and equivalent to BIC when $\beta(n) = (1/2)\log n$ [4, 5]. The penalty coefficient is to be determined based on simulation results discussed in Section 5.

5. EVALUATION

5.1. Common Condition for Simulation

We generate simulation data based on the result in Section 3.2, and evaluate the proposed algorithm. In each subsection that follows, we show a condition parameter dependence. The values of condition parameters that are not specified in each subsection are the following default values. Each evaluation consists of a hundred trials, $M_{\mathrm{t}} = 100$. Each trial consists of the data of a series of two hundred wafers, $M_{\mathrm{w}} = 200$. Each wafer has sixty-one shots, $N = 61$. The marks of twenty of the shots are measured, $n = 20$, except for Section 5.3. The transition probability of the second group is $q_2 = 0.01$ except for Section 5.4, however, the effective value is $\sim q_2 + 1/M_{\mathrm{w}}$ because the transition of the head of any series must occur. The transition amplitude of the second group is $\sigma_2 = 0.002$ except for Section 5.5. As a hidden state estimation method (see Section 4.2), option (B) is employed except for Section 5.2.

The horizontal axis of each figure is the penalty coefficient β and the vertical axis is the square root of the average of the generalization error over the M_{t} trials. The generalization error is defined as

$$G = \frac{1}{M_{\mathrm{w}}} \sum_{s=1}^{M_{\mathrm{w}}} \left(\frac{1}{N} \sum_{i=1}^{N} \left(\left(\hat{\boldsymbol{b}}(s) - \boldsymbol{b}^*(s) \right)^t \boldsymbol{e}(\boldsymbol{x}_i) \right)^2 \right),$$

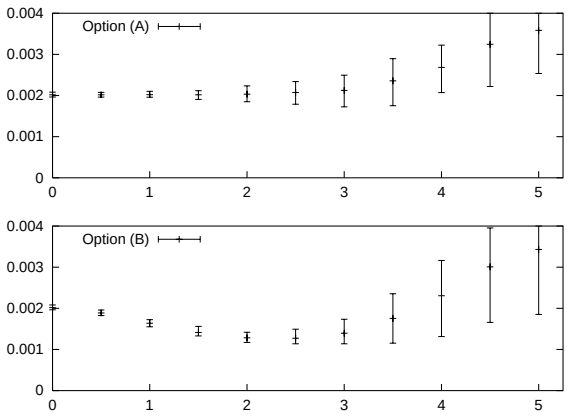

Figure 3: State estimation option dependence.

where $\hat{\boldsymbol{b}}(s)$ is the estimated hidden state vector and $\boldsymbol{b}^*(s)$ is the true one. The fifth smallest value of the M_{t} trials and the fifth largest value are also shown as an error bar.

5.2. State Estimation Option Dependence

Figure 3 shows dependence on selection from the hidden state estimation options described in Section 4.2. It shows that in the case of option (A), the penalty factor does not improve the generalization error. On the other hand, in the case of option (B), the generalization error is improved if the penalty coefficient is around two.

5.3. Sample Number Dependence

Figure 4 shows n dependence. Though the optimum value of the penalty coefficient slightly increases as n decreases, The algorithm still shows good performance if $\beta = 2$.

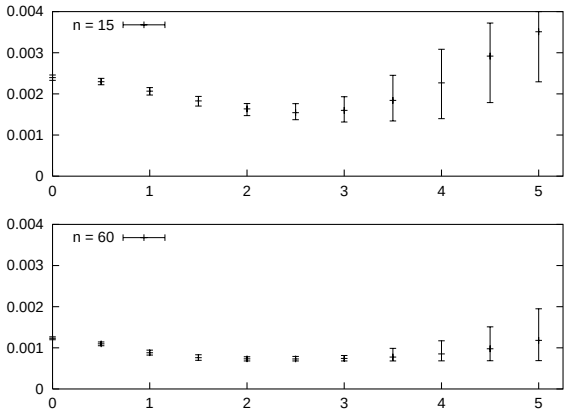

Figure 4: Sample number dependence.

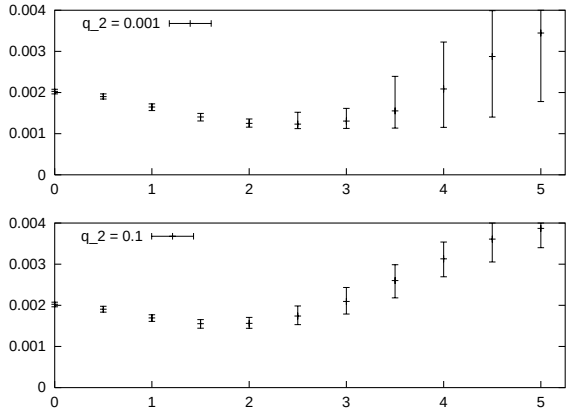

Figure 5: Transition probability dependence.

5.4. Transition Probability Dependence

Figure 5 shows q_2 dependence. Though the optimum penalty coefficient slightly decreases as q_2 increases, The algorithm still shows good performance if $\beta = 2$.

5.5. Transition Amplitude Dependence

Figure 6 shows σ_2 dependence. This shows that the optimum penalty coefficient depends on σ_2. However, the algorithm still shows good performance if $\beta = 2$. We did not evaluate dependence on another condition parameter $\tilde{\sigma}^2$, which is the variance of the observed noise. That was because any variation that does not change the ratio $\sigma_2/\sqrt{\tilde{\sigma}^2}$ does not effect upon the optimal penalty coefficient.

6. CONCLUSION

We proposed an algorithm based on the assumption that the data of alignment for an exposure tool are subject to a hidden Markov model, which is called Extended EGA Model in this paper. We evaluated the algorithm with the simulation data that are generated based on estimation of the model parameters with a limited number of real data. We conclude from the evaluation that the algorithm performs well approximately independent of some condition parameters.

In this paper, assuming that the components of the output vector, $y^{(1)}$ and $y^{(2)}$, are independent each other, we analyzed a model whose output vector is one-dimensional. As a result of that, the rotation of x axis and the one of y axis are assumed to be independent each other. However, according to natural consideration, there must be strong correlation between them actually. Development of an algorithm that considers this fact is a future work. Minor transitions of the hidden state vector such as the *trend* also should be considered. An approach utilizing the Kalman filter can be applied. This is another future work.

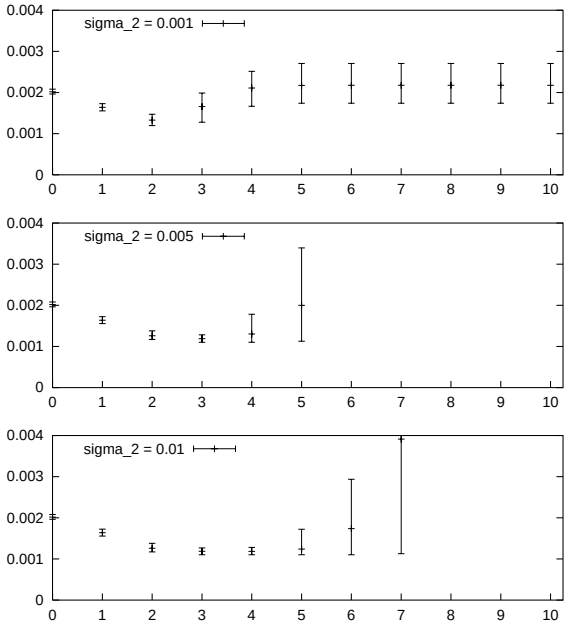

Figure 6: Transition amplitude dependence.

Acknowledgments

The authors would like to thank Kazuo Ushida, Masahiro Nei and Nobutaka Magome of Nikon Corporation for encouragement to do this work.

References

[1] T. Umatate, "Method for Successive Alignment of Chip Patterns on a Substrate," *US Patent*, 4,780,617, 1988.

[2] S. Nakajima et al., "Outlier rejection with Mixture Models in Alignment," *Proc. of SPIE*, Vol.5040, pp.1729–1741, March 2003.

[3] B. G. Leroux and M. L. Puterman, "Maximum Penalized Likelihood Estimation for Independent and Markov-Dependent Mixture Models," *Biometrics*, Vol.41, pp.545–558, 1992.

[4] H. Akaike, "A New Look at the Statistical Model Identification," *IEEE Trans. on Automatic Control*, Vol. 19, pp.716–723, 1974.

[5] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, Vol.6, No.2, pp.461–464, 1978.