

## Bayesian group analysis of plasma-enhanced chemical vapour deposition data

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2004 New J. Phys. 6 25

(<http://iopscience.iop.org/1367-2630/6/1/025>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 38.107.179.211

The article was downloaded on 20/02/2012 at 15:56

Please note that [terms and conditions apply](#).

## Bayesian group analysis of plasma-enhanced chemical vapour deposition data

**R Fischer**

Centre for Interdisciplinary Plasma Science, Max-Planck-Institut für Plasmaphysik, EURATOM Association, Boltzmannstr. 2, 85 748 Garching, Germany

E-mail: [Rainer.Fischer@ipp.mpg.de](mailto:Rainer.Fischer@ipp.mpg.de)

*New Journal of Physics* **6** (2004) 25

Received 17 December 2003

Published 19 February 2004

Online at <http://www.njp.org/> (DOI: 10.1088/1367-2630/6/1/025)

**Abstract.** A ubiquitous goal in plasma-enhanced chemical vapour deposition (PECVD) is to describe the correlation between film properties and categorical and quantitative input variables. The correlations within the high-dimensional parameter space are described using a multivariate model. Bayesian group analysis is employed to assess the grouping structures of the set of data vectors. This allows to identify sub-groups or meta-groups of predefined groups of data sets, e.g. with respect to source gases. Outliers can be identified by the necessity to form a separate group. The Bayesian approach consistently allows the handling of missing data. The grouping probabilities were compared with classical approaches such as likelihood ratio tests, the Akaike information criterion and a Bayesian variant called Bayesian information criterion. The method was applied to PECVD data of rare-earth oxide film deposition and hydrocarbon film deposition to study the evidence of grouping structures attributed to categorical quantities such as rare-earth components or source gases and quantitative variates such as bias voltage.

**Contents**

<b>1. Introduction</b>	<b>2</b>
<b>2. Bayesian group analysis</b>	<b>4</b>
2.1. Grouping hypotheses . . . . .	5
2.2. Marginal likelihood for known covariances . . . . .	5
2.3. Marginal likelihood for unknown covariances . . . . .	6
2.4. Splitting or combining groups . . . . .	7
2.5. Outlier detection . . . . .	8
2.6. Missing data . . . . .	9
2.7. AIC and BIC . . . . .	10
<b>3. Results</b>	<b>11</b>
3.1. Rare-earth oxide film deposition . . . . .	11
3.2. Hydrocarbon film deposition . . . . .	13
<b>4. Discussion</b>	<b>17</b>
<b>5. Conclusion</b>	<b>18</b>
<b>Acknowledgments</b>	<b>19</b>
<b>References</b>	<b>19</b>

**1. Introduction**

In many fields of research, a large collection of data is given for which a detailed theory is yet missing. To gain insight into the underlying theoretical description, it is important to reveal the interrelationship among the data. Such data sets typically consist of measurements on more than one variable (variate). To find and describe the underlying data structure in high-dimensional parameter space, it is usually necessary to supplement intuition with statistical techniques. Group or cluster analyses were applied to sets of data to find group structures and to classify objects into groups. The goal is to find groups of objects that have a small within-group variability relative to the between-group variability. Classical measures for finding the most distinct and compact groups can be found in [1].

A difference between group analysis and cluster analysis is given by the presence or absence of external information on ‘natural’ structures, respectively. Group analysis enables one to find evidence for genuine groups among the sample members on the basis of presence of external structures. External structures are common in designed experiments where, for example, the *a priori* information on preformed groups is given by different sets of input quantities. The goal is to decide if the preformed groups form genuine groups or they can be pooled into larger meta-structures (groupings) without loss of information. The decision to study groups of data separately or together in a merged meta-group was expected to have an impact on the conclusions to be drawn.

The present applications provide preformed groups by the design variates of the source gas used and the bias voltage applied in plasma deposition. The bias voltage is a *quantitative* variate that is characterized by a numerical value, whereas the source gas belongs to the class of *categorical* or *qualitative* variates.

Classical grouping or clustering techniques are based on tests *against* a null hypothesis where (maximum-likelihood) estimates for the group parameters enter. Bayesian group analysis

allows one to calculate the evidence *in favour of* a hypothesis where the group parameters are integrated out. The integration or *marginalization* of parameters takes into account the full variability of parameter values for evidence of a grouping hypothesis. In contrast, the classical approach validates hypotheses according to their compatibility with the single value of the respective estimates. Details about classical cluster or group analysis techniques can be found in [1].

Another difference between classical and Bayesian group analysis arises when one needs to estimate parameters that are common to all groupings. Classically, parameters that are common to all groupings and parameters that are meaningful only for individual groupings were estimated for each grouping separately. This gives as many estimates for a common parameter as the number of groupings that could be identified. The Bayesian method allows for an integration over the parameters meaningful for individual groupings, which gives only one estimate for a common parameter. Hence, the full grouping uncertainty propagates into the estimation uncertainty of the common parameter.

The present paper starts with the methodology of Bayesian group analysis in a general framework. The method can be applied to any set of data where one anticipates genuine groups within an unstructured sample of measured data or where one has to decide if preformed ('natural') groups should be analysed separately or pooled in larger groupings. Then, the method of Bayesian group analysis is applied to typical examples from plasma-enhanced chemical vapour deposition (PECVD).

The deposition of thin films by PECVD is an important technique, since film properties such as density, composition and index of refraction can be controlled by input variables such as type of gas and ion energy [2]. Material tailoring requires knowledge of the influence of input variables on film properties. Parametric dependencies can be superposed by physical mechanisms that change with the experimental design parameters, e.g. the source gas. An exhaustive scan of all input variables becomes more and more prohibitive as the dimension of the input space increases. Regularly, we have to analyse data where the sample number is small or similar to the dimension of the parameter space. Estimating the relations between input and response variates and, even more cumbersome, finding empirically different growth mechanisms become progressively more difficult the sparser the data.

Data sets for film properties are often incomplete due to limited measurement resources, limited sensitivity ranges of the measurement techniques or just due to the combination of data from different databases. Usually, missing data are simply ignored by omitting incomplete data vectors. However neglecting partially measured information makes the results less reliable compared with the case where all data, complete and incomplete, are used. Analysing all data at hand becomes even more important when the set of data vectors is mostly spoiled with missing information or, to the extreme, when only data vectors with missing information are present, e.g. owing to experimental constraints.

von Keudell *et al* [2] applied a multivariate analysis to noise-corrupted PECVD data for the deposition of a-C:H films from an rf-plasma of methane. They used a multivariate normal distribution to describe the sparse data set in the high-dimensional parameter space of input and output variables. Although the multilinear relation may be too simple for a realistic physical model, a more elaborate descriptive relation is not appropriate owing to the limited data set. Dose [3] used a multivariate model for disclosing relations between the quantities determining the properties of PECVD films of Al<sub>2</sub>O<sub>3</sub> and rare-earth oxides. The technique of principle component analysis and canonical relations was applied to analyse grouping structures of preformed groups.

von der Linden *et al* [4] developed and applied a Bayesian approach to group analysis based on multivariate models. Multivariate analysis is a powerful technique to identify linear dependences between sets of data vectors. However, the restriction to linear models may be inappropriate for strong non-linear data sets. To verify the linearity assumption, the prediction of response variables from input variables can be compared with measured response variables [2]. Many problems are globally or locally linear or can be made linear by appropriate parameter transformations.

The present paper provides extensions to the Bayesian approach described in [4] and applies the technique to PECVD data of rare-earth oxide films and a-C:H films. Instead of using a common covariance structure for all groupings, we allow for independent covariance matrices for different groupings. Secondly, we provide tools for handling missing data in the full Bayesian framework and, more user-friendly, partial Bayesian solutions for analysing incomplete data. Classical statistics fails to provide a consistent method to deal with missing data [1]. Results from the Bayesian approach will be compared with classical hypothesis tests such as the likelihood ratio test (LRT) and the Akaike information criterion (AIC). In addition, the Bayesian results are compared with an approximation to the full Bayesian technique, namely the Bayesian information criterion (BIC). Both AIC and BIC are easy to implement. Hence, they can easily be used for fast checks of grouping probabilities whenever they provide good approximations to the full Bayesian method.

## 2. Bayesian group analysis

An object is characterized by  $L$  continuous characteristics  $\mathbf{d} = (d^1, \dots, d^L)^T$  that are represented by real numbers. Discrete characteristics are naturally used to preform groups such as the type of a carrier gas. Assuming a group  $\nu$  is characterized by a mean vector  $\mathbf{m}_\nu$  and a covariance matrix  $C_\nu$ , the data points  $\mathbf{d}_\nu^i$  ( $i \in \{1, \dots, N_\nu\}$  in group  $\nu$ ) are i.i.d. multi-normal

$$p(\mathbf{d}_\nu^i | \mathbf{m}_\nu, C_\nu) = \frac{\text{etr}\{-\frac{1}{2}C_\nu^{-1}(\mathbf{d}_\nu^i - \mathbf{m}_\nu)(\mathbf{d}_\nu^i - \mathbf{m}_\nu)^T\}}{|2\pi C_\nu|^{1/2}}, \quad (1)$$

where  $\text{etr}\{X\} = \exp(\text{trace}(X))$  and  $|2\pi C_\nu|^{1/2} = \sqrt{\det(2\pi C_\nu)}$ . Note that the identity  $\text{trace}(XYZ) = \text{trace}(ZXY) = \text{trace}(YZX)$ . The likelihood for all data points  $i = 1, \dots, N_\nu$  in group  $\nu$  is given by

$$p(D_\nu | \mathbf{m}_\nu, C_\nu) = |2\pi C_\nu|^{-N_\nu/2} \text{etr}\{-\frac{1}{2}C_\nu^{-1}\bar{C}_\nu\}, \quad (2)$$

$$\bar{C}_\nu = \sum_{i=1}^{N_\nu} (\mathbf{d}_\nu^i - \mathbf{m}_\nu)(\mathbf{d}_\nu^i - \mathbf{m}_\nu)^T, \quad (3)$$

where  $D_\nu$  represents all data points in group  $\nu$ ,  $\{\mathbf{d}_\nu\}$ .  $\bar{C}_\nu$  is proportional to the sample covariance of group  $\nu$ . The covariance matrix  $C_\nu$  specifies the measurement errors and the linear relation between the variates.

### 2.1. Grouping hypotheses

The problem of grouping data according to the linear model translates into hypothesis  $H_n$ , which contains the following information:

- The data belong to  $n$  different groups as specified by  $\mathbf{d}_v^i$ .
- The number of elements in group  $\nu$  is  $N_\nu$ .
- Group  $\nu$  has mean  $\mathbf{m}_\nu$  and covariance  $C_\nu$ .
- The means  $\mathbf{m}_\nu$  of different groups are different and unknown.
- The covariance matrices  $C_\nu$  can be known or unknown. They can be common or different for all groups.

A known covariance matrix arises, e.g. for data which scatter only due to known measurement errors. Covariance matrices for the present applications are unknown.

The probability for hypothesis  $H_n$  is according to the Bayes theorem

$$P(H_n | D) \propto p(D | H_n) P(H_n). \quad (4)$$

The posterior probability for grouping hypothesis  $H_n$  factors into a marginal likelihood probability distribution function (pdf)  $p(D | H_n)$  and a prior pdf  $P(H_n)$ . The likelihood is marginalized with respect to the unknown parameters of the mean vectors  $\mathbf{m}_\nu$  and, if unknown, to the covariance matrices  $C_\nu$ . The proportionality constant drops out in model comparison and needs no further consideration.

### 2.2. Marginal likelihood for known covariances

We assume, for the moment, that the covariance matrices are known and marginalize over the unknown mean vectors only. The marginal likelihood pdf for known covariance matrices is given by

$$\begin{aligned} p(D | H_n, \{C_\nu\}) &= \prod_{\nu=1}^n \int d^L \mathbf{m}_\nu p(D_\nu | H_n, \mathbf{m}_\nu, C_\nu) p(\mathbf{m}_\nu | H_n) \\ &\approx p(\{\bar{D}_\nu\} | H_n) (2\pi)^{L(n-N)/2} \prod_{\nu=1}^n (N_\nu^{-L/2} |C_\nu|^{(1-N_\nu)/2} \text{etr}\{-\frac{1}{2} C^{-1} \bar{C}_\nu\}), \end{aligned} \quad (5)$$

where  $N = \sum_{\nu=1}^n N_\nu$  is the total number of data points. The sample mean vector  $\bar{D}_\nu$  for group  $\nu$  is

$$\bar{D}_\nu = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \mathbf{d}_\nu^i. \quad (6)$$

Equation (5) simplifies with the assumption of a common covariance matrix for all groups,  $C_\nu = C$ , to

$$p(D | H_n, C) \approx p(\{\bar{D}_\nu\} | H_n) (2\pi)^{L(n-N)/2} \left( \prod_{\nu=1}^n N_\nu^{-L/2} \right) |C|^{(n-N)/2} \text{etr}\{-\frac{1}{2} C^{-1} \bar{C}^{(n)}\}, \quad (7)$$

where the definition of the global sample covariance  $\bar{C}^{(n)}$  is

$$\bar{C}^{(n)} = \sum_{v=1}^n \bar{C}_v. \quad (8)$$

To avoid expressions that are difficult to read, the prefactors of the covariance matrices have been chosen different from the standard form found in textbooks. The group sample covariance matrices as well as the total sample covariance matrix depend on the grouping information of  $H_n$ .

The approximations employed in equations (5) and (7) are based on the assumption that the prior of the mean vectors  $p(\{\mathbf{m}_v\} | H_n)$  is slowly varying when compared with the likelihood [4]. We have chosen a flat prior for the mean within the range of the data. Hence, the prior for the mean vectors  $p(\{\mathbf{m}_v\} | H_n)$  is

$$p(\{\mathbf{m}_v\} | H_n) = \frac{\prod_{v=1}^n \Theta[\min(D) < \mathbf{m}_v < \max(D)]}{(\prod_{l=1}^L V_l)^n}, \quad (9)$$

where  $\Theta = 1$  ( $\Theta = 0$ ) if the argument is true (resp. false). The interval  $V_l = \max(D(l)) - \min(D(l))$  for the characteristic  $l$  arises from normalization. The constant prior for the mean values over the range of the data is chosen, since estimated mean values are always within the range of the data.

Note that, in [4], there are two typographical errors in the exponents of equation (6), which has to be compared with equation (7) in the present paper.

### 2.3. Marginal likelihood for unknown covariances

For the most common cases where the covariance matrices are unknown, they have to be treated as nuisance (hyper-) parameters. Within the Bayesian framework, the nuisance parameters are marginalized. We distinguish two cases: first, we assume a common covariance matrix for all groupings; second, we allow for different covariance matrices for different groupings.

When all groups are assumed to have a common unknown covariance matrix, the marginal likelihood is

$$\begin{aligned} p(D | H_n) &= \int d^{L \times L} C p(D | H_n, C) p(C) \\ &\approx p(\{\bar{D}_v\} | H_n) \pi^{nL/2 + (L(L-1) - 2NL)/4} \left( \prod_{v=1}^n N_v^{-L/2} \right) |\bar{C}^{(n)}|^{(n-N)/2} \\ &\quad \times \left[ \prod_{l=1}^L \Gamma\left(\frac{N - n - l + 1}{2}\right) \right], \end{aligned} \quad (10)$$

which is proportional to equation (11) in [4]. The prior  $p(C)$  is chosen to be a multivariate Jeffreys' prior

$$p(C) \propto |C|^{-(L+1)/2}, \quad (11)$$

which is based on Jaynes' transformation group approach on all possible similarity transformations [5]. It provides a functional form, which is invariant under all transformations

that do not change our state of ignorance. The proportionality constant drops out in the posterior odds ratio since it is the same for all hypotheses under consideration. The type of integral in equation (10) is known from the inverse-Wishart distribution [6].

The group analysis of PECVD examples shown in the present paper is based on equation (10) by assuming a common covariance matrix. The number of data vectors in some of the preformed groups  $N_v$  is less than the number of characteristics  $L$ , i.e.  $N_v < L$ , which provides insufficient information to estimate individual covariance matrices uniquely. Nevertheless, we provide a marginal likelihood for the second case.

The improper multivariate Jeffreys' prior can no longer be used where we have different covariances for different groups. The proportionality constant does not drop out in the odds ratio when we change the number of groupings. A useful proper prior for the covariances is given by the inverse-Wishart distribution. The inverse-Wishart distribution is the conjugate prior distribution for the multivariate normal covariance matrix [7].

The marginal likelihood  $p(D | H_n)$  substituted into equation (4) using the inverse-Wishart distribution as prior for  $C_v$  is

$$\begin{aligned}
 p(D | H_n) &= \int \left( \prod_{v=1}^n d^{L \times L} C_v \right) p(D | H_n, \{C_v\}) \left( \prod_{v=1}^n p(C_v) \right) \\
 &\approx p(\{\bar{D}_v\} | H_n) \pi^{L(n-N)/2} |S|^{n\mu/2} \\
 &\quad \times \prod_{v=1}^n \left\{ N_v^{-L/2} |S + \bar{C}_v|^{-(\mu+N_v-1)/2} \left[ \prod_{l=1}^L \frac{\Gamma((\mu + N_v - l)/2)}{\Gamma((\mu + 1 - l)/2)} \right] \right\} \quad (12)
 \end{aligned}$$

where the degree of freedom  $\mu$  and the scale matrix  $S$  are parameters of the inverse-Wishart distribution. To avoid a degenerate form, the degree of freedom must fulfil the criterion  $\mu \geq L$ . Details on the inverse-Wishart distribution can be found in [6, 7].

#### 2.4. Splitting or combining groups

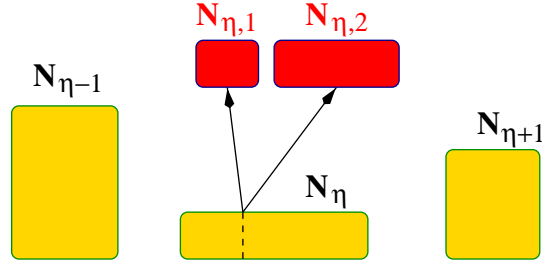
Now we are prepared to answer the key question of the introduction: what is the best classification of data points into different groups? Are there outliers which form a separate group with only a single data point? We start with the evaluation of preselected groups.

Figure 1 shows the transition from hypothesis  $H_n$  into hypothesis  $H_{n+1}$ . One of the preselected groups, group  $\eta$  with  $N_\eta$  data points, is split into two sub-groups  $\eta_1$  and  $\eta_2$  with  $N_{\eta_1}$  and  $N_{\eta_2}$  data points.

The comparison of the two hypotheses  $H_n$  and  $H_{n+1}$  is accomplished with the posterior odds ratio

$$\frac{P(H_{n+1} | D)}{P(H_n | D)} = \frac{P(H_{n+1})}{P(H_n)} \frac{P(D | H_{n+1})}{P(D | H_n)}. \quad (13)$$

The prior odds quantifies the weighting we want to give in favour of one of the two alternatives independent of the data. Throughout the paper, we have chosen the prior odds to be unity since we lack the knowledge to favour one hypothesis over the other. Results can easily be updated subsequently if expert knowledge favours one grouping over another. Note that model complexity cannot be completely attributed to prior odds. Model complexity is automatically attributed to the



**Figure 1.** Schematic illustration of the splitting of group  $\eta$ .

marginal likelihoods according to the volume of the prior covered by the likelihood distribution. The ratio of marginal likelihoods is called *Bayes factor*. Thus, the posterior odds ratio equals the ratio of the prior odds times the Bayes factor. An increase in prior volume, e.g. due to additional parameters, is automatically penalized (*Occam's razor*) if the data fit does not benefit significantly from this extra parameter space.

The posterior odds ratio for marginalized mean vectors and a marginalized common covariance matrix for all groupings [4] is given by

$$\begin{aligned} \frac{P(H_{n+1} | D)}{P(H_n | D)} &= \frac{P(H_{n+1})}{P(H_n)} \frac{P(\{\bar{D}_v\} | H_{n+1})}{P(\{\bar{D}_v\} | H_n)} \left( \pi \frac{N_\eta}{N_{\eta 1} N_{\eta 2}} \right)^{L/2} |\bar{C}^{(n)}|^{1/2} \\ &\times \frac{\Gamma((N+n-L)/2)}{\Gamma((N-n)/2)} \left( 1 - \frac{N_{\eta 1} N_{\eta 2}}{N_\eta} M_d \right)^{-(N-n-1)/2}. \end{aligned} \quad (14)$$

The odds ratio contains the multivariate Student's  $t$  distribution, which depends essentially on the Mahalanobis distance  $M_d = \Delta D_\eta^T (\bar{C}^{(n)})^{-1} \Delta D_\eta$  [1]. The difference in the sample means of the new sub-groups is  $\Delta D_\eta = \bar{D}_{\eta 2} - \bar{D}_{\eta 1}$ .

The posterior odds ratio for marginalized mean vectors and heterogeneous covariance matrices according to equation (12) is given by

$$\begin{aligned} \frac{P(H_{n+1} | D)}{P(H_n | D)} &= \frac{P(H_{n+1})}{P(H_n)} \frac{P(\{\bar{D}_v\} | H_{n+1})}{P(\{\bar{D}_v\} | H_n)} \left( \pi \frac{N_\eta}{N_{\eta 1} N_{\eta 2}} \right)^{L/2} |S|^{L/2} \\ &\times \left[ \prod_{l=1}^L \Gamma\left(\frac{\mu+1-l}{2}\right) \right]^{-1} \left\{ \prod_{l=1}^L \frac{\Gamma((\mu+N_{\eta 1}-l)/2) \Gamma((\mu+N_{\eta 2}-l)/2)}{\Gamma((\mu+N_\eta-l)/2)} \right\} \\ &\times \frac{|S + \bar{C}_\eta|^{(\mu+N_\eta-1)/2}}{|S + \bar{C}_{\eta 1}|^{(\mu+N_{\eta 1}-1)/2} |S + \bar{C}_{\eta 2}|^{(\mu+N_{\eta 2}-1)/2}}. \end{aligned} \quad (15)$$

This completes the formulae for Bayesian group analysis assuming we do not have missing data.

### 2.5. Outlier detection

Outlier detection is a special case of equations (14) and (15). Group  $\eta$  is split into two groups with  $N_{\eta 1} = 1$  and  $N_{\eta 2} = N_\eta - 1$ . A numerical test example can be found in [4].

## 2.6. Missing data

The data sets used so far are assumed to be observed completely. In this section, we consider the Bayesian approach for data with partially unobserved (missing) characteristics. We decompose the total data set,  $D_{\text{tot}}$ , into data vectors that are completely measured, namely  $D_{\text{cmp}}$ , and data vectors that have unmeasured characteristics, namely  $D_{\text{inc}}$ ;  $D_{\text{tot}} = (D_{\text{cmp}}, D_{\text{inc}})$ . The traditional method to deal with missing data is to simply omit those data vectors where only partial characteristics are measured. However one has to be aware that, by neglecting partially measured information, the results are at best less reliable compared with the case where all data are used. Analysing the complete data set is important in cases where the data set mostly contains missing information or, to the extreme, where only data vectors with missing information can be measured, e.g. owing to experimental constraints.

The Bayesian method to deal with missing data  $D_{\text{mis}}$  is given by marginalizing over the unobserved data,

$$P(D_{\text{obs}} | \theta) = \int P(D_{\text{obs}}, D_{\text{mis}} | \theta) dD_{\text{mis}}, \quad (16)$$

where  $\theta$  represents the parameters. The incomplete data vector  $D_{\text{inc}} = (D_{\text{obs}}, D_{\text{mis}})$  is split into an observed and a missing sub-vector of dimensions  $(L_{\text{obs}}, L_{\text{mis}})$ . Again, we assume that the data vector  $D_{\text{inc}, \nu}$  of grouping  $\nu$  is multi-normal-distributed with mean vector  $\mathbf{m}_\nu$  and covariance  $C_\nu$ ,

$$p(D_{\text{obs}, \nu}, D_{\text{mis}, \nu} | \mathbf{m}_\nu, C_\nu) = |2\pi C_\nu|^{-1/2} \exp\{-\frac{1}{2}(D_{\text{inc}, \nu} - \mathbf{m}_\nu)^T C_\nu^{-1} (D_{\text{inc}, \nu} - \mathbf{m}_\nu)\}. \quad (17)$$

We split the inverse covariance matrix  $C_\nu^{-1}$  into four sub-matrices  $U_{\nu i}$ ,  $V_{\nu i}$  and  $W_{\nu i}$ , where  $U_{\nu i}$  ( $V_{\nu i}$ ) denotes the square sub-matrix pertaining to  $D_{\text{obs}, \nu i}$  (resp.  $D_{\text{mis}, \nu i}$ ), and  $W_{\nu i}$  denotes the rectangular sub-matrix of  $C_\nu^{-1}$  pertaining to the correlation between  $D_{\text{obs}, \nu i}$  and  $D_{\text{mis}, \nu i}$ . The marginalized likelihood (equation (16)) reads

$$p(D_{\text{obs}, \nu i} | \mathbf{m}_\nu, C_\nu) = (2\pi)^{-L_{\text{obs}, \nu i}/2} |\tilde{U}_{\nu i}|^{1/2} \exp\{-\frac{1}{2} \mathbf{z}_{\nu i, \text{obs}}^T \tilde{U}_{\nu i} \mathbf{z}_{\nu i, \text{obs}}\}, \quad (18)$$

where  $\mathbf{z}_{\nu i, \text{obs}} = D_{\text{obs}, \nu i} - \mathbf{m}_{\nu i, \text{obs}}$ , the projection matrix  $\tilde{U}_{\nu i} = U_{\nu i} - W_{\nu i}^T V_{\nu i}^{-1} W_{\nu i}$  and  $\mathbf{m}_{\nu i, \text{obs}}$  is the sub-vector of  $\mathbf{m}_\nu$  pertaining to  $D_{\text{obs}, \nu i}$ .

For calculating the grouping probabilities, we have to combine the missing-data likelihood (equation (18)) with the complete-data likelihood (equation (1)) and marginalize over the mean values  $\mathbf{m}_\nu$  and the covariance matrix elements entering  $C_\nu^{-1}$ ,  $U_\nu$ ,  $V_\nu$  and  $W_\nu$ . Note that the residual mean values ( $\mathbf{m}_{\nu i, \text{obs}}$ ) and the sub-matrices ( $U$ ,  $V$ ,  $W$ ) depend on the indices of the missing values of the incomplete data vectors, which may differ for the various incomplete data vectors. Since this multidimensional integration can no longer be performed analytically, we have to use either Monte Carlo techniques or approximate the integration with partial maximum-likelihood estimates for the missing-data likelihoods. We have chosen the second approach since the partial maximum-likelihood approach is easy to implement and we do not expect the full approach to modify the results for the present applications significantly.

The missing-data likelihood is evaluated at the maximum-likelihood estimates of the sample mean vector  $\mathbf{m}_\nu^* = \bar{D}_{\text{cmp}, \nu}$  and the sample covariance matrix  $C_\nu^* = (1/N_\nu) \sum_i^{N_\nu} (D_{\text{cmp}, \nu i} - \bar{D}_{\text{cmp}, \nu})(D_{\text{cmp}, \nu i} - \bar{D}_{\text{cmp}, \nu})^T$ , where the estimates are calculated only from the complete data sets in group  $\nu$ ,  $D_{\text{cmp}, \nu}$ . For a singular sample covariance  $C_\nu^*$  in group  $\nu$ , e.g. for  $N_\nu < L$ , we have to

assume a common covariance structure for all groups. The estimate for a common covariance matrix is given by  $C^* = (1/N) \sum_v N_v C_v^*$ . In principle, the estimates  $m_v^*$  and  $C_v^*$  can be evaluated from all data vectors  $D_{\text{tot}}$ , but the numerics is less straightforward.

For obtaining the marginal likelihood, we have to multiply equation (10) or (12) with the maximum-likelihood replacement of the missing-data likelihood,

$$P(D_{\text{tot}} | H_n) \approx P(D_{\text{cmp}} | H_n) \times P(D_{\text{obs}} | H_n, \{m_v^*, C_v^*\}). \quad (19)$$

The posterior odds ratio for the complete and incomplete data vectors is given by equations (14) and (15) multiplied by the ratio of the maximum-likelihood replacement of the missing-data likelihoods. The maximum-likelihood approximation accounts for the incomplete data in the grouping probability similar to statistical LRTs. But LRT does not account for the full parameter space in model comparison as done with posterior odds ratios. LRT is a test on a single point defined from the estimates, whereas the Bayesian approach marginalizes over the parameter space.

The combination of Bayesian and classical statistical model comparison tests in the present paper forms a trade-off between full Bayesian methods and practicability aspects. For data sets where most or all data vectors have unobserved characteristics, it may become necessary to establish the full Bayesian formalism using Monte Carlo integration techniques.

## 2.7. AIC and BIC

Results of the Bayesian analysis will be compared with two frequently used methods for hypothesis testing and model selection. The likelihood probability increases with the number of groups provided, since an increase in the number of parameters allows for a better fit to the data. The LRT statistics measures the gain in fitting the data with the maximum-likelihood estimates. LRT does not account for model parsimony. An information criterion (IC) is used to favour models with a reduced number of parameters. Both AIC and BIC [8] penalize model complexity by adding an additional term to the maximized log-likelihood,

$$\text{IC} = -2 \ln(p(D | m^*, C^*)) + k * N_p, \quad (20)$$

where  $k = 2$  for AIC and  $k = \ln N_d$  for BIC.  $N_p$  denotes the number of estimated parameters and  $N_d$  the number of observed data. Here,  $N_d$  is the number of data characteristics, i.e.  $N_d = NL$ , and not the number of data vectors  $N$ . Evidence in favour of a model increases with decrease in the value of AIC or BIC.

It has been shown that AIC tends to overestimate the number of parameters needed (see [9] and references therein). Comparing the penalizing terms of AIC and BIC indicates that BIC tends to favour simpler models than those chosen by AIC. A convention for calibrating BIC differences is shown in table 1. The difference in the BIC values for two models may be viewed as a rough approximation to minus twice the logarithm of the Bayes factor for unspecified prior densities. On practical grounds BIC has been proven to be more justified than any other IC. Details of the performance of AIC and BIC can be found in [9].

The approaches LRT, AIC and BIC can be consistently applied only to complete data vectors. They do not provide a unique extension for dealing with missing data [1].

**Table 1.** Calibration of BIC differences in terms of evidence [9].

BIC difference	<2	2–6	6–10	>10
Evidence	Weak	Positive	Strong	Very strong

**Table 2.** Twenty measurements of rare-earth oxide film deposition [10].

Run	Rare earth	Carrier gas	$T_{\text{sub}}$ (°C)	$P_{\text{rf}}$ (W cm <sup>-2</sup> )	$\bar{R}_{\text{dep}}$ (μg cm <sup>-2</sup> min <sup>-1</sup> )	C content (%)	Metal content (rel. theor. value)
12	Y	Ar/H <sub>2</sub> O	350	1.0	5.15	3.25	0.9504
13	Y	Ar/H <sub>2</sub> O	350	1.5	2.2	2.05	0.9949
14	Y	Ar/H <sub>2</sub> O	375	1.0	4.35	1.6	0.9670
15	Y	Ar/H <sub>2</sub> O	400	1.0	3.45	1.55	0.9886
16	Y	Ar/H <sub>2</sub> O	400	1.5	4.8	1.0	0.9975
19	Dy	Ar/H <sub>2</sub> O	400	0.75	4.05	3.7	0.9219
20	Dy	Ar/H <sub>2</sub> O	400	1.5	4.0	1.0	0.9770
21	Er	Ar/H <sub>2</sub> O	400	1.5	3.2	1.0	0.9703
22	Y	N <sub>2</sub> O	350	1.0	3.75	5.3	0.8933
23	Y	N <sub>2</sub> O	350	1.5	3.3	1.0	0.9720
24	Y	N <sub>2</sub> O	400	1.0	3.0	2.3	0.9161
25	Y	N <sub>2</sub> O	400	1.5	3.5	1.0	0.9644
27	Dy	N <sub>2</sub> O	350	1.5	4.05	3.9	0.9460
28	Dy	N <sub>2</sub> O	400	1.0	6.5	6.5	0.9265
29	Dy	N <sub>2</sub> O	400	1.5	3.5	1.0	0.9816
30	Er	N <sub>2</sub> O	350	1.5	3.35	2.8	0.9291
31	Er	N <sub>2</sub> O	400	1.5	4.8	3.3	0.9829
32	Y	CO <sub>2</sub>	400	1.5	1.75	11.3	0.9931
34	Dy	CO <sub>2</sub>	400	1.5	3.35	2.85	0.9506
35	Er	CO <sub>2</sub>	400	1.5	4.1	3.5	0.9543

### 3. Results

#### 3.1. Rare-earth oxide film deposition

In rare-earth oxide film deposition, Weber *et al* [10] have taken 25 measurements with rare-earth components of yttrium, dysprosium and erbium, and the carrier gases carbon dioxide, nitrous oxide and argon/water vapour. Table 2 shows the 20 data vectors with measurements for the metal content. As depicted in table 3, the whole set of data may be arranged either according to the rare-earth species or according to the carrier gas, in either case leading to  $n = 3$  groupings of otherwise identical data. The dimension of the data characteristics is  $L = 5$ . Some predefined groups have fewer data points than data characteristics, i.e.  $N_v \leq L$ , which makes us choose a common covariance structure for all groups. In principle, sparsely occupied groups can be allowed to have the same covariance structure, whereas densely occupied groups may have different covariances. The marginal likelihoods for this situation can be calculated easily, although this is beyond the scope of the present paper.

In table 3, the log-probabilities for the permutations of the preformed data groups are given relative to the case where all data belong to one group. The carrier gases (rare-earth components) in curly braces indicate which preformed groups are combined to one group, respectively.

**Table 3.** Group analysis of rare-earth oxide film deposition data. Data vectors for components within curly braces form groups.  $\{N_v\}$ , number of data vectors in group  $v$ ;  $n$ , number of groups;  $N_p$ , number of parameters;  $P_g$ , grouping probability;  $P_g^{\text{ML}}$ , maximum-likelihood probability.

Choice	$\{N_v\}$	$n$	$N_p$	$\ln P_g$	$\ln P_g^{\text{ML}}$	–AIC	–BIC
One group	20	1	20	0	0	0	0
{Ar/H <sub>2</sub> O, N <sub>2</sub> O} {CO <sub>2</sub> }	17, 3	2	25	2.9	7.2	4.5	–8.6
{Ar/H <sub>2</sub> O} {N <sub>2</sub> O, CO <sub>2</sub> }	8, 12	2	25	2.8	8.8	7.6	–5.4
{Ar/H <sub>2</sub> O, CO <sub>2</sub> } {N <sub>2</sub> O}	11, 9	2	25	–1.5	4.1	–1.7	–14.8
{Ar/H <sub>2</sub> O} {N <sub>2</sub> O} {CO <sub>2</sub> }	8, 9, 3	3	30	4.3	14.8	9.6	–16.5
{Y, Dy} {Er}	16, 4	2	25	–1.8	2.6	–4.9	–17.9
{Y} {Dy, Er}	10, 10	2	25	0.1	5.9	1.8	–11.2
{Y, Er} {Dy}	14, 6	2	25	–3.0	2.0	–6.0	–19.0
{Y} {Dy} {Er}	10, 6, 4	3	30	–2.7	6.9	–6.2	–32.3

Results for the posterior odds,  $P_g$  (where  $g$  stands for grouping), show that the different groupings have similar probabilities. Using the calibration of the BIC differences, where  $\ln P_g$  has to double, three separate groupings with respect to carrier gases show strong evidence ( $2 \times 4.3$ ) and groupings with respect to rare-earth components show positive evidence in favour of all-in-one group ( $2 \times (-1.8 \dots -3.0)$ ). Despite this evidence, the difference in probabilities is small compared with results from other applications [4]. A sensitivity study shows that the posterior odds values for this application depend crucially on the volume of the prior of the mean vectors. We assume that the true mean value of the groups is within the interval spanned by the data. Increasing the prior interval for the mean values by a moderate factor of 2 decreases the grouping probabilities for two or more groups in favour of an all-in-one grouping. Hence, the sensitivity study shows that the different groupings are not well separated. This indicates that the small number of data is not sufficient to favour clearly one grouping over another.

Nevertheless, a trend can be seen. Groupings with respect to carrier gases have more evidence compared with groupings with respect to rare-earth components. There might be a sub-structure related to the carrier gases. In fact, the three-group situation where each carrier gas forms a separate group has the highest probability. However, assuming that we have a certain grouping and inferring physical parameters from that may, therefore, be misleading.

To gain more confidence in our interpretation, the Bayesian results will be compared with the frequently used approaches of LRT, AIC and BIC. Table 3 shows the maximum log-likelihood values ( $\ln P_g^{\text{ML}}$ ) and the negative AIC and BIC quantities (negative for maximization) relative to the values for all-in-one grouping. The maximum log-likelihood values increase with the number of parameters  $N_p$ , as expected. AIC penalizes model complexity proportional to the number of parameters. The ordering of groupings according to AIC follows the Bayesian result. In particular, the grouping where the three carrier gases form separate groups shows the largest maximum log-likelihood, which is not overruled by the penalizing term. In contrast, BIC never favours individual groups. According to the BIC calibration convention, there is at least positive (5.4) to very strong evidence (32) that all data should be grouped together. The reason is that BIC does not consider explicit prior assumptions. Our prior for the covariance matrix is chosen to be uninformative, but the prior for the mean values is chosen to be informative. As addressed

above, we assume that the true mean value of the groups is within the interval of the observed data values. BIC represents an approximation to the Bayes factor with *large* prior volumes. In this sense, BIC gives a conservative estimate of model complexity in the sense of being most economical of all parameters. Hence, evidence from BIC for less parsimonious groupings is rather robust, although support for less complex groupings does not necessarily mean we have to favour them. Informative prior pdfs may overrule the parsimony criteria provided by BIC.

The search for instructive combinations of preformed groups will fail when the initial groups are misclassified. A misclassified initial grouping can be identified by looking for the outlier probability of individual data points. The largest outlier probability of value very close to 1 is found for measurement (run) 32 for all grouping structures proposed, owing to the large carbon content compared with the distribution of others. Measurement 32 can be treated as a separate group. Separation of measurement 32 into a fourth group shows that measurement 13 is also a candidate for an outlier in the sense that it does not fit into the rest of the preformed groups. This is due to the small value of  $\bar{R}_{\text{dep}}$ . The rest of the data vectors do fit within one grouping.

In conclusion, the data set provides a rather poor basis for learning from groupings. This may be due to the fact that we have, in the mean, only six data points within a group of five characteristics. On the one hand, pooling data into one group may conceal valuable physical processes; on the other, an analysis of separated groups of data may imply significance where there is none. Bayesian group analysis helps us to learn about the significant amount of information provided by the data.

### 3.2. Hydrocarbon film deposition

Schwarz-Selinger *et al* [11] have studied PECVD of hydrocarbon films on single-crystalline silicon substrates to examine the dependence of film properties on deposition parameters as well as on the choice of the source gas. Table 4 shows the 21 data vectors with measurements for seven hydrocarbon source gases (methane ( $\text{CH}_4$ ), ethane ( $\text{C}_2\text{H}_6$ ), propane ( $\text{C}_3\text{H}_8$ ), n-butane ( $\text{n-C}_4\text{H}_{10}$ ), iso-butane ( $\text{iso-C}_4\text{H}_{10}$ ), ethylene ( $\text{C}_2\text{H}_4$ ), acetylene ( $\text{C}_2\text{H}_2$ )) and the dc self-bias voltage  $V_b$  that correlates with the ion energy. The source gases differ in hydrogen to carbon ratio (H/C in gas), carbon chain length and hybridization of the carbon atoms. The film properties were characterized by real-time *in situ* ellipsometry, and the composition of the films was determined quantitatively by ion-beam analysis.  $n_p$ ,  $n_s$  and  $k_p$ ,  $k_s$  represent the parallel (p) and perpendicular (s) components of the real (n) and imaginary (k) parts of the refractive index. The density ( $\rho$ ) is calculated from the stoichiometry and film thickness.  $n_{\text{H+C}}$  represents the total particle number density. Details on the experiment can be found in [11]. In addition, Schwarz-Selinger [12] provided data  $I(k)$  for the integral of the extinction coefficient  $k$  of the C–H stretching vibration over the wavelength range 2700–3200  $\text{cm}^{-1}$  for layers deposited at various bias voltages  $V_b$  and source gases (see figure 5 in [11]).

Three data vectors have missing characteristics and they are, at first, dismissed. In a second step, they will be considered.

Predefined groups are naturally given by the categorical variable of the source gas and the quantitative variable of  $V_b$ . The categorical variable of the source gas can be transformed into a quantitative variable by exploiting the H/C ratio. However, one has to be aware of the fact that the H/C ratios provide only partial information about the source gases since the hybridization differs too. There are three input values for  $V_b$  that can be combined in five ways for having an

**Table 4.** Twenty-one measurements of hydrocarbon film deposition [11] for the source gases methane (m), ethane (e), propane (p), n-butane (nb), iso-butane (ib), ethylene (ey) and acetylene (ay). The measured quantities are the dc self-bias voltage  $V_b$  (V), the parallel (p) and perpendicular (s) components of the real (n) and imaginary (k) parts of the refractive index, the density  $\rho$  ( $\text{g cm}^{-3}$ ), the total particle number density  $n_{\text{H+C}}$  ( $10^{28} \text{ m}^{-3}$ ), the ratios H/C and H/(H + C) in the film and the integral of the extinction coefficient  $k$  of a C–H stretching vibration.

Gas	H/C in gas	$V_b$	$n_p$	$n_s$	$k_p$	$k_s$	$\rho$	$n_{\text{H+C}}$	H/C in film	H/(H + C) in film	$I(k)$
m	4	0	1.565	1.580	0.0005	0.0009	1.0	8.5	0.79	0.44	6.61
e	3	0	1.615	1.616	0.0010	0.0015	1.1	9.1	0.75	0.43	6.20
p	2.66	0	1.595	1.6	0.0007	0.0009	1.0	8.5	0.79	0.44	6.88
nb	2.5	0	1.59	1.60	0.0007	0.0009	0.9	8.2	0.89	0.47	7.11
ib	2.5	0	1.59	1.60	0.0015	0.002	1.1	8.9	0.92	0.48	7.06
ey	2	0	1.72	1.74	0.0018	0.0032	1.1	9.5	0.79	0.44	5.99
ay	1	0	1.800	1.805	0.014	0.015	1.4	11.1	0.69	0.41	4.12
m	4	−30	1.807	1.825	0.0165	0.0185	1.5	12.0	0.69	0.41	4.96
e	3	−30	1.845	1.86	0.02	0.022	1.6	12.8	0.69	0.41	4.50
p	2.66	−30	1.72	1.75	0.006	0.0085	–	–	–	–	–
nb	2.5	−30	1.729	1.74	0.0017	0.0022	–	–	–	–	6.23
ib	2.5	−30	1.69	1.71	0.0006	0.0008	–	–	–	–	6.41
ey	2	−30	1.945	1.96	0.019	0.021	1.7	13.7	0.64	0.39	4.44
ay	1	−30	2.09	2.1	0.037	0.045	1.8	14.0	0.54	0.35	2.63
m	4	−200	2.1	2.20	0.075	0.10	1.7	12.1	0.43	0.33	2.55
e	3	−200	2.22	2.29	0.1	0.11	1.9	13.2	0.47	0.32	2.95
p	2.66	−200	2.23	2.23	0.096	0.14	2.2	14.8	0.41	0.29	2.70
nb	2.5	−200	2.25	2.3	0.086	0.13	2.1	14.6	0.45	0.31	2.70
ib	2.5	−200	2.26	2.28	0.085	0.10	1.9	13.2	0.43	0.30	2.59
ey	2	−200	2.335	2.4	0.1	0.15	2.2	14.8	0.37	0.27	2.11
ay	1	−200	2.455	2.5	0.13	0.14	2.4	14.6	0.27	0.21	1.35

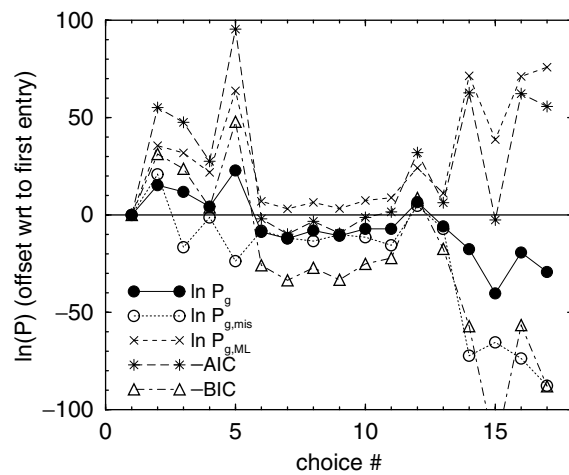
all-in-one grouping up to having three single groups. For the seven source gases, the number of combinations is 877, which, for practical reasons, restricts the number of combinations that can be shown explicitly.

Table 5 shows a subset of possible arrangements of the source gases and all arrangements with respect to  $V_b$ . The dimension of the data characteristics is  $L = 8$ . From the first two characteristics, ‘H/C in gas’ is used for groupings w.r.t.  $V_b$  and the characteristic  $V_b$  is used for groupings w.r.t. the source gases. Only two of the four characteristics  $\rho$ ,  $n_{\text{H+C}}$ , ‘H/C in film’ and ‘H/(H + C) in film’ are independent. From the four dependent values, the density  $n_{\text{H+C}}$  and the ratio ‘H/C in film’ are chosen as the independent data for group analysis. As in the previous example, some of the predefined groups have fewer data vectors than data characteristics, i.e.  $N_v \leq L$ , which makes us choose a common covariance structure for all groupings. Figure 2 depicts the probabilities shown in table 5 w.r.t. the choice (grouping) number.

There is clear evidence that the groups sorted w.r.t.  $V_b$  have to be treated as separate groups. Among the groupings 1–5, the maximum at  $\ln P_g = 23$  is for choice 5 (see also the first five closed circles in figure 2). This result is insensitive to changes in the prior support and is supported by

**Table 5.** Group analysis of hydrocarbon film deposition data. Data vectors for components within curly braces form groups.  $P_g^{\text{mis}}$ , grouping probability including data vectors with missing information.

Choice	$\{N_v\}$	$n$	$N_p$	$\ln P_g$	$\ln P_g^{\text{mis}}$	$\ln P_g^{\text{ML}}$	-AIC	-BIC
1. One group	18	1	35	0	0	0	0	0
2. $\{0, -30\} \{-200\}$	11, 7	2	42	15	21	36	55	31
3. $\{0\} \{-30, -200\}$	7, 11	2	42	12	-17	32	48	24
4. $\{0, -200\} \{-30\}$	14, 4	2	42	4.3	-1.3	22	28	3.9
5. $\{0\} \{-30\} \{-200\}$	7, 4, 7	3	49	23	-24	64	95	48
6. $\{e, p, nb, ib, ey, ay\} \{m\}$	15, 3	2	42	-8.7	-8.1	7.0	-2.0	-26
7. $\{m, p, nb, ib, ey, ay\} \{e\}$	15, 3	2	42	-12	-12	3.1	-9.8	-34
8. $\{m, e, nb, ib, ey, ay\} \{p\}$	16, 2	2	42	-8.0	-13	6.4	-3.3	-27
9. $\{m, e, p, ib, ey, ay\} \{nb\}$	16, 2	2	42	-11	-10	3.3	-9.4	-33
10. $\{m, e, p, nb, ey, ay\} \{ib\}$	16, 2	2	42	-7.1	-11	7.4	-1.3	-25
11. $\{m, e, p, nb, ib, ay\} \{ey\}$	15, 3	2	42	-7.2	-16	8.8	1.5	-22
12. $\{m, e, p, nb, ib, ey\} \{ay\}$	15, 3	2	42	6.6	4.7	24	32	8.7
13. $\{m, e, p, ey, ay\} \{nb, ib\}$	14, 4	2	42	-5.9	-7.2	11	6.3	-17
14. $\{p\} \{nb\} \{ib\} \{ey\} \{ay\} \{m, e\}$	3, 2, 2, 2, 3, 6	6	70	-18	-72	71	63	-57
15. $\{m\} \{e\} \{p\} \{ib\} \{ey\} \{nb, ay\}$	3, 3, 2, 2, 3, 5	6	70	-40	-65	39	-2.6	-121
16. $\{m\} \{e\} \{p\} \{ey\} \{ay\} \{nb, ib\}$	3, 3, 2, 3, 3, 4	6	70	-19	-74	71	62	-57
17. $\{m\} \{e\} \{p\} \{nb\} \{ib\} \{ey\} \{ay\}$	3, 3, 2, 2, 2, 3, 3	7	77	-29	-88	76	56	-88



**Figure 2.** Group analysis of hydrocarbon film deposition data.

AIC and BIC values (stars and triangles in figure 2). The logarithms of the maximum-likelihood probabilities  $\ln P_g^{\text{ML}}$  (crosses) only reflect the fit to the data, which becomes better the more parameters are provided, but does not include a term penalizing model complexity.

The support for three individual groups is due to the large distance between the mean values of the eight characteristics compared with the correlation lengths within the groups. The strong evidence for individual groups should not necessarily be interpreted as due to the presence of a hidden mechanism for film deposition that distinguishes different  $V_b$  values. The result can

be interpreted as a consequence of the large difference in  $V_b$  values between the groups. This interpretation is supported by the large probabilities for combinations of data for neighbouring  $V_b$  values as compared with the data set where  $V_b = 0\text{ V}$  and  $V_b = -200\text{ V}$  are grouped. The grouping only reflects the dependence of the film properties on  $V_b$ , which is not evenly covered. For learning more about covariances related to changes in the self-bias voltage  $V_b$ , additional measurements covering the intermediate values are necessary. In this regard, Bayesian group analysis helps to identify the experimental design parameters, which should be measured on a finer grid to avoid physical misinterpretation of clusters of data that are only due to the choice of the design parameters.

The situation changes when the partially observed data vectors are added. Separate groups with respect to  $V_b$  are no longer supported. There is clear evidence only for the grouping  $\{0, -30\}$   $\{-200\}$ . The reasons are that the incomplete data values are not within the range of other data values in that group and that there is a large overlap of the refractive index of the data with  $V_b = 0\text{ V}$  and the incomplete data with  $V_b = -30\text{ V}$ . The approaches LRT, AIC and BIC are not useful here, since they do not provide a reasonable extension for dealing with missing data.

There is no evidence that the data pre-sorted with respect to the source gases form separate groups, except for grouping number 12. Acetylene (ay,  $\text{C}_2\text{H}_2$ ) has to be sorted into a separate group. The reason is given by the characteristic  $I(k)$ , which is systematically smaller for acetylene compared with other source gases.

The same data set without  $I(k)$  provides no clear evidence for acetylene forming a separate group in the sense of a positive value for  $\ln P_g$ . However, even for this reduced data set, there is weak evidence for acetylene playing a special role, since  $\ln P_g$  is approximately 4–8 times larger for acetylene in a separate group than that for other source gases (values not shown in table 5). The extinction coefficient of the C–H stretching vibrations for the films deposited from acetylene is mainly responsible for forming a separate group in acetylene.

Separating other source gases than acetylene from the joint group yields smaller grouping probabilities. Separating all source gases into individual groups or combining only two source gases within one group is clearly ruled out by small grouping probabilities; the largest and smallest grouping probability for six groups is explicitly shown in table 5. There is no evidence of a separate group consisting of n-butane (nb) and iso-butane (ib). In contrast with the groupings w.r.t.  $V_b$ , the results for groupings w.r.t. the source gases are not affected by the missing data vectors. Including incomplete data vectors still supports acetylene forming a separate group.

The physical reason for acetylene forming a separate group is given by the decreased H/C ratio. The physical properties of the adsorbed films mainly depend on the H/C film ratio, which depends on the H/C ratio of the source gas [11]. Although the data set contains the ‘H/C in film’ characteristic, the interdependence between  $I(k)$  and the H/C ratio of the source gas is most prominent.

AIC favours individual groups and provides strongest evidence for some groupings with six and seven groups. AIC fails to penalize model complexity properly, owing to the well-known behaviour of overestimating the number of parameters when the number of data is high. In addition, AIC shows a large scatter of the evidences compared with results from the Bayesian approach (cf  $\ln P_g$  with  $-AIC$ ). This can be clearly seen for the groupings 13–17. This is due to the fact that the AIC evidence is calculated at a single point, i.e. the maximum-likelihood value ( $\ln P_g^{\text{ML}}$ ) for the parameters. The Bayesian approach marginalizes over the parameters. This takes into account the full variability of the parameter values, which are consistent with the data.

BIC confirms the full Bayesian result with even stronger evidence. As explained above, this is due to an uninformative prior pdf implicitly used in deriving BIC. BIC seems to provide a suitable approximation to the full Bayesian approach at least for the present applications.

The result for the hydrocarbon film deposition is that the present data set does not support different deposition mechanisms depending on the type of source gas. This quantitative result based on grouping probabilities with multivariate distributions coincides with the conclusions of Schwarz-Selinger *et al* [11]. The choice of the source gas has, indeed, a distinct influence on the optical properties of the PECVD hydrocarbon films. However, 'all films can be characterized by their unique correlation between their hydrogen content  $H/(H + C)$  and other film properties such as optical constants, density, etc. On the basis of a given film structure, it is therefore not possible to deduce the precursor gas used for deposition . . .'. Bayesian probability theory allows one to quantify this statement with grouping probabilities using complete and incomplete data vectors.

#### 4. Discussion

A major question in the two applications of Bayesian group analysis on PECVD data is the evidence for genuine groups. Is it necessary to pool the data or are there indications for real groups which have to be studied separately? The relation of the within-group variability to the between-group variability determines if the groups can be treated as compact and distinct. The significance for individual groups increases with the separation of the groups with respect to the group extensions. A central goal of the present applications was to clarify if different physical mechanisms can be identified for the various choices of experimental design parameters. Both the rare-earth oxide and the hydrocarbon film deposition data do not support different deposition mechanisms depending on the type of source gas (acetylene will be discussed below). The reason may be a lack of real groups or a sample size that is too small. Both applications show a rather poor ratio of the number of data points to the number of characteristics. In designing future experiments to find real grouping structures, significant sample size relative to the dimension of the parameter space should be considered. Bayesian group analysis on simulated data sets may help in identifying reasonable design parameters.

The reason for acetylene forming a separate group is the large relative distance of the H/C ratio of acetylene to the H/C ratio of the other hydrocarbons. It is expected that an extended data set including hydrocarbons with H/C ratios between 1 and 2 would fill up the gap and would pool all data into one group. The same is expected to occur if the self-bias voltage is sampled on a finer grid. The analysis shows that neither the carbon chain length nor the hybridization of the carbon atoms is identified to cause a real group structure.

Irrespective of the lack of evidence of physically real groups, the method found outliers. Outliers are identified by the necessity to form a separate group. Auxiliary methods to identify outliers with their subsequent deletion are not required. Outliers are self-consistently treated along with the regular data.

Bayesian group analysis, like Bayesian model selection in general, is readily understandable by even non-statisticians, since interpretation is based on probabilities. Bayesian model selection allows one to find evidence *in favour of* a model or hypothesis, whereas classical approaches such as *p* values test a model *against* a null hypothesis. It is well known that classical model selection schemes are difficult to interpret properly [13]. In addition, classical model selection

tools, such as  $p$  values and AIC, do not guarantee consistency. The Bayesian scheme guarantees selection of the true model if the true model is among the studied models and if enough data are observed. A general discussion of the Bayesian approach to model selection in comparison with classical schemes can be found in [13].

The present work compares the full Bayesian approach, which integrates over parameter space, with two frequently used methods for hypothesis testing and model comparison, namely AIC and BIC. The classical approaches work at the single point of the most probable parameter value, neglecting the parameter space with smaller posterior probabilities. Since the most probable value is not the *true* parameter value but only an estimate of the true value, the complete parameter space compatible with the grouping structure has to be considered for the validation of grouping. Additionally, integration over the parameter space introduces a property called Occam's razor, which automatically penalizes model complexity in favour of simpler models. Both the classical method AIC and the approximate Bayesian method BIC provide an additional penalizing term for model complexity. Results from the present applications show that AIC provides a poor criterion for decreasing the number of parameters properly. As is known from the literature [9], AIC tends to overestimate the number of parameters needed. Hence, it is not useful for group analysis. Comparison of the penalizing terms of AIC and BIC indicates that BIC performs better in decreasing the number of parameters. However, the approximate Bayesian model selection scheme BIC is known to be conservative, since it may underestimate the number of parameters (groups). The present applications provide informative prior knowledge of the mean values of the groups. Informative prior knowledge of parameters weakens the penalizing Occam factor in Bayes factors, which may result in support for more groups than estimated from BIC. Nevertheless, BIC can be used for a fast check of grouping structures if the approximative Bayesian criterion is calibrated for the typical application to be analysed.

The drawback of BIC (similar to AIC) is that it cannot be applied to incomplete data sets. BIC does not provide a recipe for dealing with missing data. As shown in the second application, considering data vectors with missing entries may alter the conclusions drawn from complete data sets only. Incomplete measurements may be of special importance, since the physical situation can be at an operational threshold such that only parts of the characteristics can be measured reliably. In the present case, the *in situ* measurements could be performed, but the *ex situ* measurements were no longer possible due to the fragility of the deposited films. Furthermore, combining data sets from various experiments or combining recent experiments with historical data typically results in incomplete data sets. In those situations, a combined analysis is possible only with a consistent approach for handling missing data. The Bayesian method introduces and marginalizes the missing data values to propagate their full uncertainties into the final results. The observed parts of the incomplete data vectors do, at least, increase the reliability of the result, but can also change the evidence of grouping structures as observed with the hydrocarbon film deposition data.

## 5. Conclusion

Bayesian probability theory was applied to quantify the grouping probabilities of a set of data vectors within a multivariate model. The approach was extended to allow for arbitrary covariance matrices for different groups and to handle missing data. Outliers can be identified by the necessity to form a separate group. The grouping probabilities are compared with classical approaches of

LRTs and AIC and a Bayesian variant called BIC. At least for the present applications, BIC has proven to be a useful and easy to implement tool for group analysis. The method was applied to PECVD data of rare-earth oxide film deposition and hydrocarbon film deposition to study the evidence of grouping structures attributed to categorical quantities such as rare-earth components or source gases and quantitative variates as bias voltage.

Finally, Bayesian group analysis provides a quantification of the evidence of structures within a set of (incomplete) data vectors. These structures are often identified by naked eye or by simple two-parameter plots in the case of low-dimensional parameter spaces. Bayesian group analysis is consistent with intuition; however, in high-dimensional parameter spaces, intuition has to be supplemented with a Bayesian approach.

### Acknowledgments

The author is indebted to V Dose, W Jacob and T Schwarz-Selinger for stimulating discussions and the latter two for providing the data. S Gori is also acknowledged for solving the combinatorial problem on ‘How many groupings can be performed with  $N$  elements’.

### References

- [1] Krzanowski W J 1988 *Principles of Multivariate Analysis* (Oxford: Clarendon)
- [2] von Keudell A, Annen A and Dose V 1997 *Thin Solid Films* **307** 65
- [3] Dose V 1993 *Appl. Phys. A* **56** 471
- [4] von der Linden W, Dose V and Ramaswami A 1998 *Maximum Entropy and Bayesian Methods* ed G J Erickson, J T Rychert and C R Smith (Dordrecht: Kluwer) p 87
- [5] Jaynes E T 1983 *E T Jaynes: Papers on Probability, Statistics and Statistical Physics* ed R D Rosenkrantz (Dordrecht: Reidel) p 114
- [6] O’Hagan A 1994 *Kendall’s Advanced Theory of Statistics, Bayesian Inference* 1st edn (New York: Wiley) p 293ff
- [7] Gelman A, Carlin J B, Stern H S and Rubin D B 1995 *Bayesian Data Analysis* (London: Chapman and Hall)
- [8] Schwarz G 1978 *Ann. Stat.* **6** 461
- [9] Kass R E and Raftery A E 1995 *J. Am. Stat. Assoc.* **90** 773
- [10] Weber A, Suhr H, Schumann H and Köhn R 1990 *Appl. Phys. A* **51** 520
- [11] Schwarz-Selinger T, von Keudell A and Jacob W 1999 *J. Appl. Phys.* **86** 3988
- [12] Schwarz-Selinger T 2003 private communication
- [13] Berger J and Pericchi L 2001 *Model Selection* ed P Lahiri (Institute of Mathematical Statistics Lecture Notes—Monograph Series vol 38) (Beachwood, OH: Institute of Mathematical Statistics) p 135