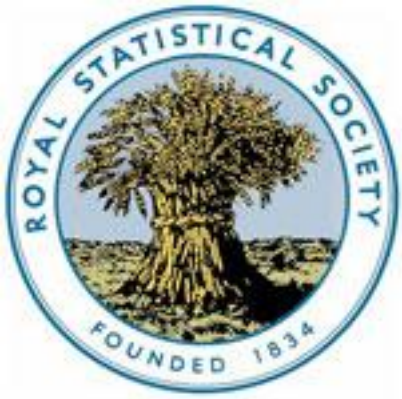

The Statistician (1994)
43, *No.* 1, *pp.* 111–127

Statistical analysis of spectra from electron spectroscopy for chemical analysis

By CHRISTIAN RITTER†

Université Catholique de Louvain, Belgium

[Received July 1992. Revised September 1993]

SUMMARY

Electron spectroscopy for chemical analysis is a key technique in the study of modified material surfaces. Analysis of the resulting spectra consists in decomposing multiple peaks into subpeaks, whose functional form is known up to a few parameters. Statistical inference consists in estimating these parameters and derived quantities, such as peak area ratios, and in assessing the accuracies of these estimates. Purely likelihood-based peak decomposition is notorious for problems of identifiability. In practice, however, additional knowledge exists about some of the parameters and can, when incorporated in the model as an informative prior, produce unique decompositions. Modern tools for Bayesian statistics, such as profile diagnostics, Laplacian approximations of marginals and Markov chain algorithms for sampling from the posterior, can then be used to obtain inference reaching beyond point estimates and approximate standard errors.

Keywords: Additive Gaussian–Lorentzian mixtures; Laplacian approximation; Markov chain simulation; Penalized likelihood; Prior information; Profile methods; Weighted non-linear least squares

1. General introduction

This is a study of a practical problem for which prior information arises naturally and must be included in the analysis to provide sensible results. In the first part, we introduce electron spectroscopy for chemical analysis (ESCA) and show how the problem of analysing ESCA spectra can be formulated in a Bayesian framework. In the second part, we conduct a Bayesian analysis of a specific data set. We use a wide range of techniques, beginning with simple normal approximations, and ending with Markov chain simulations. As a result, we arrive at a much deeper understanding of the ESCA problem.

In many research problems, data acquisition and evaluation proceed in stages. The results of the previous stage carry on to the next stage, and knowledge accrues over time. There is a great need for statistical concepts which are consistent with this structure. The Bayesian approach has this capacity, provided that the corresponding computations can be conducted in a generic fashion, allowing wide flexibility in the choice of the statistical model.

We demonstrate, in the context of the ESCA problem, that such a Bayesian analysis environment can be assembled from available building blocks. Simple and approximate methods, such as normal approximations about the posterior mode, can be augmented by tiers of more sophisticated, more computing-intensive and nearly exact techniques, such as profile methods, marginalization using the Laplacian approximation and Markov chain simulations. In combination, these methods allow problem-driven statistical analyses, not only for the ESCA problem, but for a much wider range of applications.

† *Address for correspondence*: Center for Operations Research and Econometrics, and Institut de Statistique, Université Catholique de Louvain, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium.
E-mail: ritter@stat.ucl.ac.be

 0039–0526/94/43111

2. Introduction to electron spectroscopy for chemical analysis for polymers

ESCA is an important tool for studying the chemistry of the surfaces of materials. In the ESCA process, these surfaces are irradiated with low energy X-rays, causing the emission of photoelectrons from the surface, which can be detected and classified by their kinetic energies. ESCA spectra give the counts y_i^* of photoelectrons at each of the energies x_i, $i = 1, \ldots, n$, over a desired energy range $[x_a, x_b]$ for a prespecified duration of irradiation.

The energy retained by the photoelectrons is the difference between the energy of the X-ray photons and the energy needed for freeing the electrons from their orbitals. Chemical compositions and bonding structures give characteristic spectra. Therefore, ESCA can be applied to *identify* elements and chemical structures.

Furthermore, in different bonding configurations, orbitals of the same element, such as 1s carbon in carbon–carbon, carbon–oxygen or carbon–nitrogen bonds, have slightly different binding energies, thereby splitting the ESCA peak into subpeaks. The relative areas of those subpeaks are proportional to the frequencies of the corresponding bonds in the molecule. Therefore, ESCA spectroscopy can also be used in *quantitative* analyses of complex polymer surfaces.

3. Structure of data and statistical model

Provided that we choose sufficiently long scanning times, the distribution of the y_i^* can be approximated by a normal distribution with energy-dependent mean $\eta(x_i)$ and variance $w_i^{-1} \approx y_i^*$. Often, a background b_i is subtracted from the raw signal y_i^* to form the adjusted counts $y_i = y_i^* - b_i$. A parametric model $\eta(x_i) = \eta(x_i; \boldsymbol{\theta})$ leads to the likelihood

$$L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}, \mathbf{w}) \propto \exp\{-\tfrac{1}{2}S_{\mathbf{x},\mathbf{y},\mathbf{w}}(\boldsymbol{\theta})\}$$

with the weighted sum of squares

$$S_{\mathbf{x},\mathbf{y},\mathbf{w}}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{y_i - \eta(x_i; \boldsymbol{\theta})\}^2 w_i,$$

which we simply call $S(\boldsymbol{\theta})$, and therefore to weighted non-linear least squares as a fitting method.

A commonly used template for ESCA subpeaks of simple non-metallic elements, such as carbon, is the additive Gaussian–Lorentzian function (Wertheim and DiCenzo, 1986) of the form

$$f(x; \beta, \gamma, \rho) = \rho \exp\left[-\ln 2\left\{\frac{2(x-\beta)}{\gamma}\right\}^2\right] + (1-\rho)\left[1 + \left\{\frac{2(x-\beta)}{\gamma}\right\}^2\right]^{-1}.$$

Here, β is the location of the peak, $\gamma > 0$ the bandwidth (full width at half-maximum) and $0 < \rho < 1$ the mixture ratio. The maximum height of this function is 1; its integral is

$$\gamma\left\{\rho\sqrt{\left(\frac{\pi}{4 \ln 2}\right)} + (1-\rho)\frac{\pi}{2}\right\}.$$

Assuming linear superposition for a known number m of subpeaks, we obtain the model

$$\eta(x; \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho}) = \sum_{j=1}^{m} \alpha_j f(x; \beta_j, \gamma_j, \rho_j), \qquad j = 1, \ldots, m.$$

The choice of parameters α_j, γ_j and ρ_j subject to the constraints $0 < \alpha_j$, $0 < \gamma_j$ and $0 < \rho_j < 1$ makes the minimization of $S(\boldsymbol{\theta})$ and subsequent inference difficult. This can be avoided

by replacing $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\rho})$ with $(\mathbf{a}, \mathbf{b}, \mathbf{g}, \mathbf{r})$ defined as $\alpha_j = \exp a_j$, $\beta_j = b_j$, $\gamma_j = \exp g_j$ and $\rho_j = \exp r_j/(1 + \exp r_j)$.

4. Difficulties related to peak decomposition and model specification

There are two main complications in the analysis of ESCA spectra. First, some of the electrons are scattered and lose energy, owing to collisions in the top atomic layers of the surface. These form a background intensity in the observed spectra which is difficult to model or to adjust for. Second, peaks from orbitals of similar bonds overlap, producing composite peaks which require curve fitting for analysis. The difficult problems involving peak decomposition and background removal have been under investigation for many years. Some of the key references are Perram (1968), Vanderginste and De Galan (1978), Baruya and Maddams (1978), Maddams (1980) and Powell and Seah (1990).

A typical example of overlapping subpeaks is presented in Fig. 1. The two plots show a

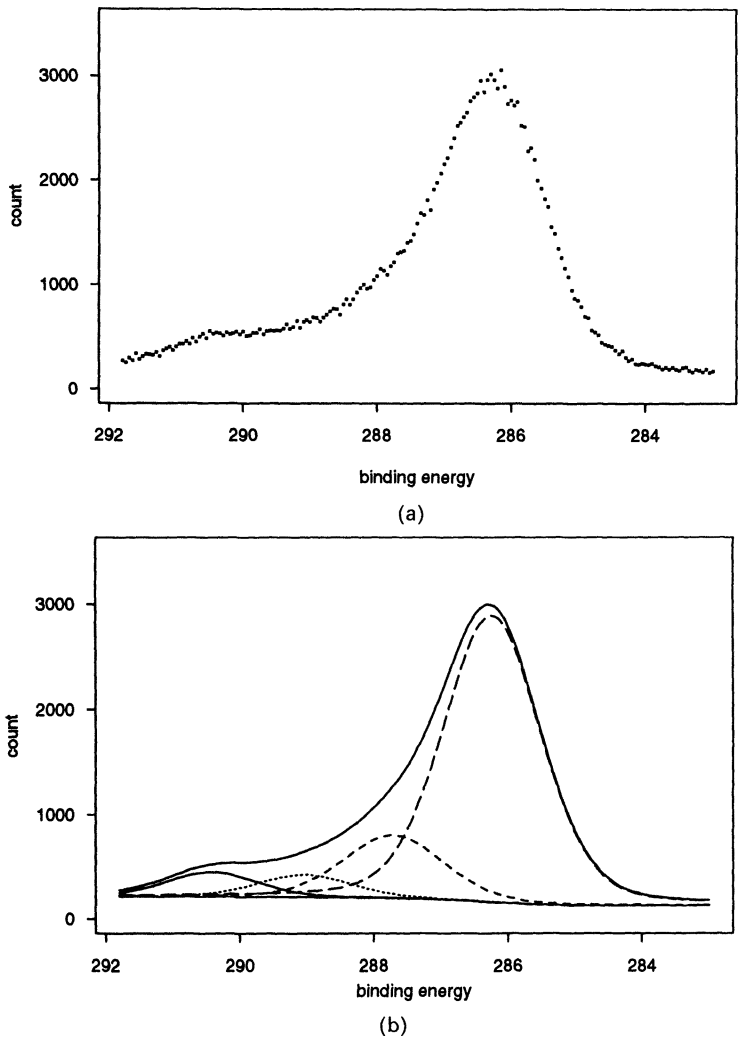

Fig. 1. 1s carbon peak for PPMMA: (a) raw data; (b) decomposed peak with background (commonly, in plots of ESCA spectra, the energy (horizonal) axis is a decreasing scale since higher binding energies correspond to lower observed energies in the electron analyser)

complex peak, which originates from the 1s carbon orbital in plasma-polymerized methylmethacrylate (PPMMA), a polymer used in microelectronics manufacturing, and its decomposition into a background and four single peaks. This decomposition is very good, but not unique.

Even for a known number of subpeaks, it can be difficult or impossible to obtain a unique weighted least squares fit. In our example, a large class of decompositions into four subpeaks provides fits of almost indistinguishable quality. Fig. 2 displays four such decompositions, and Fig. 3 shows that they produce nearly equivalent fits. The decomposition in Fig. 2(d) is the most realistic, having subpeaks with similar widths, whereas the decomposition in Fig. 2(a), corresponding to the smallest weighted sum of squares among the displayed fits, has subpeaks of greatly differing widths and does not lead to a sensible chemical interpretation.

If a unique weighted least squares fit can be found, the non-linearity of the model makes it difficult to assess the precision of the estimates of model parameters and derived quantities such as the ratios of peak areas.

Finally, there is also uncertainty about the methods for removing the background. Therefore the model may have to be augmented by another parametric or nonparametric

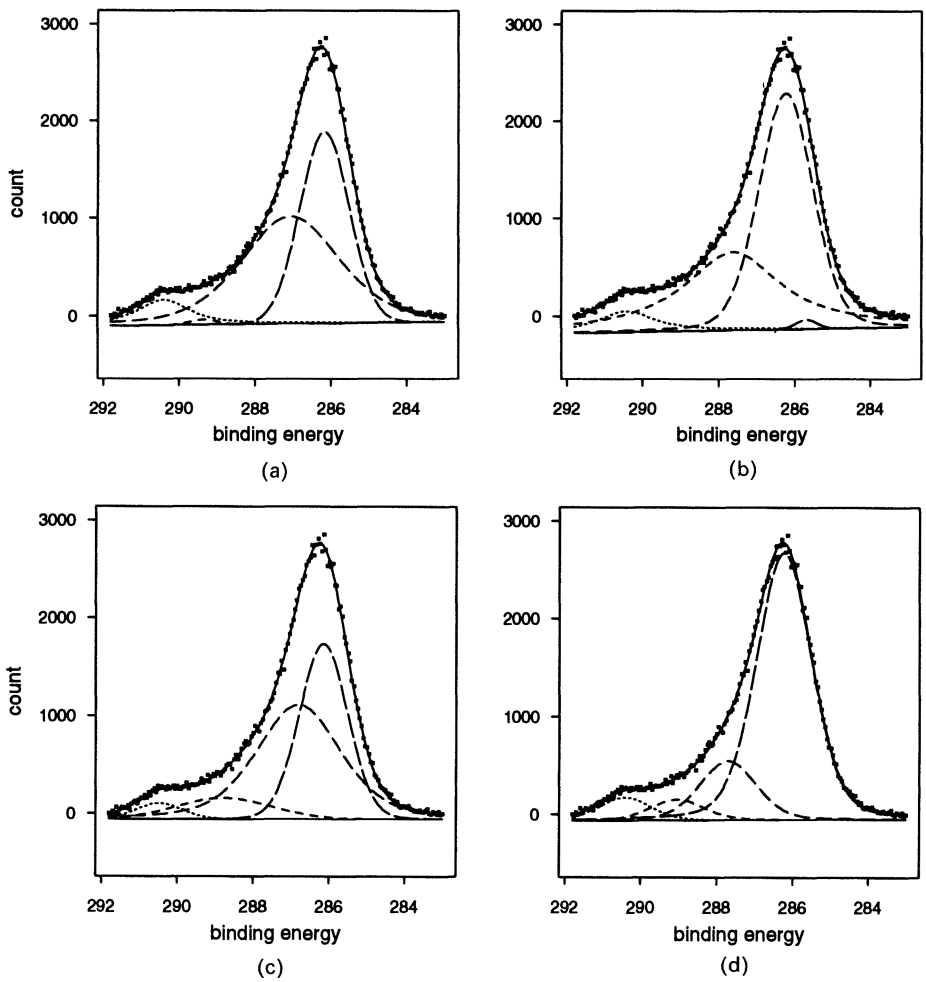

Fig. 2. Four different decompositions: ·····, ---------, – – – –, ————, subpeaks; ————, superposition and background; ■, data points

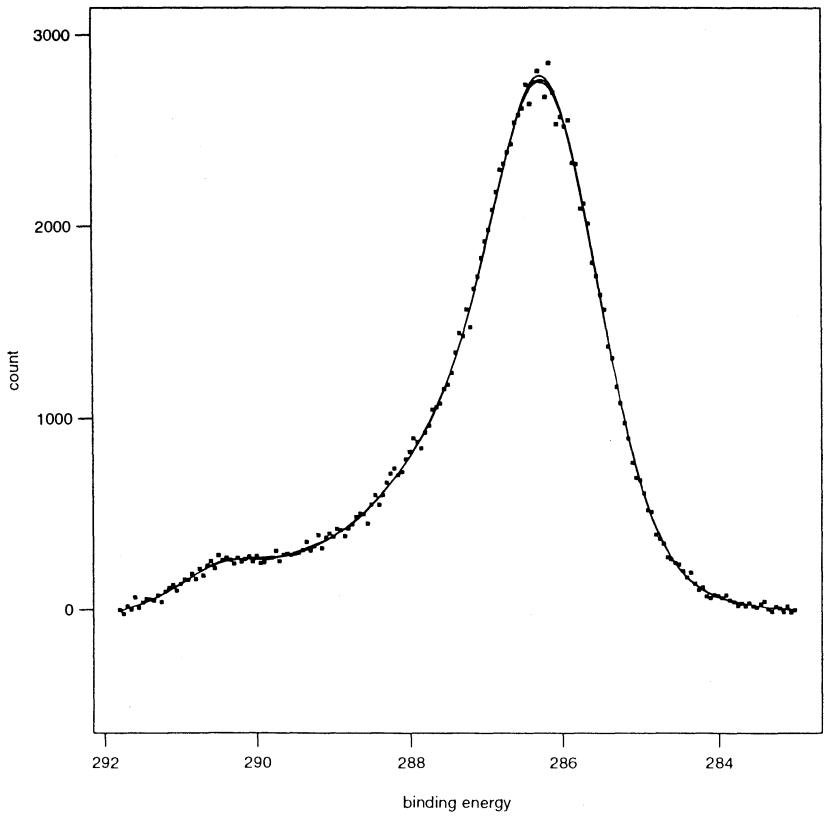

Fig. 3. Four fits superimposed (———) together with the data (■)

component. Several scenarios for removing the background may have to be studied and compared. In our analysis, we adjust the raw counts y_i^* by an integrated background estimate b_i (Powell and Seah, 1990). Furthermore, we augment the model by the linear adjustment term $i + s(x_i - x_a)$.

5. Prior information

Researchers working in ESCA know from their experience with standard samples that close peaks involving the same element have similar, but not necessarily identical, widths and mixture ratios. Moreover, the approximate peak positions for the spectra of certain model compounds are known from previous experiments and have been tabulated. If no unique fit can be achieved, the experienced researcher uses this auxiliary information informally to discriminate between different fits of equal quality. Here, we show that by using Bayesian techniques such auxiliary information can be formally incorporated into the estimation procedure. In doing so, we can not only arrive at plausible fits but may also be better able to assess the precision of the subsequent parameter estimates.

5.1. *Simple quadratic penalties*

As we have just mentioned, we usually have information about some of the parameters. Typically, much of this information concerns spacings of parameters, but not their precise values. For example, we know quite precisely that the peak widths and mixture ratios must be similar but we know much less about their values. These values depend partly

on properties of the measuring instrument, which can vary and which cannot all be controlled. Similarly, we know much about the spacing of the possible subpeaks, but not about their absolute positions, since the entire spectrum can be shifted owing to surface charging. Here, we discuss how simple quadratic penalties can be constructed which are mild with respect to global shifts and strict with respect to changes in the parameters relative to each other.

Suppose that $\boldsymbol{\theta}^{(0)}$ is an *a priori* estimate of the parameter vector $\boldsymbol{\theta}$ and suppose that the deviations of the components θ_i from $\theta_i^{(0)}$ can be represented by a standard deviation σ whereas the deviations of the differences $\tau_{ij} = \theta_i - \theta_j$ from the $\tau_{ij}^{(0)} = \theta_i^{(0)} - \theta_j^{(0)}$ can be represented by a standard deviation ω (usually smaller than σ). Then, the information about the precision can be summarized by a quadratic penalty

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})' \Sigma_P^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})$$

to the sum of squares with penalty matrix

$$\Sigma_P = \begin{pmatrix} \sigma^2 & t & \cdots & t \\ t & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & t \\ t & \cdots & t & \sigma^2 \end{pmatrix},$$

where $t = \sigma^2 - \omega^2/2$.

In Bayesian terms, this penalty corresponds to a multivariate normal prior distribution with mean vector $\boldsymbol{\theta}^{(0)}$ and covariance matrix Σ_P. Usually, we require that parameters for which a multivariate normal prior is to be chosen are unconstrained. This holds for the transformed parameters **a**, **b**, **g** and **r**.

5.2. *Constructing prior distribution for our example*

5.2.1. *Peak heights.* Little is known beforehand about the peak heights and therefore about the parameters a_j and we assume a flat prior for them.

5.2.2. *Peak locations.* For our polymer, we can expect subpeaks corresponding to C—C, C—O, C=O, O—C—O and O—C=O bonds, and therefore construct a model with four peaks (O—C—O and C=O create subpeaks at almost identical locations). Using the tabulated peak positions from Clark and Thomas (1976a, b) and our judgment of the precision of the absolute peak positions and their spacing, we can construct a quadratic penalty of the previously introduced type with $\mathbf{b}^{(0)} = (289.0, 287.8, 286.5, 285.0)'$, $\sigma_b = 2$ and $\omega_b = 1$. In our example, it is apparent that the whole spectrum is shifted by about 1.2 eV owing to surface charging, and we adjust by this amount.

5.2.3. *Peak widths and shapes.* In our measurements, 1s carbon peaks are commonly between 1 eV and 2 eV wide and can be approximated by Gaussian–Lorentzian mixtures of 60–95% Gaussian content. From previous experiments we derive an average peak width of 1.8 eV and an average mixture ratio of 88% Gaussian. Furthermore, we know that typically the peak widths and mixture ratios do not differ by more than 0.1 eV and 5% respectively. Therefore, we choose priors with $\mathbf{g}^{(0)} = (0.6, 0.6, 0.6, 0.6)'$, $\sigma_g = 0.5$ and $\omega_g = 0.1$ for the widths, and $\mathbf{r}^{(0)} = (2, 2, 2, 2)'$, $\sigma_r = 2$ and $\omega_r = 0.5$ for **r**, the vector of the transformed mixture ratios.

5.3. *Combined penalty*

Together, we obtain the combined penalty

$$P(\mathbf{a}, \mathbf{b}, \mathbf{g}, \mathbf{r}) = (\mathbf{b} - \mathbf{b}^{(0)})' \Sigma_b^{-1} (\mathbf{b} - \mathbf{b}^{(0)}) + (\mathbf{g} - \mathbf{g}^{(0)})' \Sigma_g^{-1} (\mathbf{g} - \mathbf{g}^{(0)}) + (\mathbf{r} - \mathbf{r}^{(0)})' \Sigma_r^{-1} (\mathbf{r} - \mathbf{r}^{(0)})$$

with

$$\mathbf{b}^{(0)} = \begin{pmatrix} 290.2 \\ 289.0 \\ 287.7 \\ 286.2 \end{pmatrix}, \qquad \Sigma_b = \begin{pmatrix} 4.0 & 3.5 & \cdots & 3.5 \\ 3.5 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 3.5 \\ 3.5 & \cdots & 3.5 & 4.0 \end{pmatrix},$$

$$\mathbf{g}^{(0)} = \begin{pmatrix} 0.6 \\ 0.6 \\ 0.6 \\ 0.6 \end{pmatrix}, \qquad \Sigma_g = \begin{pmatrix} 0.250 & 0.245 & \cdots & 0.245 \\ 0.245 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0.245 \\ 0.245 & \cdots & 0.245 & 0.250 \end{pmatrix},$$

and

$$\mathbf{r}^{(0)} = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 2 \end{pmatrix}, \qquad \Sigma_r = \begin{pmatrix} 4.000 & 3.875 & \cdots & 3.875 \\ 3.875 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 3.875 \\ 3.875 & \cdots & 3.875 & 4.000 \end{pmatrix}.$$

The combined penalty corresponds to independent multivariate normal priors for the parameters **b**, **g** and **r**, whereas a flat prior is chosen for **a**. The resulting posterior can be written as a non-linear least squares likelihood for an augmented model with pseudodata. Thus, the usual non-linear least squares algorithms can be used to find the posterior mode, if it exists.

6. Finding posterior mode and approximate standard errors

The posterior density, corresponding to the prior introduced previously, has a unique mode which can be found by non-linear least squares. Table 1 shows the resulting parameter estimates and their approximate standard errors. These standard errors were computed from the normal approximation to the posterior distribution at the mode. The corresponding fit is shown in Fig. 4. The fit is good and the residuals do not reveal systematic deviations. The first-order autocorrelation coefficient of the residuals is $r = -0.09$, which is not significantly different from 0.

TABLE 1
Parameter estimates and approximate standard errors

Parameter	*Value*	*Standard error*	*Parameter*	*Value*	*Standard error*
a_1	5.46	0.08	b_1	−290.42	0.06
a_2	5.37	0.18	b_2	−289.03	0.16
a_3	6.40	0.07	b_3	−287.70	0.09
a_4	7.91	0.01	b_4	−286.23	0.02
g_1	0.46	0.08	r_1	1.24	0.52
g_2	0.50	0.09	r_2	1.29	0.51
g_3	0.55	0.08	r_3	1.35	0.50
g_4	0.54	0.01	r_4	1.21	0.19
i	−62.90	18.77			
s	2.27	1.99			

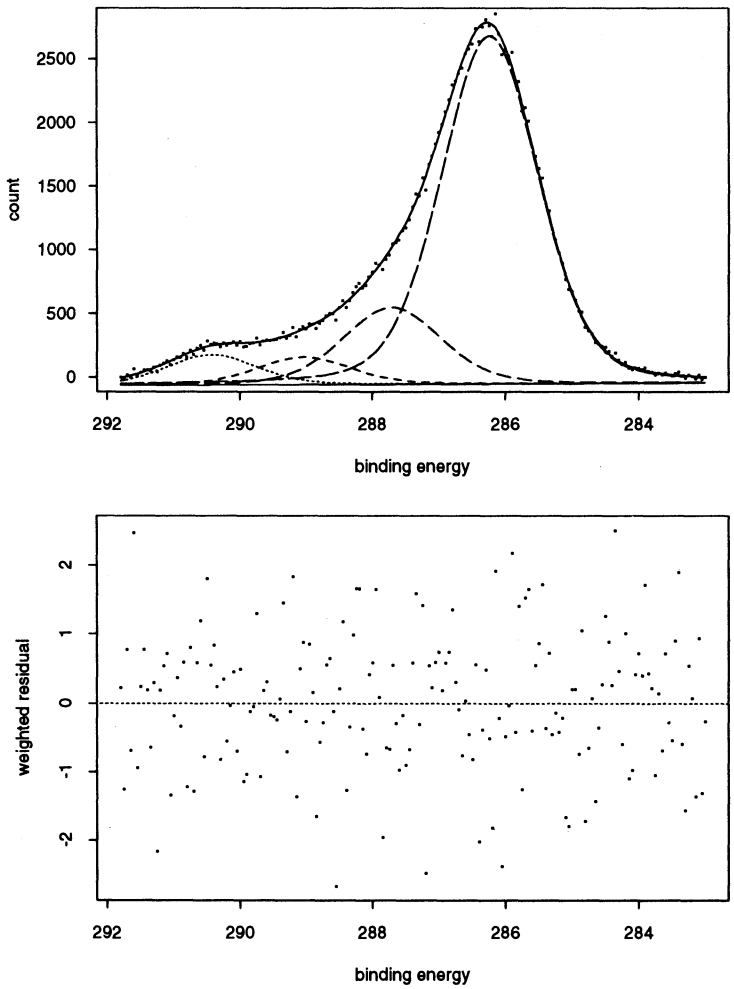

Fig. 4. Fit corresponding to the posterior mode and weighted residuals

7. Using posterior profiles for assessing quality of normal approximation to posterior

Profile methods (Bates and Watts, 1988; Ritter, 1992) are very general tools for summarizing a function F of several variables $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ in a neighbourhood of a maximum $\hat{\boldsymbol{\theta}}$. In our context this function is the posterior density. In the profiling operation each parameter component θ_i is varied in turn, and for each value $\theta_i = t$ the function F is maximized with respect to the other components. The resulting conditional maximizer $\tilde{\boldsymbol{\theta}}_i(t)$, taken as a function of t, is called the ith profile trace, and $\tilde{F}_i(t) = F\{\tilde{\boldsymbol{\theta}}_i(t)\}$ is called the ith profile value. Accordingly, we call the structure $\{\tilde{\boldsymbol{\theta}}_i(\,), \tilde{F}_i(\,)\}$ the ith profile. The profile value can be interpreted as the shadow of F when seen from the θ_i-axis. Posterior profiles can be used to investigate whether a normal approximation of the posterior is appropriate; in addition, more elaborate inference techniques can be based on posterior profiles.

However, the plain posterior profile $\{\tilde{\boldsymbol{\theta}}_i(\,), \tilde{\pi}_i(\,)\}$ is not always a good way of summarizing the posterior along the θ_i-axis, since it ignores that for some values of θ_i the conditional maximum $\tilde{\boldsymbol{\theta}}_i(\theta_i)$ may be narrower than for others. If the conditional maximum is wide, many other settings of the remaining parameters correspond to fits of almost equal quality. If the conditional maximum is narrow, only few other settings are as good as $\tilde{\boldsymbol{\theta}}_i(\theta_i)$. The *modified*

profile adjusts for such imbalances by multiplying $\tilde{\pi}_i(\,)$ by a width adjustment factor of the form $|\tilde{\mathbf{R}}_{-i}(\theta_i)|^{-1/2}$ where

$$\tilde{\mathbf{R}}_{-i}(t) = \frac{\partial^2}{\partial \boldsymbol{\theta}_{-i} \partial \boldsymbol{\theta}'_{-i}} \log \pi(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}_i(t)},$$

and $\boldsymbol{\theta}_{-i}$ is the vector $\boldsymbol{\theta}$ without its ith component. The determinant of the Hessian matrix $\tilde{\mathbf{R}}_{-i}(\,)$ is a measure of the local width of the conditional maximum $\tilde{\boldsymbol{\theta}}_i(\,)$; it is small when the maximum is wide. The modified profile values are proportional to the Laplacian approximations to the componentwise marginals (Leonard, 1982; Tierney *et al.*, 1989; Leonard *et al.*, 1989).

In our example, the modified posterior profiles can be computed for all parameters over large intervals. Fig. 5 shows the resulting modified profile values (full curves) contrasted with

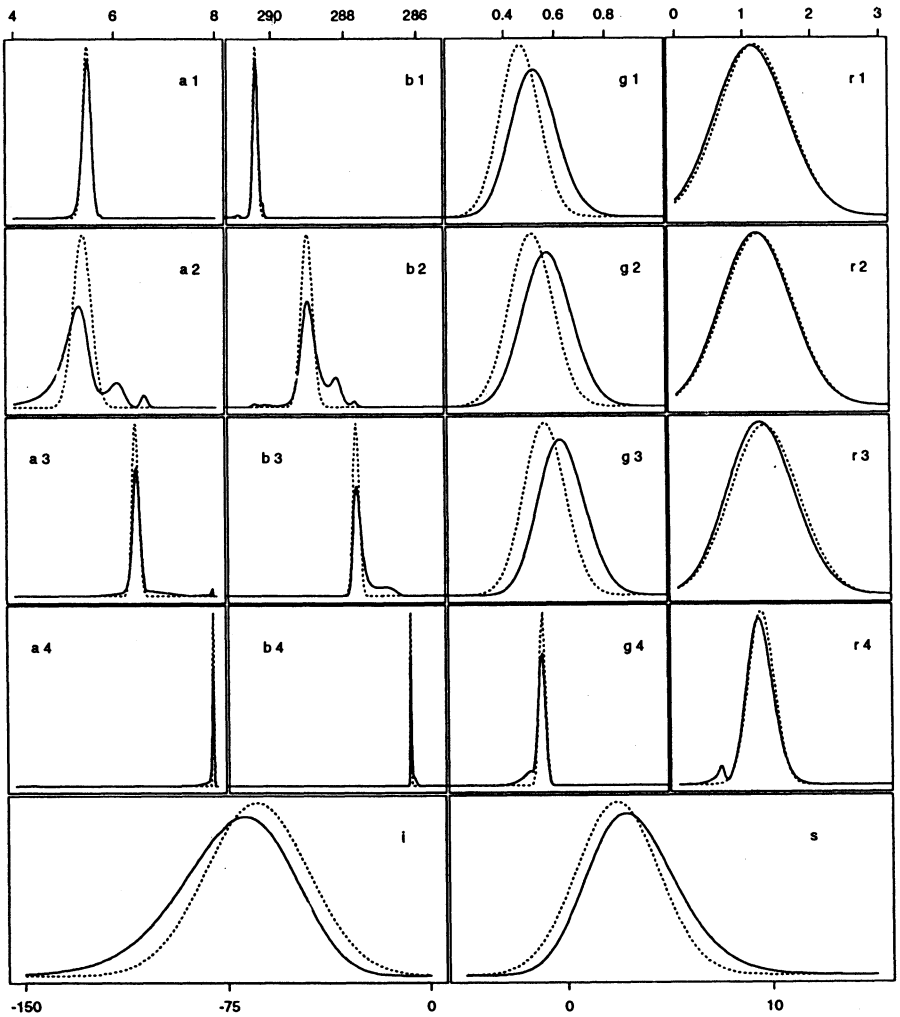

Fig. 5. Modified profiles (———) and normal approximations (·····): all curves have been normalized to integrate to 1 over the intervals displayed

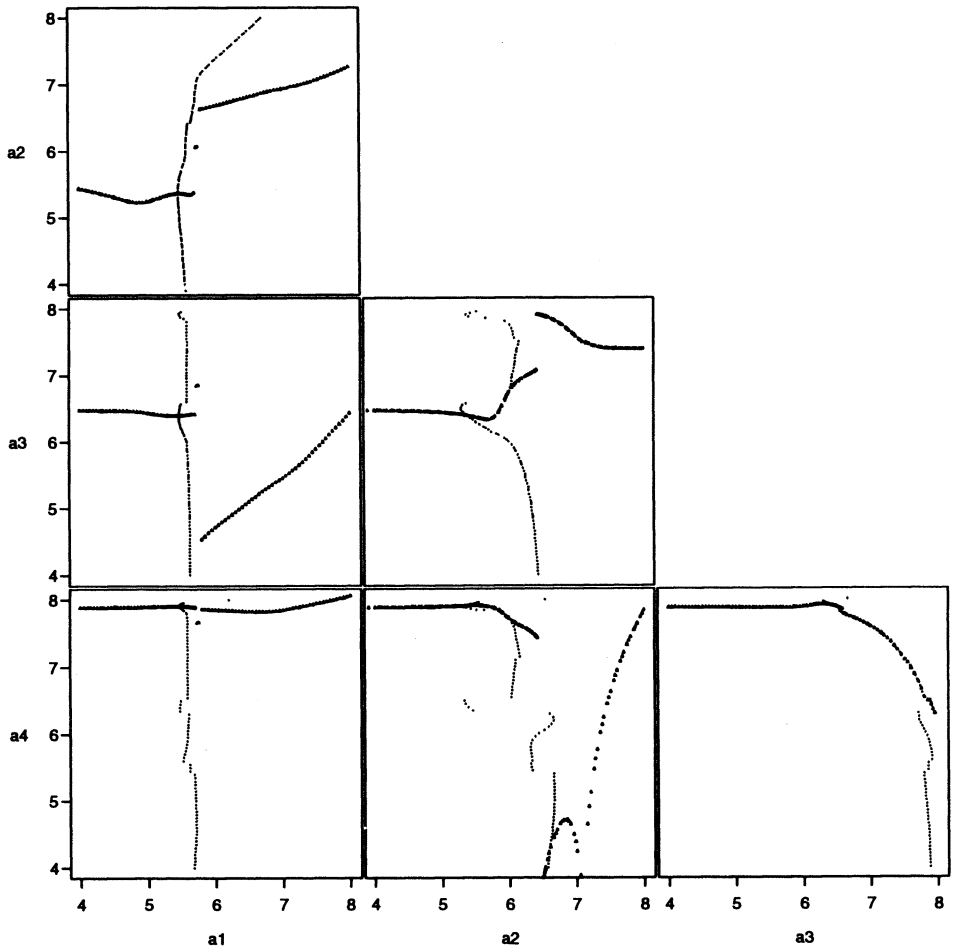

Fig. 6. Pairwise projections of the profile traces of the log-height parameters a_j

the marginals of the normal approximation corresponding to the posterior at the posterior mode (dotted curves).

Fig. 6 shows the pairwise plots of the profile traces of the log-height parameters a_1, a_2, a_3 and a_4. If the posterior is approximately multivariate normal, the posterior profile traces are nearly straight lines.

From Figs 5 and 6, we can conclude that the modified profile values are not well approximated by the normal approximations. Moreover, the profile traces corresponding to the height parameters have sudden bends and even jumps. The plots provide a clear warning that inference about linear or non-linear functionals of the parameters derived by the delta method from the approximate standard errors and correlations may be strongly misleading.

8. Using an interesting posterior profile to construct scenarios for decompositions

As is evident from the modified profile of the parameter a_2, there are three distinct probable configurations: the first at about $a_2 = 5.3$, the second at about 6.0 and the third at about 6.6. Fig. 7 shows the corresponding fits. Fig. 7(a), for $a_2 = 5.3$, shows a decomposition into four

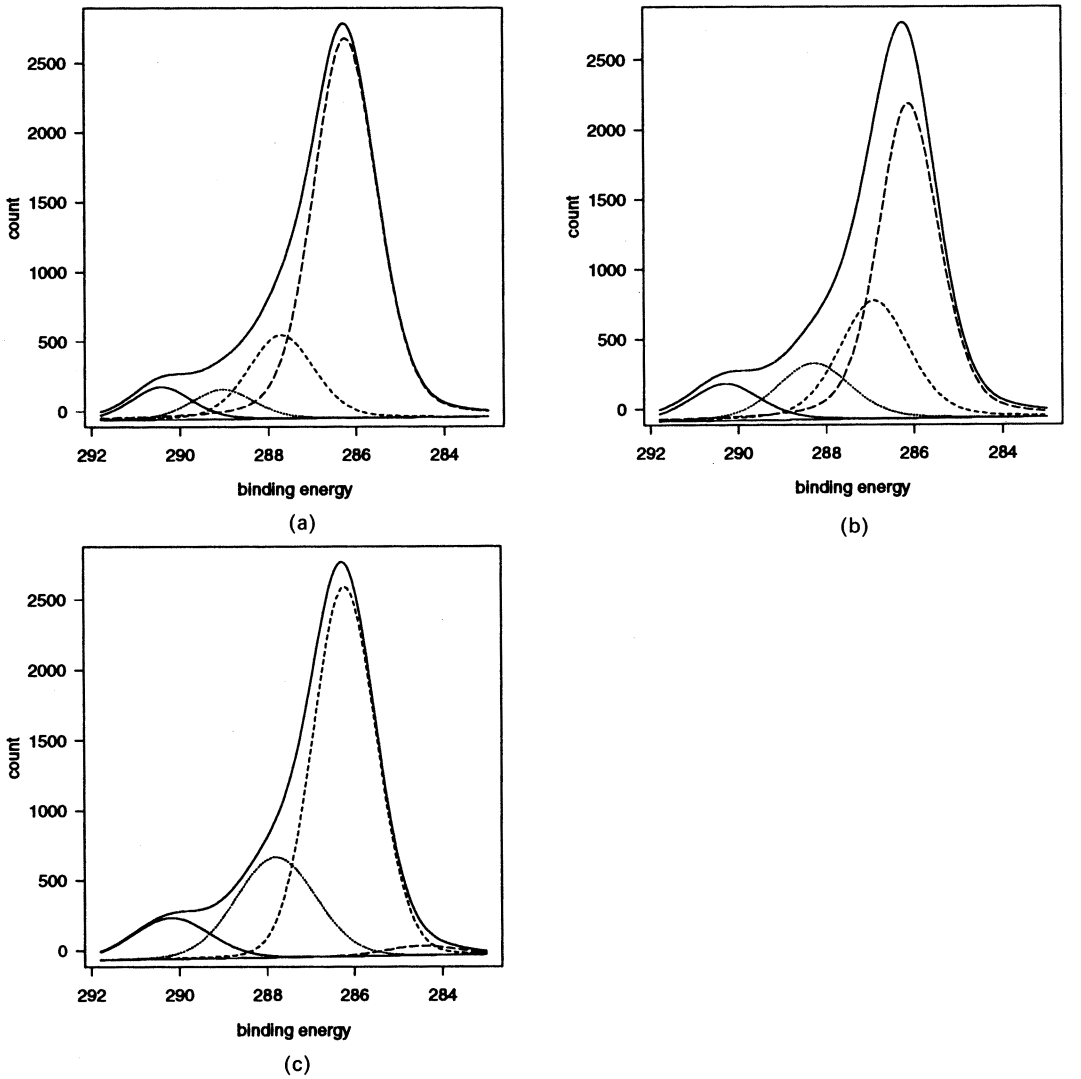

Fig. 7. Three fits corresponding to the three modes in the marginal distribution of a_2

distinct subpeaks which is consistent with our auxiliary knowledge. Fig. 7(b), for $a_2 = 6.0$, shows a decomposition into three subpeaks where the right (carbon–carbon) subpeak is split. Fig. 7(c), for $a_2 = 6.6$, shows a decomposition into three subpeaks with a very small fourth peak in the right-hand tail of the spectrum.

Since there are two minor modes corresponding to three-peak fits and a dominant mode corresponding to a four-peak fit, we conclude that the spectrum together with our auxiliary knowledge provide strong but not conclusive evidence of four peaks and therefore of the existence of the O—C—O or C=O bonds.

9. Using Markov chains to deepen analysis

9.1. *General comments*

The posterior can also be investigated by sampling parameter vectors $\{\boldsymbol{\theta}^{(i)}\}_{i=1}^{N}$ from it and studying the corresponding fits. If a large number of such parameter vectors has been drawn,

pairwise scatterplots of the $\theta^{(k)}$ can be used to explore the important part of the parameter space containing fits which are nearly equivalent to the optimal fit. Such displays are convenient tools for detecting unusual features of the posterior such as multiple modes. However, it is usually difficult to sample from the posterior directly, unless it has a simple form. Here, we use two methods that were specifically designed for complicated posteriors: the griddy Gibbs sampler (Ritter and Tanner, 1992) and the griddy Hastings algorithm (Ritter, 1992), also known as the sliding lattice approach (Tierney, 1991). Both are members of the large family of algorithms for creating ergodic Markov chains with desired stationary distributions. For additional details on such algorithms, we refer to Metropolis *et al.* (1953), Hastings (1970), Geman and Geman (1984), Tanner and Wong (1987), Gelfand and Smith (1990) and Müller (1991).

9.2. *Preparing our example for Markov chain exploration*

To use Markov chain sampling techniques in our example, we need to ensure that we are working with a proper posterior. In our example we do this by restricting the parameter space to a finite domain containing all reasonable parameter vectors. For the log-height parameters, we choose intervals from 2 to 8.5. Peaks of log-height less than 2 are invisible in the noise, whereas a peak of log-height 8.5 is clearly higher than any part of the observed spectrum. For the locations we impose the intervals [284, 291.5]. The first value is clearly below the carbon–carbon peak locations, and the second value is clearly higher than the expected position of the O—C=O peak.

For log-peak-widths and for the logistic transforms of the mixture ratios we choose the intervals $[-0.5, 1.5]$ and $[0, 8]$ respectively and for the intercept and the slope we consider the intervals $[-200, 100]$ and $[-20, 20]$ as sufficiently large.

9.3. *Conducting Markov chain analysis of posterior*

In our example, we construct 100 parallel chains. This approach is convenient but not very efficient; more efficient implementations use only one long chain. We choose multiple chains since we want to monitor the progress of the sample from iteration to iteration easily.

We start with 100 parameter vectors uniformly distributed over the parameter domain and use the griddy Gibbs sampler with 15 grid points until for all parameters the plots of the sample quantiles against the number of iterations indicate that the sample has stabilized. In our case this occurs after about 200 iterations. We continue the sampling by using the griddy Hastings algorithm with five grid points. Since the posterior is quite complicated, convergence is very slow, and even after 5000 iterations the width parameters g_1 and g_2 have not completely stabilized.

In Fig. 8 the traces of the 0th, 25th, 50th, 75th and 100th percentiles of the location parameters indicate that the first peak quickly assumes a position at about 290 eV whereas the other three peaks initially share the position of the main peak of the spectrum at about 286.5 eV. Later, the second and the third peak begin to move away from the fourth peak. Eventually, the second peak assumes positions between 289 eV and 288 eV, the third peak between 288 eV and 286.5 eV, and the fourth peak locks in at about 286.2 eV.

To compare the results from the Markov chain run with the results obtained by profiling we generate estimates of marginal densities of all parameters from the Markov chain sample after the 5000th iteration. Fig. 9 shows the modified profiles and the marginal histograms from the Markov chain sample of size 100 (see Fig. 11, later, for a comparison of the modified profiles (full curves) and smooth density estimates of the Markov chain sample (dotted curves) obtained by the method of Kooperberg and Stone (1991)).

We conclude that the overall agreement is quite good. Possible reasons for the remaining differences in some of the parameters are that either the Laplacian approximation is not good for these parameters or that the Markov chain had been stopped prematurely.

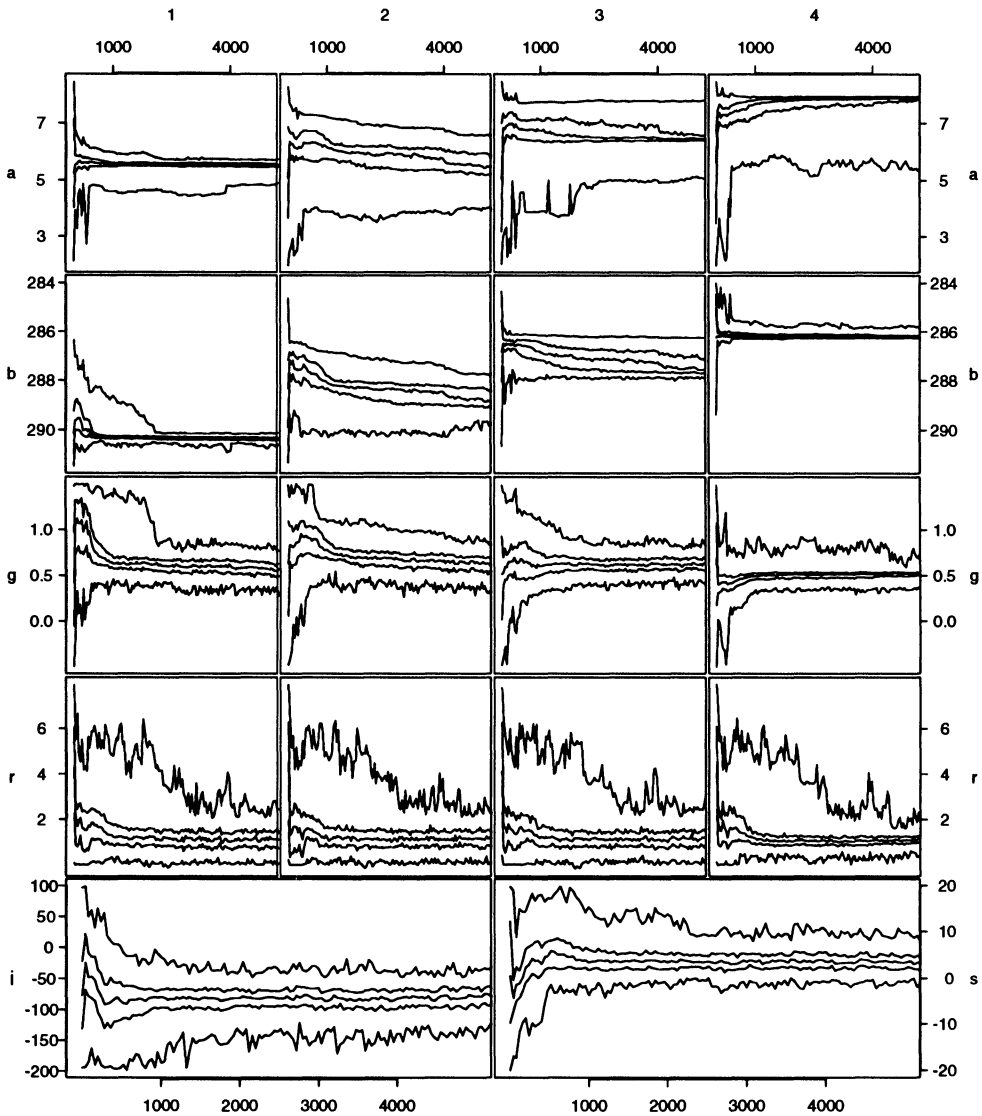

Fig. 8. Quartiles (0th, 25th, 50th, 75th, 100th percentiles) *versus* number of iterations, for all parameters (first 200 iterations, griddy Gribbs sampler; remaining iterations, griddy Hastings algorithm)

Fig. 10 shows the pairwise plots of the Markov chain sample of the height parameters at the last iteration. This plot is in agreement with the profile traces shown in Fig. 6.

9.4. *Assessing effect of ignoring prior information*

Using the Markov chain approach, we can also investigate what happens when the prior information is ignored. Since without a prior no unique fit can be obtained, the profile methods cannot be used here. Fig. 11 shows smooth density estimates derived from a Markov chain sample of the posterior (dotted curves) and smooth density estimates derived from a Markov chain sample of the plain likelihood (broken curves). The Markov chain sample of the plain likelihood (restricted to the same finite domain as the posterior and therefore made integrable) was generated similarly to the Markov chain sample of the posterior by starting with 100

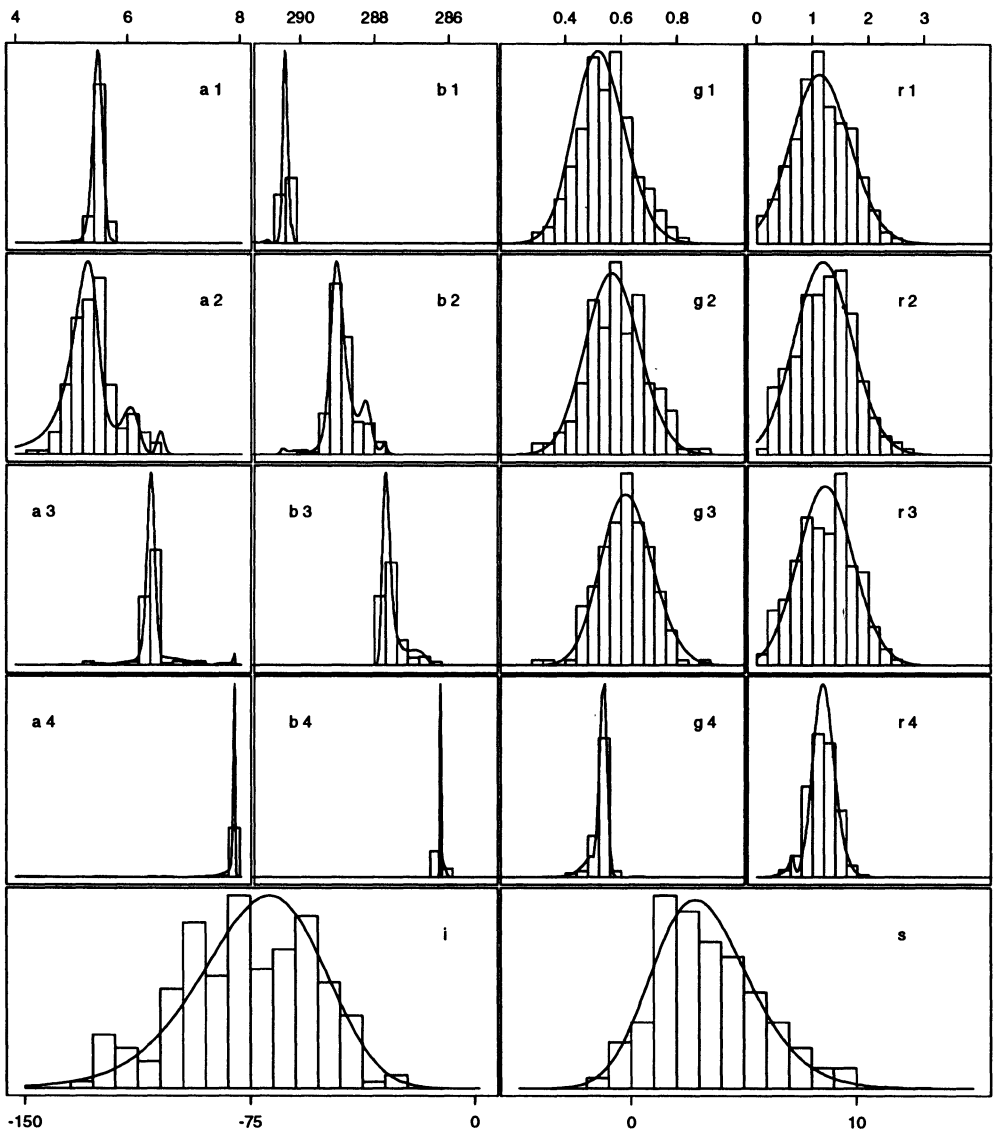

Fig. 9. Approximate marginals for all parameters: smooth curves, Laplace's approximation; histogram, Markov chain sample (smooth representation of the marginals of the Markov chain sample can be found in Fig. 11)

uniformly distributed parameter vectors, and then running the griddy Gibbs sampler for several hundred iterations followed by 5000 iterations of the griddy Hastings algorithm. Here also, although most of the parameters had stabilized, full convergence could not be achieved. The smooth density estimates were computed by using log-spline-density estimation (Kooperberg and Stone, 1991).

The striking features of Fig. 11 are the complete loss of precision regarding the mixture parameters and a preference for mixtures with a widened second peak, where the second peak is moved to the right, and the third and fourth peaks are either forming a split peak, or the third peak takes over the position of the fourth peak, pushing it into the right-hand tail of the spectrum. Such decompositions do not correspond to sensible chemical interpretations.

We therefore conclude that the observed spectrum contains very little information about

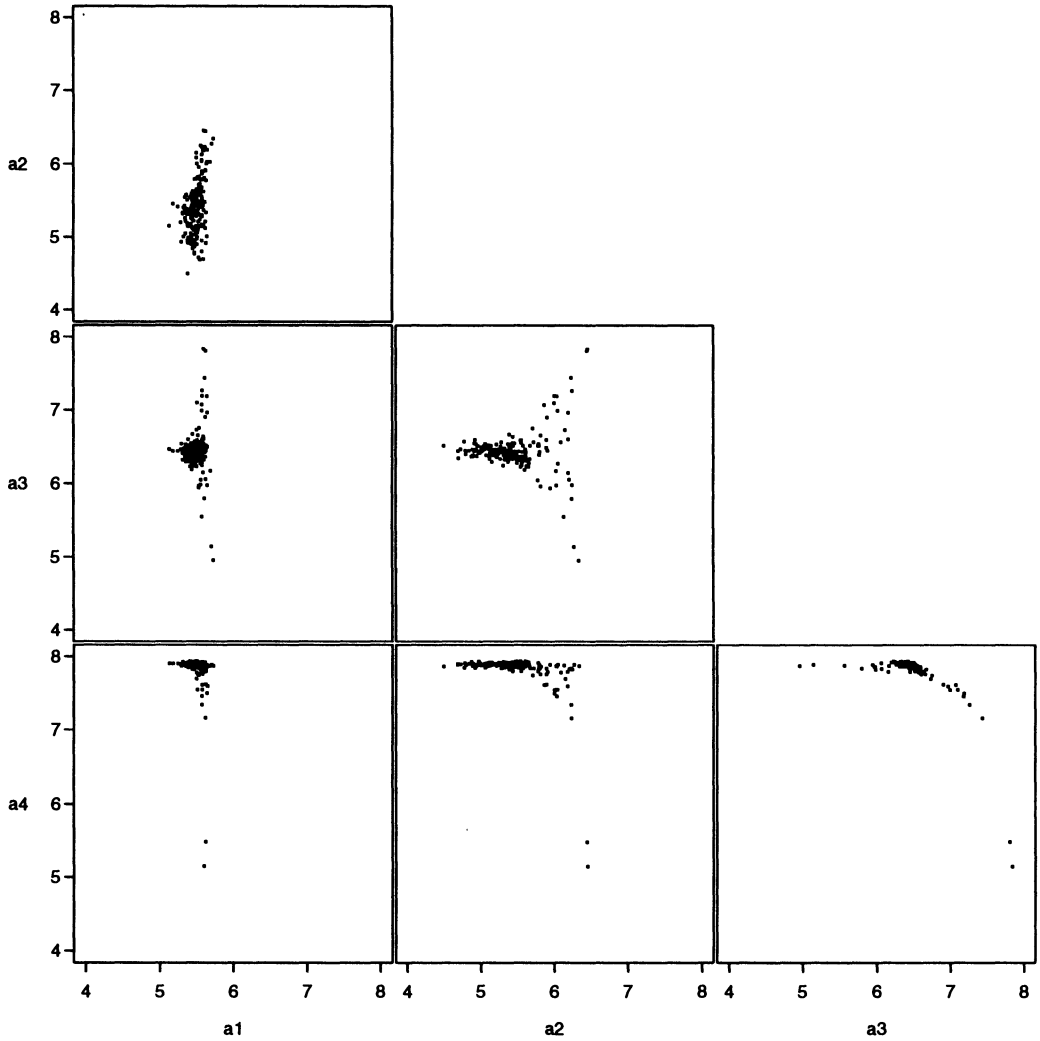

Fig. 10. Pairwise scatterplots of the log-height parameters a_j of the Markov chain sample after 5000 iterations

the mixture ratios and by itself does not suggest a sensible peak structure. Only by introducing all available auxiliary information can the decomposition produce useful results.

10. Conclusions

Using modern statistical techniques, we can investigate decompositions of ESCA spectra in great detail. To obtain safe and sensible interpretations, it is essential to incorporate as much expert information as possible into the fitting process. In well-conditioned cases, simple summaries consisting of parameter estimates and standard errors are sufficient. If there is doubt, the profile methods can be used to assess the adequacy of these summaries. If there is discussion about the prior information in the fitting, Markov chain simulations can be used to investigate what happens if this information is weakened. Here, it is not necessary that a unique fit still exists. In this context, even if the Markov chain algorithms do not converge completely, the intermediate results can provide useful information for the researcher.

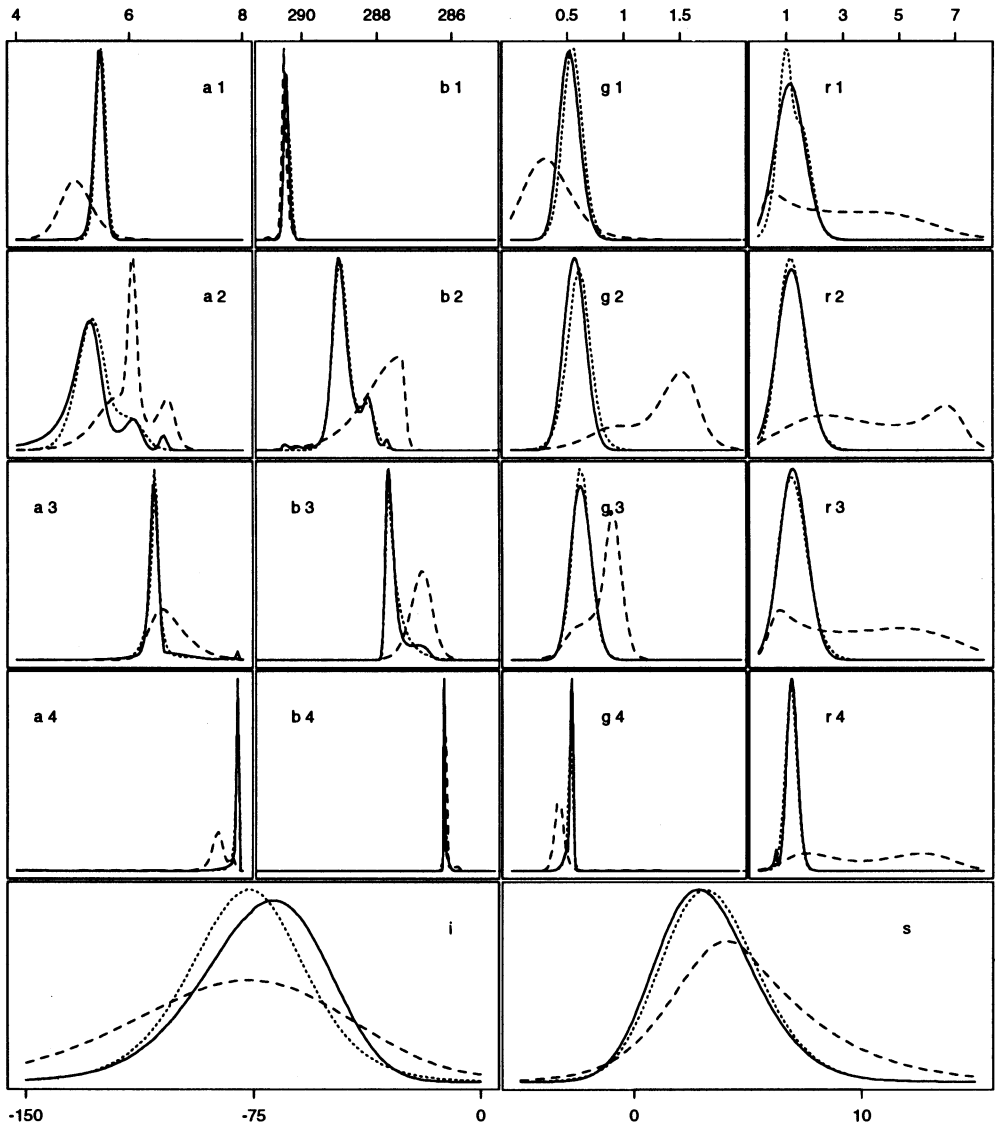

Fig. 11. Effect of ignoring the prior information: modified profiles of the penalized problem (———), smooth density estimates of a Markov chain sample of the posterior (·····) and smooth density estimates of a Markov chain sample of the plain likelihood (---------)

A major drawback of using Markov chain methods to study ESCA decompositions is the large amount of computing involved. In our example 5000 iterations in the Markov chain were conducted, corresponding to roughly 100 million evaluations of the posterior. Profile methods are much less intensive of computer time, but they only work if a unique fit can be achieved, and if the profiles can be computed sufficiently far from the posterior mode.

Acknowledgements

The author would like to thank Søren Bisgaard, Milan Buncick and Jeff Tobin from the Engineering Research Center of Plasma Aided Manufacturing, and Ngoc Tran from the

Material Sciences Department at the University of Wisconsin–Madison for their help and encouragement. The author would also like to thank Douglas Bates, Martin Tanner and Tom Leonard for introducing him to some of the new statistical methods used in this research. This work was supported by the National Science Foundation under grant ECD-8721545 and by the European Community under grant ERBCHBGCT920167.

References

Baruya, A. and Maddams, W. F. (1978) An examination of the uniqueness of Gausian and Lorentzian profiles. *Appl. Spectrosc.*, **32**, 563–566.

Bates, D. M. and Watts, D. G. (1988) *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Clark, D. T. and Thomas, H. R. (1976a) Applications of ESCA to polymer chemistry: X, Core and valence energy levels of a series of polyacrylates. *J. Polym. Sci. Polym. Chem.*, **14**, 1671–1700.

——— (1976b) Applications of ESCA to polymer chemistry: XI, Core and valence energy levels of a series of polymethacrylates. *J. Polym. Sci. Polym. Chem.*, **14**, 1701–1713.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, **6**, 721–741.

Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Kooperberg, C. and Stone, C. J. (1991) Logspline density estimation for censored data. *Technical Report 226*. Department of Statistics, University of Washington, Seattle.

Leonard, T. (1982) Comments on "A simple predictive density function". *J. Am. Statist. Ass.*, **77**, 657–658.

Leonard, T., Hsu, J. S. J. and Tsui, K. W. (1989) Bayesian marginal inference. *J. Am. Statist. Ass.*, **84**, 1051–1058.

Maddams, W. (1980) The scope and limitations of curve fitting. *Appl. Spectrosc.*, **34**, 245–267.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Müller, P. (1991) A generic approach to posterior integration and Gibbs sampling. *Technical Report 91-09*. Department of Statistics, Purdue University, West Lafayette.

Perram, J. W. (1968) Interpretation of spectra. *J. Chem. Phys.*, **49**, 4245.

Powell, C. J. and Seah, M. P. (1990) Precision, accuracy, and uncertainty in quantitative surface analyses by Auger-electron spectroscopy and x-ray photoelectron spectroscopy. *J. Vac. Sci. Technol.* A, **8**, 735–763.

Ritter, C. (1992) Modern inference for nonlinear least squares regression. *PhD Thesis*. Department of Statistics, University of Wisconsin, Madison.

Ritter, C. and Tanner, M. (1992) Facilitating the Gibbs sampler: the Gibbs stopper and the Griddy Gibbs sampler. *J. Am. Statist. Ass.*, **87**, 861–868.

Tanner, M. and Wong, W. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Statist. Ass.*, **82**, 528–541.

Tierney, L. (1991) Markov chains for exploring posterior distributions. *Technical Report 560*. School of Statistics, University of Minnesota, Minneapolis.

Tierney, L., Kass, R. and Kadane, J. B. (1989) Approximate marginal densities of nonlinear functions. *Biometrika*, **76**, 425–433.

Vanderginste, V. G. M. and De Galan, L. (1978) Critical evaluation of curve fitting in infrared spectroscopy. *Ann. Chem.*, **47**, 2124–2132.

Wertheim, G. K. and DiCenzo, S. B. (1986) Least squares analysis of photoemission data. *J. Elect. Spectrosc. Reltd Phen.*, **40**, 57–67.