Estimating SCB-MW distributions from partial data

Report IWDE 01 - 04

Estimating SCB-MW distributions from partial data

J. Molenaar

December 2001

Instituut
Wiskundige Dienstverlening
Eindhoven

Technische Universiteit Eindhoven

Estimating SCB-MW distributions from partial data

J. Molenaar
September 2001

Background and abstract. *Since 1999 DOW and Eindhoven University of Technology (EUT) cooperate in a project entitled "Property prediction from molecular structure". The information about the molecular structure of a polymer blend is contained in the SCBD-MWD (Short Chain Branching-Molecular Weight) distribution. The experimental determination of this distribution via cross fractionation is an expensive and time consuming task. During the meeting of June, 2001, the idea was proposed to estimate this distribution from integrated partial information, which is much simpler obtained. The present report deals with the mathematical aspects of the idea to reconstruct the full distribution from partial data. It is shown that the method can be very successful if the distribution to be reconstructed consists of a limited number of smooth peaks.*

Introduction

In the project 'Property prediction from molecular structure' – a cooperation between DOW and EUT – one tries to predict the solid state properties of polymer blends from information on the molecular structure of the blends. As determining molecular parameters the Molecular Weight Distribution (MWD) and the Short Chain Branching Distribution (SCBD) are used. The MWD indicates the fraction of molecules with a given molecular weight. The structures of two chains with the same molecular weight may differ considerably, since the number and location of the branches may vary. In the context of this project we only deal with chains consisting of a long backbone with short branches. The SCBD gives the fraction of the molecules with a given branching content. It can be measured by the Crystaf device. We shall not explain this method here in detail. The polymer is fractionated at varying temperatures and to each temperature a definite branching content corresponds. The Crystaf device yields per temperature the total molecular weight of the fraction of molecules with a specific branching content. The SCBD is therefore a function of the temperature. Both the MWD and the SCBD contain partial information. The full information on the molecular structure of the blend is contained in the SCB-MW distribution which gives the fraction of molecules with given branching content and given molecular weight. This full distribution can be measured via fractionating according to branching followed by molecular weight analysis or vice-versa. However, this experimental procedure is expensive and time consuming. Still, the full SCB-MW distribution is needed, since some macro-properties of the blend essentially depend on the full distribution and not only on MWD or SCBD. An example of such a property is Tie-Chain. The experimental procedures to obtain MWD and SCBD separately are standard and relatively cheap nowadays. This raises the question whether it is possible to reconstruct the full distribution from the partial ones. From an experimental and financial point of view such a reconstruction method is highly attractive. In the present report we deal with the mathematical and practical aspects of the reconstruction and show that this approach is very promising.

Problem formulation

For the sake of conciseness we shall denote the full SCB-MW distribution by $F(m,T)$ with m denoting molecular weight and T temperature. The distribution F thus depends on two independent variables and can be plotted as a surface in the 3-dimensionale place. A characteristic example is given in Fig. 1. For F it holds that $F(m,T) \geq 0$ for all (m,T). The distribution typically consists of a limited number of smooth peaks. Partial information about F is contained in the functions

$$G(m) \equiv \int F(m,T)\, dT \ ,$$

$$H(T) \equiv \int F(m,T)\ dm\ . \tag{1}$$

where m and T range over the relevant intervals. From the definitions it is clear that G contains information on F integrated along lines parallel to the T-axis, and H contains information on F integrated along lines parallel to the m-axis.

In terms of F, G and H the central question is:

Is it possible to reconstruct F from given G and H data?

We first discuss a naive approach that certainly fails, but provides some insight in the problem. After that it is shown how the question can be answered positively.

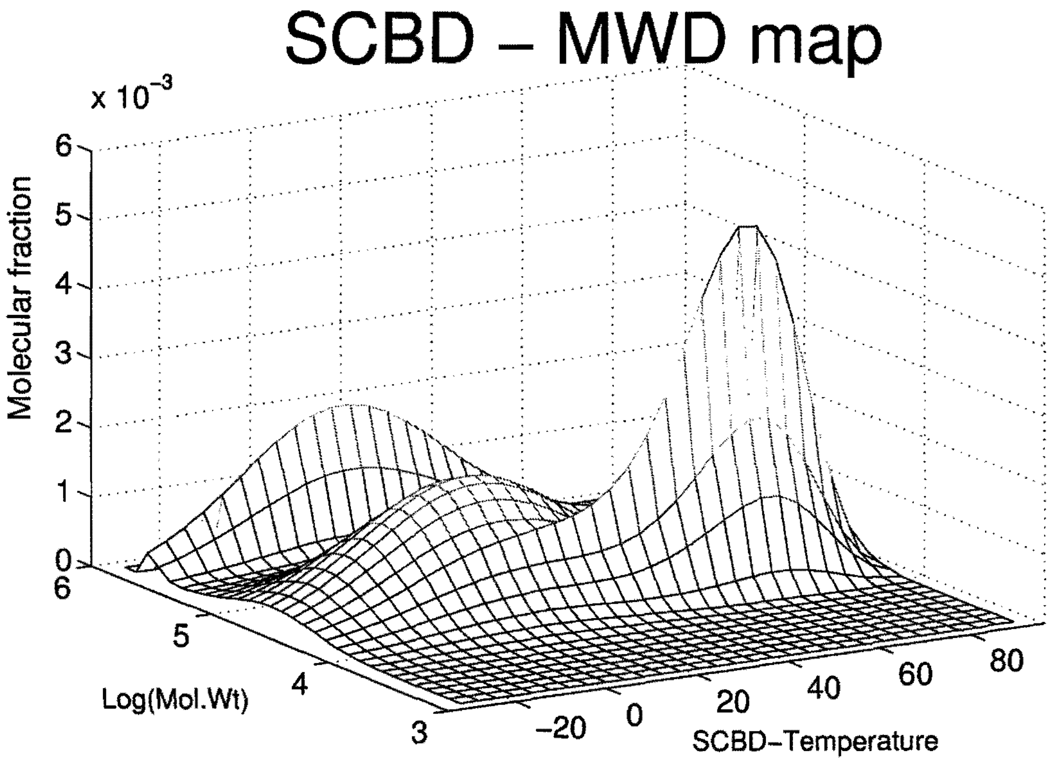

Fig. 1. Example of an SCB-MW distribution with 3 peaks.

Naive approach

In practice m and T vary over discrete grid points m_i and T_j. Let us for convenience assume that the number of grid points is the same along both axes and that both grids are uniform. The corresponding grid values of F are denoted as

$$F_{ij} \equiv F(m_i, T_j), \quad 1 \leq i, j \leq N \, . \tag{2}$$

Let us approximate the integrals in (1) by Riemann sums. Then, the values of G and H are estimated by

$$G_i \equiv G(m_i) = h_1 \sum_{j=1}^{N-1} F(m_i, T_j) = h_1 \sum_{j=1}^{N-1} F_{ij}, \quad 1 \leq i \leq N \, . \tag{3}$$

$$H_j \equiv H(T_j) = h_2 \sum_{i=1}^{N-1} F(m_i, T_j) = h_2 \sum_{i=1}^{N-1} F_{ij}, \quad 1 \leq j \leq N \, . \tag{4}$$

Here, h_1 and h_2 are the step sizes along the m and the F-axis, respectively. We can formulate the central question as follows. Given N values of both G and H, so in total $2N$ data points, can the grid values F_{ij} be estimated? The number of unknowns F_{ij} is N^2. The number of (linear) equations (3) and (4) relating the unknowns to the data is $2N$. Since $N^2 > 2N$ for $N > 2$, this system of equations is underdetermined and has not a unique solution. This crude approach shows that in principle too little information is available. This shortage of data can be compensated by adding extra information, as shown in the following.

Reconstruction of one peak

Let us first deal with the situation that F consists of only one smooth peak. This situation will usually reveal itself in the data by G and H both showing one smooth peak. A crucial step is to make assumptions about the form of the peak of F. We assume that this peak is Gaussian shaped and can be represented by a distribution of the form

$$P(m, T) = \alpha \exp[-\beta(m - \overline{m})^2 - \gamma(T - \overline{T})^2] \tag{5}$$

with

α: height of the peak
β: peak width of the cross-section through the top and parallel to the T-axis
γ: peak width of the cross-section through the top and parallel to the m-axis
$\overline{m}$: m-coordinate of the top
$\overline{T}$: T-coordinate of the top.

We remark that expression (5) is not the most general representation of a Gaussian peak in two variables. For the present purpose of investigating the general principles of reconstruction inclusion of extra terms (e.g., the cross term $(m - \overline{m})(T - \overline{T})$) is not yet relevant. Such a refinement can be incorporated if the regression models used in the project require it.

According to (5), a peak is characterized by 5 parameters. To reconstruct the peak, these 5 parameters must be estimated from the partial data. Substituting (5) into (1) (with $F \equiv P$) we find that

$$
\begin{aligned}
G(m) &= \int_{-\infty}^{+\infty} P(m,T)\, dT \\
&= \alpha e^{-\beta(m-\overline{m})^2} \int_{-\infty}^{\infty} e^{-\gamma(T-\overline{T})^2}\, dT \\
&= \alpha\sqrt{\tfrac{\pi}{\gamma}}\; e^{-\beta(m-\overline{m})^2}\ .
\end{aligned} \tag{6}
$$

Similarly, for $H(T)$ we obtain

$$
H(T) = \alpha\sqrt{\tfrac{\pi}{\beta}}\; e^{-\gamma(T-\overline{T})^2}\ . \tag{7}
$$

So, $G(m)$ contains the 4 parameters α, β, γ, and $\overline{m}$, and $H(T)$ the 4 parameters $\alpha, \beta, \gamma, \overline{T}$. This implies that not all parameters can be estimated from only G data or only F data. Since the number of parameters is 5, we need at least 5 data points. These could be $4G$ and $1H$ data points, or $3G$ and $2H$ data points, or $2G$ and $3H$ data points, or $1G$ and $4F$ data points. In practice one has much more data and the estimation could be performed in the least squares sense. For details see the Appendix.

We conclude that reconstruction of one Gaussian peak from partial data is possible and requires in principle not many data points. The estimation procedure involves solving a set of equations. In the Appendix it is shown that this procedure is very simple and fast.

Reconstruction of two peaks

Above we have shown that reconstruction of one peak is possible if one introduces a specific representation of the peak. So, the form of the peak is assumed to belong to a restricted class. Since the details of the SCB-MW distribution are probably not very relevant for the prediction of polymer blend properties, the choice of this class is not crucial. Only a few characteristics of the peaks determine blend properties like Elastic Modulus and Tie Chain. In view of these considerations we represent a distribution F with two peaks as the sum of two Gaussian peaks:

$$
F(m,T) = P_1(m,T) + P_2(m,T) \tag{8}
$$

with P_1 and P_2 of the form (5). So, P_1 contains the 5 parameters $\alpha_1, \beta_1, \gamma_1, \overline{m}_1, \overline{T}_1$ and P_2 the parameters $\alpha_2, \beta_2, \gamma_2, \overline{m}_2, \overline{T}_2$. Representation (8) will be most reliable if the two peaks are clearly separated. If the peaks have much overlap a more subtle representation might be required.

Analogous to (6) and (7), we have for two peaks:

$$G(m) = \alpha_1 \sqrt{\frac{\pi}{\gamma_1}}\, e^{-\beta_1 (m-\overline{m}_1)^2} + \alpha_2 \sqrt{\frac{\pi}{\gamma_2}}\, e^{-\beta_2 (m-\overline{m}_2)^2} \tag{9}$$

$$H(T) = \alpha_1 \sqrt{\frac{\pi}{\beta_1}}\, e^{-\gamma_1 (T-\overline{T}_1)^2} + \alpha_2 \sqrt{\frac{\pi}{\beta_2}}\, e^{-\gamma_2 (T-\overline{T}_2)^2} \tag{10}$$

Both G and H contain 8 parameters out of the 10 parameters in total. The coordinates $\overline{m}_1$ and $\overline{m}_2$ must be estimated from G data, and the coordinates $\overline{T}_1$ and $\overline{T}_2$ from H data. Since 10 parameters must to be estimated the minimal number of data points is 10. In practice more data is available.

We remark that problems arise if the two peaks are nearly identical in shape. In Fig. 2a and 2b the top positions of the two identical peaks are indicated. In Fig. 2a the coordinates of the two peaks are $(\overline{m}_1, \overline{T}_1)$ and $(\overline{m}_2, \overline{T}_2)$, whereas in Fig. 2b these coordinates are $(\overline{m}_1, \overline{T}_2)$ and $(\overline{m}_2, \overline{T}_1)$. Both configurations are clearly different, but they give rise to the same partial data $G(m)$ and $H(T)$. So, in this special case the partial data do not provide enough information to discriminate between the two configurations. It is not expected that this situation will often appear in a practical data set. If the estimation procedure yields parameter sets which are nearly identical for both peaks, the reconstruction under consideration should be treated with extra care.

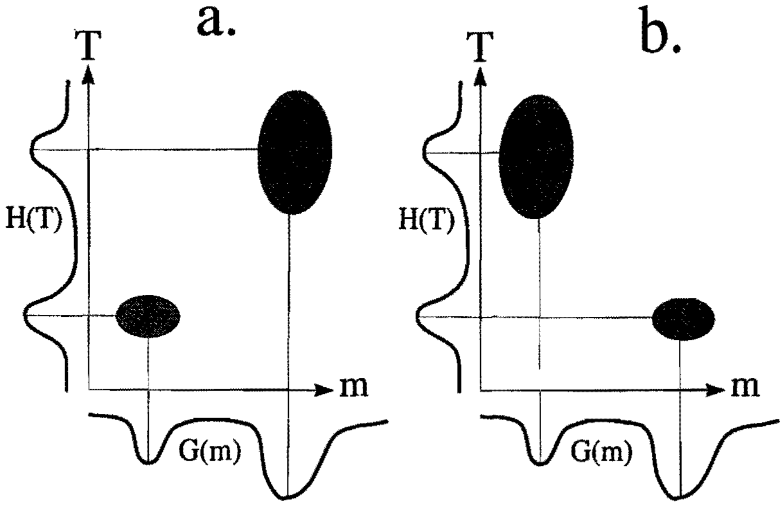

Fig. 2. Sketch of distributions with two peaks. The different distributions under a. and b. give rise to similar partial data if the peaks are very similar in height and shape.

Reconstruction of N peaks ($N > 2$)

The reconstruction of N peaks ($N > 2$) can be done along the same lines as outlined above. One could use, e.g., the representation

$$F(m,T) = \sum_{k=1}^{N} P_k(m,T) \tag{11}$$

with each P_k containing the parameters $\alpha_k, \beta_k, \gamma_k, \overline{m}_k$, and $\overline{T}_k$. So, for N peaks the distribution F contains $5N$ parameters which must be estimated from at least $5N$ data points. In the Appendix an efficient algorithm is presented to perform the estimation. Since many partial data points are available it is to be expected that the procedure works equally well for N peaks as for 1 peak, as long as the peaks have not much overlap.

For overlapping peaks the representation (11) must be refined. Other complications may be met if the peak positions are such that the $\overline{m}$-coordinates or the $\overline{T}$-coordinates of different peaks are nearly equal. In Fig. 3 such a situation with 4 peaks is sketched. Even if the peaks considerably differ in the parameters α, β, and γ, the reconstruction may be hard in this case. One then needs extra information, e.g. the number of peaks.

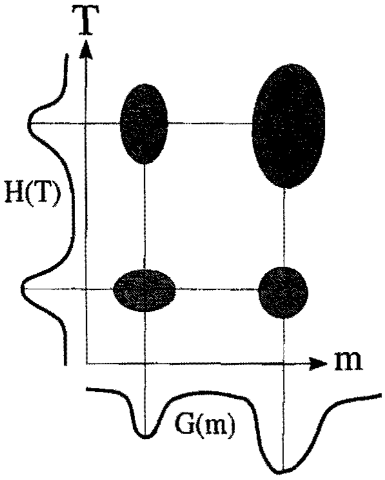

Fig. 3. Sketch of a situation in which the reconstruction procedure outlined in the Appendix, section A4, is not directly applicable.

Conclusions

We conclude that reconstruction of the SCB-MW distribution from partial (integrated) data is possible if an appropriate representation of the distribution is available. Such a representation contains parameters which characterize the peaks. The problem then reduces to the estimation of these parameters. The estimates of these parameters are in general given by the solution of a set of nonlinear equations. The complexity of the reconstruction procedure depends heavily on the number of peaks and their configuration. The present analysis yields the following insights:

a) Reconstruction of one peak is fast and simple since the estimation of the parameters can be reduced to solving five linear equations.

c) Reconstruction of two or more peaks is simple if the problem may be reduced to repeated application of the procedure under a). However, if the configuration is as shown in Fig. 3 the estimation procedure is more complicated. It is then of great help if in advance the number of peaks is known. The general way to apply reconstruction in these cases is to use a general estimation procedure based on the least-squares approach as described in the Appendix. However, in practice one could rather develop an extension of the method under a) for one peak.

Future research concerns the implementation of the reconstruction method proposed here and testing of its applicability to measured data.

Appendix

In this Appendix we present some mathematical details of the reconstruction procedure. The resulting formulae can be used in the numerical implementation of the method. Before dealing with the problem in its general form, we give a simple application for illustrative purposes. Thereafter, we treat the problem in a general setting.

A1. Reconstruction of one peak

In general the data are given at N_1 points m_i, $i = 1, ..., N_1$ and at N_2 points T_j, $j = 1, ..., N_2$ and we use the notations

$$G_i \equiv G(m_i) = \alpha\sqrt{\tfrac{\pi}{\gamma}}\; e^{-\beta(m_i - \overline{m})^2}, \quad i = 1, ..., N_1 \ , \tag{A1}$$

$$H_j \equiv H(T_j) = \alpha\sqrt{\tfrac{\pi}{\beta}}\; e^{-\gamma(T_j - \overline{T})^2}, \quad j = 1, ..., N_2 \ . \tag{A2}$$

Let us first look at the reconstruction of one peak, with as data 4G values and one H value. So, we have $N_1 = 4$, $N_2 = 1$. We shall show that the parameters $\alpha, \beta, \gamma, \overline{m}$, and $\overline{T}$ can be found via explicit formulae. Taking the natural logarithm of both sides of (A1) we obtain

$$\ln G_i = \ln\alpha + \tfrac{1}{2}\ln\pi - \tfrac{1}{2}\ln\gamma - \beta(m_i - \overline{m})^2, \quad i = 1, ..., 4 \ . \tag{A3}$$

The parameter $\overline{m}$ follows from the equation

$$\frac{\ln G_1 - \ln G_2}{\ln G_3 - \ln G_4} = \frac{(m_1 - \overline{m})^2 - (m_2 - \overline{m})^2}{(m_3 - \overline{m})^2 - (m_4 - \overline{m})^2} \ , \tag{A4}$$

which has the explicit solution

$$\overline{m} = \frac{1}{2}\,\frac{m_1^2 - m_2^2 + p(m_4^2 - m_3^2)}{m_1 - m_2 + p(m_4 - m_3)} \ , \tag{A5}$$

where the factor p is shorthand notation for the lefthand side of (A4):

$$p \equiv \frac{\ln G_1 - \ln G_2}{\ln G_3 - \ln G_4} \ . \tag{A6}$$

If an estimate for $\overline{m}$ is known, the parameters α, β, γ directly follow from the equations (A3), which are linear in these parameters. For this purpose standard techniques such as Gauss elimination can be applied to the first three equations of (A3). Eventually, the parameter $\overline{T}$ is found from the H data:

$$\ln H_1 = \ln\alpha + \tfrac{1}{2}\ln\pi - \tfrac{1}{2}\ln\beta - \gamma(T_1 - \overline{T})^2 \ , \tag{A7}$$

which is obtained by taking logarithms of both sides of (A2). So, the simple case of reconstruction of one Gaussian peak is thus straightforward.

A2. Reconstruction in general

Next, we present the general formulation of the reconstruction algorithm. It should be emphasized that this formulation is given for completeness' sake. In practice, application of the general formulae without incorporating extra expert knowledge may lead to badly converging numerical procedures.
In the following we denote the number of peaks by N. If we integrate representation (11) over the m and T variable respectively, we obtain expressions similar to (9) and (10):

$$G(m) = \sum_{k=1}^{N} \alpha_k \sqrt{\frac{\pi}{\gamma_k}}\, e^{-\beta_k(m-\overline{m}_k)^2} \ , \tag{A8}$$

$$H(T) = \sum_{k=1}^{N} \alpha_k \sqrt{\frac{\pi}{\beta_k}}\, e^{-\gamma_k(T-\overline{T}_k)^2} \ . \tag{A9}$$

In total we thus have $5N$ unknown parameters $\alpha_k, \beta_k, \gamma_k, \overline{m}_k, \overline{T}_k$, $k = 1, ..., N$, which have to be estimated from the $N_1 + N_2$ data points

$$G_i \equiv G(m_i), \quad i = 1, ..., N_1 \ ,$$

$$H_j \equiv H(T_j), \quad j = 1, ..., N_2 \ .$$

The minimal number of necessary data points is $N_1 + N_2 = 5N$. In practice one usually has $N_1 + N_2 \gg 5N$. This excess of information suggests to solve the estimation problem via the least squares method. In this approach one introduces the object function

$$E \equiv \sum_{i=1}^{M_1} (G_i - G(m_i))^2 + \sum_{j=1}^{M_2} (H_j - H(T_j))^2 \ . \tag{A10}$$

This function depends on the parameter set $\alpha_k, \beta_k, \gamma_k, \overline{m}_k$, and $\overline{T}_k$, $k = 1, ..., N$ and measures for specific values of the parameters the discrepancies between the representations $G(m)$ and $H(T)$ and the data points. If E attains its minimal value,

the corresponding parameters are optimal.
A necessary condition following from this minimization is that all derivaties of E with respect to the parameters vanish. This leads to $5N$ nonlinear equations that should be satisfied. However, in practice one rather minimizes the function of E directly in the $5N$-dimensional parameter space. For this, many standard methods are available. The success of this approach strongly depends on the quality of the data and the initial guess which can be provided for the parameters. These initial values must follow from a separate (rough) estimation procedure.

A3. Reconstruction in practice

As already mentioned above, the procedure outlined in section A2 is hard to implement and does not give the guarantee of success provided that the initial values are chosen with great care. In practice, one could use extra information or SCB-MW distributions. In these distributions the peaks are mostly well separated from each other. Furthermore, in the application under consideration, i.e. prediction of bulk properties, the detailed form of the peaks is not important and a representation in the form of a sum of Gaussian peaks as in (8) and (11) suffices. Another important aspect is that in general the number N of peaks is known in advance. It may happen that one of the peaks is ver low or even invisible, but also then the reconstruction procedure can be applied with the N value fixed.
The measured $G(m)$ and $H(T)$ profiles often provide direct information about the coordinates of the peaks. Given a rough estimate of these coordinates for one peak, information near this peak can be used to find reliable estimates for all parameters of the peak by means of the procedure worked out in section A1.
A reconstruction algorithm, which is expected to yield satisfactory results, is given below.

A4. Reconstruction algorithm

1. Find from $G(m)$ and $H(T)$ data rough estimates for the peak coordinates $(\overline{m}_k, \overline{T}_k)$, $k = 1, ..., N$.

2. Select per peak 5 data points which contain nearly only information on that peak and not on other peaks. Use the algorithm sketched in section A1 to estimate the parameters, including improved estimates for $\overline{m}$ and $\overline{T}$, for each peak separately.

3. Apply the algorithm sketched in section A2 using the results from step 2 as initial values.

Remarks

a) It will often happen that step 3 can be omitted.

b) The procedure in step 2 may fail in a situation as sketched in fig. 3. For that particular case the single-peak procedure in section A1 is not applicable. However, it is not hard to extend this procedure to a two-peak procedure. We shall not work out this extension here in detail, but use can be made of the fact that estimates for the peak coordinates $(\overline{m}_1, \overline{T}_1)$ and $(\overline{m}_2, \overline{T}_2)$ can be deduced from the G and H profiles directly.