Chapter 1

Evolution of Statistical Models of Overlap

1.1 Introduction

Whether dealing with the spatial decomposition of a chemical mixture, a temporal distribution of occurrences from a radioactive source, or some other random process where the events due to the detection technique have a measurable bandwidth, event coordinates may be obliterated and obscured due to overlap. In situations where resolution is an artifact of the process rather than the sensor, it is easily visualized how a group of events with some associated breadth, all occupying a small localized region, may fuse giving the illusion of one occurrence. It also follows that any distinguishing marker of the resulting signature of the conglomerate after detection may not coincide with the equivalent marker of any of the single event signatures had they occurred independently. Thin-layer chromatography, gas (Fig. 1.1) and liquid chromatography and capillary electrophoresis are just some examples of processes with inherent overlap.

To date many models have been set forth to gauge the extent of overlap in separations and yield some sort of educated guess as to the number of underlying components that are present. These models, under reasonable levels of saturation or relative component density, for the most part tend to yield excellent predictions drawing on statistics and the separation science knowledge base. The focus here is two equations previously put forth to model the overlap of one and two-dimensional separations that were recently modified by Davis [2] to deal with nonhomogeneous distributions. The result is more robust algorithms that can gauge the quality of a separation by predicting the number of concealed components that are resident.

Figure 1.1: A portion of a gas chromatogram taken of lime oil.

1.2 Theory

Since we are dealing with the random errors that affect the precision of an outcome, we must consider the effect the uncertainty in the number of observable peaks will have on the precision of the determination of the number of underlying components. That is to say if one were to analyze a particular separation, containing a finite number of components whose positions are governed by some global distribution function, then the arrangement of peaks in the separation is only one of an infinite number of possible assortments. Being that the outcome becomes unique only as the number of components becomes infinite, the need to consider the effect of the statistical fluctuations in repeating a separation under the same governing frequency becomes clear. The frequency, describing where the single component profiles are likely to be, in this case parallels the role of the probability density function when modeling a separation mathematically.

With this objective in mind, we present the least squares regression expression in its statistically rigorous form with weighting factors. These factors, being just the reciprocals of the variances of the number of observable peaks, dictate the need for a method of finding these variances when the single component profiles of a separation may be described by a variable Poisson density. We thus outline, to the best of our knowledge, the first derivation of the uncertainty in the number of these peaks or the variance of a thinned Poisson process of variable density. Described also is the Bayesian approach for determining the most likely number of components from a separation containing a given number of visible peaks, which presents itself as a natural by-product of the analysis.

1.2.1 The One-Dimensional Model for Separations

To facilitate our understanding of the one-dimensional model consider a point Poisson process of infinite extent and constant density λ describing component positions. The probability of m components occurring within any interval t is described by the Poisson probability density function

$$P(m) = \frac{(\lambda t)^m e^{-\lambda t}}{m!}; \quad m = 0, 1, 2, \cdots . \tag{1.1}$$

It follows that the probability that no components will occur within an arbitrary fixed interval of x_o is $e^{-\lambda x_o}$ by setting $t = x_o$ and $m = 0$ in the above expression. This is also the probability that the interval between any two adjacent components, which we shall call a deviate, will be greater than x_o. The average number of deviates on a length, Δx, greater than x_o will be $\Delta x \lambda e^{-\lambda x_o}$. This is the mean number of deviates on the length, which is $\Delta x \lambda$ since for a continuous process the deviate density equals the component density, multiplied by the probability, $e^{-\lambda x_o}$, that any one deviate exceeds x_o. If one wishes to find the mean number of deviates that exceed x_o on a measure of a point Poisson process over which the event density is not constant, one would simply sum local means on all the segments of that measure, which are made small enough so that one may assume the density is constant over each:

$$\lim_{\Delta x \to 0} \sum_k \Delta x_k \lambda_k e^{-\lambda_x x_o} \Longrightarrow \int \lambda(x) e^{-\lambda(x) x_o} \, \mathrm{d}x. \tag{1.2}$$

If we assume for the moment (this point will be clarified later) that there is a one to one correspondence between the number of visible peaks, p, in our model and deviates exceeding x_o, then we may write

$$\overline{m} \int f(x) e^{-\overline{m} f(x) x_{oi}} \, \mathrm{d}x = \overline{p}_i, \tag{1.3}$$

which is the Davis equation [2]. Here the substitution $\overline{m} f(x) = \lambda(x)$ has been made where $f(x)$ is the frequency, or probability density function, for components normalized, by definition, over the span of x. Integration of both sides reveals $\overline{m}$ to be the mean number of components, whose positions are described by a variable Poisson process with a density function $\lambda(x)$, within the span of x. This is also the expectation value of m, the actual number of components within the span. Note that m is integral but $\overline{m}$ need not be. As it originally appears, the integration of the Davis equation takes place over the interval [0,1]. Here, without loss of generality, the integration is taken to be over the span of the dependent variable x, which is also the extent of the separation within our model, with the understanding that $f(x)$ is normalized on this domain, so that rescaling is not necessary.

In an actual separation the placement of the m components is governed by such things as their chemical nature, method of sample introduction, the physical properties of the separation medium and so on. The

Figure 1.2: A typical p-plot. The initial plateau, represented by the unfilled circles, is due to overlap and is not included in the fit. Filled circles are placed at the centers of the ‘shelves’ that are defined through the function of $p(d_o)$. This pattern was generated from a uniform random density and the error bars shown represent $\pm\sigma_p$ calculated from equation (1.20).

combined effect of all these considerations may be represented by a frequency function. Although this function must be estimated, in an actual analysis, from the separation itself [2, 4], we assume here that we have the ability to deduce it exactly. The parameter x_o, described in more detail later, is representative of the bandwidth of each single component profile. The wider the bandwidth the less peaks are visible due to a fusion of multiple components into only one observable signal. The reader is reminded again at this point that fixing all the values that appear on the left side of equation (1.3) will still yield different values for the number of visible peaks, p, under multiple, fixed procedure, separations of identical component mixtures. Although each arrangement of single component locations is still governed by the same frequency function, the exact placement from trial to trial will vary just as successive rolls of a die would offer different numbers even though the probabilities have not changed. We take a set of so generated separations to constitute an ensemble. As the number of members in our ensemble increases without bound the average number of visible peaks tends toward $\overline{p}$ appearing on the right side of equation (1.3). The subscript i is used to distinguish between parameters relating to the same separation but taken under conditions where different values of x_o apply. As is described in Davis' paper [2, 5], we may in effect artificially increase the bandwidth and calculate for each x_{oi} the number of peaks p_i that are visible. In so doing we are left with more data to fit. In Fig. 1.2 one can see p_i plotted against d_{oi}, which is the counterpart to x_{oi} in two dimensional separations.

We chose the method of least squares when fitting the left side of equation (1.3) to $\{p_i\}$ as a means of estimating the free parameter m. Throughout this chapter braces are taken to symbolize the set of all members over the range of the free subscript. If we take each $\{p_i\}$ as having been sampled from a parent normal distribution $N_i(\mu_{pi}, \sigma_{pi}^2)$ and further assume that $\{\sigma_{pi}^2\}$ may be found, maximum-likelihood estimation may be invoked. By associating $\{\mu_{pi}\}$ with expression (1.3) for $\overline{p}_i$ we write the likelihood function as

$$L(p_i, \sigma_{pi}; m) = \prod_i \frac{1}{(2\pi\sigma_{pi}^2)^{1/2}} e^{-\frac{1}{2}(\frac{p_i-\mu_{pi}}{\sigma_{pi}})^2}. \tag{1.4}$$

It is clear that the above expression takes on its maximum value upon minimizing the statistic

$$\sum_i \frac{1}{\sigma_{pi}^2}(p_i - \overline{m}\int f(x)e^{-\overline{m}f(x)x_{oi}}\,\mathrm{d}x)^2, \tag{1.5}$$

Figure 1.3: The top diagram windows a portion of a Poisson process and shows all the events as dots. The bottom diagram shows the same process where the clusters have been replaced by single dots residing at the weighted center of all the events constituting the cluster.

hence we have the criterion for the most probable $\overline{m}$ which is also our best guess at m.

1.2.2 Weighting Factors on Fitted Data

Key to the above analysis, and also for gauging the uncertainty in the estimation of m as we will see later, is the determination of $\{\sigma_{pi}^2\}$. Before embarking on this task, let us state our objective clearly by describing in detail the mathematical model that will be used. Consider a one dimensional window, of arbitrary dimension, incident upon a point Poisson process of density $\lambda(x)$ which need not necessarily be constant. Further take that process to be thinned in accordance with the parameter x_o. By thinning we mean that any two Poisson points, or events, separated by a distance less than the resolution parameter, x_o, are unresolvable from one another, illustrated in Fig. 1.3. Note also that these intervals may be temporal if the separation were carried out in time. We take a group of such unresolvable events to constitute a 'cluster'. A cluster may also be a single event separated from its nearest neighbors on both sides by a large enough distance. In general the left-most event in a cluster will be separated from the event on its left by a distance greater than x_o and the right-most event in a cluster will be separated from the event on its right by a distance of greater than x_o. No distance separating two events within a cluster will exceed the resolution parameter. One may take an ensemble of such prepared systems to find the variance σ_p^2 of the number of clusters in the usual way.

First windowing a region of constant density and unit length from a continuous Poisson point process, we seek to arrive at the variance, $E(p^2) - [E(p)]^2$, through a knowledge of the probability of occurrence of every possible state. The idea is to first develop the ability of describing the variance of clusters, which are analogous to visible peaks, on a small interval of constant density before moving on to larger intervals with variable densities. We then break up these larger intervals into tiny regions so that the density of any given region may be taken as constant. Under the assumption that the variances of individual regions are mutually independent, the variances on these smaller regions may be simply summed to arrive at the variances of p on larger intervals of variable density. As it turns out, however, this assumption of summable variances is not always valid. This issue of the variances not being consistently additive is an important one and will be dealt with further on. As before, m is representative the original number of windowed events and p the number of clusters in the thinned process. Upper case P is reserved for probability. The probability of each state may be written

$$P(p \cap m) = P(m)P(p \mid m), \tag{1.6}$$

where the first term on the right is just the Poisson density function:

$$P(m) = \frac{\lambda^m e^{-\lambda}}{m!}, \tag{1.7}$$

which is followed by the the conditional probability that p clusters will occur given that m events already have.

We may take advantage of the condition of m Poisson occurrences in the second term on the right of (1.6) to make it more manageable. With the aid of a theorem, we may describe the distribution of these Poisson events with uniformly distributed random variables. This description, in turn, will allow us to take advantage of a probability equation specifically derived for uniformly distributed random variables, presented later. The theorem [6] asserts that given a Poisson counting process takes on a value n at t, $N(t) = n$, then the n arrival times $s_1, \dots, s_n$ have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval $(0, t)$. The interpretation we shall utilize is that given n deviates have occurred in a Poisson process initiated at the origin and ending with the n^{th} event, then each of the n deviates shares an identical distribution with each of the n intervals formed by distributing $n-1$ uniform randomly occurring events in the same span. The word deviate here is defined to be the interval between the events, which is commonly referred to as an interarrival time in Poisson statistics when dealing with a temporal process. Our first deviate is then the distance from where our process was initiated to the first event. To see that the assumption follows from the theorem, let us consider ending the Poisson process, initiated at the origin, at an infinitesimal distance shy of the n^{th} event. This is a special case since the theorem has no prerequisite as to where in relation to any of the Poisson occurrences the coordinate t should be placed. By placing t essentially at the n^{th} event, the interval between the $(n-1)^{\text{st}}$ event and t becomes a deviate that has the same exponential distribution as all the other $n-1$ deviates in the process. The last interval would not have the same distribution as the other deviates if we ended the process arbitrarily between events. Finally, because both the $n-1$ Poisson 'arrival times' have the same distribution as the order statistics corresponding to $n-1$ independent random variables uniformly distributed on the interval $(0, t)$, and the n defined intervals on $(0, t)$ for the uniform process are all identically distributed, the n Poisson deviates must all have the same distribution as the n defined uniform intervals.

Armed with this information we rescale $(x' = x/t)$ so that our n^{th} event falls at unity, then the likelihood that exactly k of the n deviates are greater than a temporarily rescaled x_o becomes an exercise in geometrical probability. This problem was addressed by Kendall [7] but equation (2.36) in reference [7] is misprinted in that all the combinations inside the braces should read as $n-k$ deviates taken s at a time giving the probability as

$$\frac{n!}{k!} \sum_{s=0}^{(k+s)x_o<1} \frac{(-1)^s}{s!(n-k-s)!} [1-(k+s)x_o]^{n-1}. \tag{1.8}$$

It is noted that the terms, in this and other applicable sums, are taken to drop out for $s > n-k$ since the factorial for negative integers evaluates to $\pm\infty$, depending on whether the left or right sided limit is considered, as defined by the Gamma function. The sum may thus be terminated if this alternate condition is met first.

At this point we may deal with the Poisson component distribution of an actual separation of constant density in the following way. We let our interval of arbitrary length t originate one the first (left most) component location and terminate on the m^{th} (last and right most) component location. Since the Poisson process is now predefined, initiating the length on the first component rather than at an arbitrary origin guarantees that the first interval on the length is still a full deviate which is governed by the exponential distribution. Because our length has an integral number of Poisson deviates, we may again take advantage of the above theorem which allows us to take the now $m-1$ deviates as having the same distribution as the $m-1$ intervals defined in letting $m-2$ events occur uniformly random in the same length t. After rescaling, we use equation (1.8) to give us the probability that $p-1$ of $m-1$ deviates are equal to or exceed x_o, which is equivalent to the probability of getting p clusters from the m components.

At this point one can visualize doing the above analysis in pieces. We could use multiple lengths for regions that are described by different densities and obtain probabilities via equation (1.8) so long as each length terminated on a component location at both ends as shown in Fig. 1.4. It is, however, much more convenient to be able to work with constant lengths of arbitrary size in dealing with distributions having continuously varying densities. So that we may justify the use of these optimal lengths, the following argument is offered. We visualize a continuous Poisson point process of constant density that is completely subdivided into concurrent intervals of identical length. It is clear that any boundary, separating two intervals, will in general not coincide with an event. In this situation we may take the number of events and the number of deviates per interval to be the same on average. Because there is a one to one correspondence between deviates and points for continuous Poisson point process, we may say that the deviate density equals the point density. This may be realized by always taking the space between two events that contains an interval boundary as belonging to the interval on the left. The right interval will not be cheated of space since it in turn will in general receive a contribution from its right hand side. Another way of making this idea tangible would be to say that on average, if we considered many intervals, the interval boundaries would bisect the interevent distances. If we bring the interval ends together,

Figure 1.4: A Poison process that is split up into different sized lengths such that the ends of those lengths are always terminated on an event of the process.

each contributing its half spacing to give us a full deviate, we will have a circle on which clearly there are an equal number of events and deviates. This idea is illustrated in Fig. 1.5 and the argument will allow us to make use of the current single interval equations and apply them to interval sequences.

Within a continuous distribution we consider each deviate greater than x_o to represent the start of a new cluster since it will by definition be resolvable from the one before it. The way is now clear to substitute p for k and m for n in equation (1.8) and write for a single interval of length t

$$P(m) = \frac{(\lambda t)^m e^{-\lambda t}}{m!}, \tag{1.9}$$

$$P(p \mid m) = \frac{m!}{p!} \sum_{s=0}^{(p+s)\frac{x_o}{t}<1} \frac{(-1)^s}{s!(m-p-s)!} \left[1-(p+s)\frac{x_o}{t}\right]^{m-1}. \tag{1.10}$$

Thus

$$\sum_m \sum_p P(p \cap m) = e^{-\lambda t} \sum_{m=0}^{\infty} (\lambda t)^m \sum_{p=0}^{m} \frac{1}{p!} \sum_{s=0}^{(p+s)\frac{x_o}{t}<1} \frac{(-1)^s}{s!(m-p-s)!} \left[1-(p+s)\frac{x_o}{t}\right]^{m-1}, \tag{1.11}$$

where the number of clusters, p, is limited by the number of events and the infinite sum over m is due to the fact that it is possible, although not probable, to have any number of events in a Poisson interval even if $m >> \lambda t$. The summing in equation (1.11) over the probability of every possible state must equate to unity. This fact is verified explicitly by Rowe [13].

Under the assumption that the individual random variables $\{\sigma_{pj}^2\}$, representing the variances in a sequence of intervals, are independent, the variance of a segment comprised of n of these intervals may be written

$$\sigma_p^2 = \sum_{j=1}^{n} \sigma_{pj}^2 = \sum_{j=1}^{n} \left(\sum_{m=0}^{\infty} \sum_{p=0}^{m} p^2 P(p \cap m) - \left[\sum_{m=0}^{\infty} \sum_{p=0}^{m} p P(p \cap m) \right]^2 \right). \tag{1.12}$$

Once more, a different λ_j may be used in each interval yielding a tool for dealing with variable density Poisson segments of length nt.

Although it is desirable to make our interval length parameter, t, small when working with sharply varying densities, one is limited by the value $2x_o$. Below this critical value, variances will no longer be additive, as will be evident from what follows. Fortunately there is a way around this difficulty. The solution is built on the fact that at, or less, than the limiting value of $t = 2x_o$

$$E(p^2) = e^{-\lambda t} \sum_{m=0}^{\infty} (\lambda t)^m \sum_{p=0}^{m} \frac{p^2}{p!} \sum_{s=0}^{(p+s)\frac{x_o}{t}<1} \frac{(-1)^s}{s!(m-p-s)!} \left[1-(p+s)\frac{x_o}{t}\right]^{m-1} = E(p); t \le 2x_o. \tag{1.13}$$

The reason for this equality is that the contributions from the terms inside the sum over p vanish for $p > 1$ in accordance with the upper limit on the innermost sum over s. Since $p^2 = p$ for $\{p : 0, 1\}$ we arrive at the conclusion expressed in equation (1.13). After explicit evaluation of this expression at $t = 2x_o$ we see

$$E(p) = E(p^2) = e^{-\lambda 2x_o} \sum_{m=1}^{\infty} \frac{(\frac{1}{2})^{m-1}(\lambda 2x_o)^m}{(m-1)!} = \lambda 2x_o e^{-\lambda 2x_o} \sum_{m=1}^{\infty} \frac{(\lambda x_o)^{m-1}}{(m-1)!} = \lambda 2x_o e^{-\lambda x_o}; t = 2x_o. \tag{1.14}$$

In general using equation (1.11) to arrive at the expectation value for p, through incorporating another factor of p inside the sum over p, will yield

$$E(p) = \lambda t e^{-\lambda x_o}, \tag{1.15}$$

Figure 1.5: The top diagram shows a sequence of random events spanning three intervals. We consider the central interval and see that on average its borders will fall halfway between two events. On average then the two ‘half deviates’, d1 and d2, will sum to a pseudo deviate. We may bring the two half deviates together by adjoining the ends of the interval together to form a circle. On this circle then there will be an equal number of events and deviates.

which is valid for any value of $t > x_o$ by induction and is known to be true for any value of t from Poisson statistics.

We may now write the variance, $E(p^2) - [E(p)]^2$, on an interval of length $x_o < t \leq 2x_o$ and constant density, as a closed form expression:

$$\sigma_p^2(\lambda, t \leq 2x_o) = \lambda t e^{-\lambda x_o}(1 - \lambda t e^{-\lambda x_o}). \tag{1.16}$$

Although valid for all $x_o < t \leq 2x_o$, this expression for variance is not linear in t and hence the variance calculated on intervals of size $t < 2x_o$ will not be additive. This result is not surprising when one realizes that the events whose variance we are considering, deviates greater than or equal to x_o which will be the number of clusters, are not point events but rather events that have a measurable extent. A deviate of $3x_o$, a reference point of which occurs in an interval of $t = x_o$, will preclude the identical reference point of any other deviate from occurring within at least one of the two neighboring intervals. Starting from the variance on a constant density interval at the critical length of $t = 2x_o$,

$$\sigma_p^2(\lambda, t = 2x_o) = 2x_o\lambda e^{-\lambda x_o}(1 - 2x_o\lambda e^{-\lambda x_o}), \tag{1.17}$$

we can again prove by induction that the variance on any constant density interval of length $t \leq 2x_0$, is

$$\sigma_p^2(\lambda, t \geq 2x_o) = \lambda t e^{-\lambda x_o}(1 - 2x_o\lambda e^{-\lambda x_o}). \tag{1.18}$$

Because this equation is linear in t it may be divided by t to give a variance density:

$$\frac{\sigma_p^2}{t} = \lambda e^{-\lambda x_o}(1 - 2x_o\lambda e^{-\lambda x_o}), \tag{1.19}$$

which can be integrated over intervals where the density is free to vary:

$$\sigma_p^2 = \int \lambda(x)e^{-\lambda(x)x_o}[1 - 2x_o\lambda(x)e^{-\lambda(x)x_o}]\,\mathrm{d}x. \tag{1.20}$$

It may at first seem strange that we derived equation (1.20) from an expression valid only at larger intervals. Let us relate this to the task of finding the mass of a pile of sand. In this case one would without hesitation take the mass to be

$$\rho \int \mathrm{d}v. \tag{1.21}$$

Faith in this mass must be based on the assumption that ρ correctly describes the density of the sand which is estimated by taking a large known mass of sand and dividing by the volume it occupies. Any density fluctuations on a small scale are washed out by the law of averages when considering sufficient volumes. This is what we have done with equation (1.19). If we tried to use equation (1.21) to give us information about sand on volumes on the order of sand grain sizes, or even less, we would find the information incorrect. On these tiny volumes we may no longer consider the sand as a fluid but rather must acknowledge its structure, the individual grains, and describe it by a suitable equation, as equation (1.16) does for variances on 'cluster sized' intervals.

It deserves mention that lower limit of $t > x_o$ placed on equations (1.15) and (1.16) are the consequence of the model being used in conjunction with the scaled Kendall expression. Equation (1.10) will correctly predict a probability of zero for any interval containing a deviate longer than the interval itself. In our model, however, we are considering interval sequences. Two neighboring events occurring several intervals apart may indeed be separated by a distance greater than x_o, even if the individual intervals are smaller, which we may choose to associate with a particular interval. With $t > x_o$, an interval containing a single event will yield a single deviate of greater than x_o since each half spacing will add to the length of the interval itself. The probabilities of such an interval containing no deviates greater than x_o are immaterial since this situation will no contribute to our sum over all states ($E(p) = E(p^2) = 0 \quad \text{since} \quad p = 0$). Equation (1.16), although qualified only for $x_o < t \leq 2x_o$, was found to yield predictions in agreement with simulations for $t \leq x_o$. These simulations were carried out on the basis of associating deviates with the intervals that contained their centers. Equation (1.15), as stated, is known valid for all t.

Having now a theoretical expression for $\{\sigma_{pi}^2\}$, we are in a position to advance in our discussion to the uncertainty of m. To facilitate our understanding, one possible sequence of steps leading to the well known expression for the variance of a dependent parameter is outlined. This expression is of great utility in the physical sciences and similar steps leading to this equation can be found in many good texts dealing with data analysis [8]. To start with let us return to the notion of an ensemble of n identically prepared systems mentioned above. Every choice of x_{oi} will yield a set, $\{p_{ij}\}$, which will begin to define a distribution as n tends toward infinity. Here the first indice, i, identifies the particular resolution parameter and the second, j, the specific ensemble member. We also know that, given any system, m is defined by $\{p_i\}$ through the minimization of equation (1.5). This fact is expressed by writing m as a function of $\{p_i\}$ for an ensemble member:

$$m_j = f(\{p_{ij}\}). \tag{1.22}$$

Taking these two things in conjunction, let us postulate that the best value for m is found by using the expectation values from our ensemble in the calculation

$$\overline{m} = f(\{E(p_i)\}). \tag{1.23}$$

In rewriting the sample variance,

$$\sigma_m^2 = \frac{1}{n-1}\sum_{j=1}^{n}(m_j - \overline{m})^2, \tag{1.24}$$

let us express to first order

$$m_j - \overline{m} \approx \sum_i \left(p_{ij} - \frac{1}{n}\sum_{k=1}^{n} p_{ik}\right)\left(\frac{\partial m}{\partial p_i}\right). \tag{1.25}$$

Combining the two previous equations gives us

$$\sigma_m^2 \approx \frac{1}{n-1}\sum_{j=1}^{n}\left[\sum_i (p_{ij} - E(p_i))\left(\frac{\partial m}{\partial p_i}\right)\right]^2, \tag{1.26}$$

which upon neglecting cross terms reduces to

$$\sigma_m^2 \approx \sum_i \sigma_{pi}^2 \left(\frac{\partial m}{\partial p_i}\right)^2. \tag{1.27}$$

Justification for the omission of these cross terms hinges on lack of correlation. Take, for example, any of the mixed partials,

$$2\left(\frac{\partial m}{\partial p_{i'}}\right)\left(\frac{\partial m}{\partial p_i}\right)\sum_{j=1}^{n}(p_{i'j} - \overline{p}_{i'})(p_{ij} - \overline{p}_i), \tag{1.28}$$

and consider the sum. Each of the two differences, for any j, has a 50/50 chance of being positive or negative. On the whole then the product is also equally likely to be less or greater than zero. Once more for a large sample there will be a well randomized assortment of values and thus this sum will tend toward zero as n grows without bound.

Returning now to our specific problem of finding the uncertainty in the estimate for the number of single component peaks, we employ techniques of calculus and set the partial derivative of equation (1.5) with respect to $\overline{m}$ equal to zero:

$$0 = \sum_i \frac{2}{\sigma_{pi}^2}\left(p_i - \overline{m}\int f(x)e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x\right)\left[\overline{m}x_{oi}\int (f(x))^2 e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x - \int f(x)e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x\right]. \tag{1.29}$$

Here we have treated σ_{pi}^2 as a constant, specifically ignoring its dependence on $\overline{m}$ only to make the equations short enough to present the closed form technique. As soon as the equality in equation (1.29) is set, $\overline{m}$ becomes a constant taking on the value that minimizes equation (1.5). Our goal is to find

$$\sigma_{\overline{m}}^2 = \sum_j \sigma_{pj}^2 \left(\frac{\partial \overline{m}}{\partial p_j}\right)^2, \tag{1.30}$$

where we take j, replacing the i in equation (1.27), to be now just a dummy index rather than the usual ensemble member label. To get the partials in equation (1.30), we must implicitly differentiate both sides of equation (1.29) with respect to p_j and then solve. Writing out only the $i = j$ term we have

$$\begin{aligned}0 = \frac{2}{\sigma_{pj}^2}\left(p_j - \overline{m}\int f(x)e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x\right)\frac{\partial\overline{m}}{\partial p_j}\Bigg[-\overline{m}x_{oj}^2\int (f(x))^3 e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x + x_{oj}\int (f(x))^2 e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x \\ +x_{oj}\int (f(x))^2 e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x\Bigg] + \frac{2}{\sigma_{pj}^2}\left(1 - \frac{\partial\overline{m}}{\partial p_j}\int f(x)e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x + \frac{\partial\overline{m}}{\partial p_j}\overline{m}x_{oj}\int (f(x))^2 e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x\right) \\ \times\left[\overline{m}x_{oj}\int (f(x))^2 e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x - \int f(x)e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x\right]\end{aligned} \tag{1.31}$$

The rest of the $i \neq j$ terms all have the same form with the exception that the '1' vanishes. Solving for our partials we get

$$\begin{aligned}\frac{1}{\sigma_{pj}^2}\left[\int f(x)e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x - \overline{m}x_{oj}\int (f(x))^2 e^{-\overline{m}x_{oj}f(x)}\,\mathrm{d}x\right] = \frac{\partial\overline{m}}{\partial p_j}\sum_i \frac{1}{\sigma_{pi}^2}\times \\ \Bigg\{\left(p_i - \overline{m}\int f(x)e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x\right)\left[2x_{oi}\int (f(x))^2 e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x - \overline{m}x_{oi}^2\int (f(x))^3 e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x\right] \\ + \left[\overline{m}x_{oi}\int (f(x))^2 e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x - \int f(x)e^{-\overline{m}x_{oi}f(x)}\,\mathrm{d}x\right]\Bigg\}.\end{aligned} \tag{1.32}$$

Substituting expressions (1.20) and (1.32) into our sum, (1.30), we finally arrive at a value and thus we can place our uncertainty on $\overline{m}$.

To be rigorously correct, however, the explicit $\overline{m}$ dependence of σ_{pi}^2 must be considered for a Poisson model and equation (1.29) will take on a much lengthier form. Faced with this, it is felt that the partials in equation (1.30) can be more easily evaluated numerically. The manner in which the j^{th} partial was obtained was to recalculate the least squares $\overline{m}$ from equation (1.5) for each of two modified sets of p-plot coordinates, $\{p_i\}$. The modification leaves all but the $i = j$ coordinates unaltered. The particular, p_j, coordinate is replaced by a new coordinate which is in turn higher then lower than the original. From the two calculated $\overline{m}'^{\text{s}}$ one can gauge the partial dependence as $\Delta\overline{m}/\Delta p_j$. Care must be taken in selecting Δp_j. One would like Δp_j small enough to guarantee an accurate derivative without being so small that the value obtained is effected by the precision limitations of the algorithm. From tracing minimum changes on successive trials Δp_j was chosen to be 10% of the number of coordinates being fit since more coordinates would drop the dependence of m on any one.

It is added that the latter part of equation (1.6) may be inverted using Bayes' formula:

$$P(m_k \mid p) = \frac{P(m_k)P(p \mid m_k)}{\sum_i P(m_i)P(p \mid m_i)}, \tag{1.33}$$

to provide one with an alternate method for estimating m given p. This idea of Bayesian estimation is commonly employed [1, 3, 9] when considering *a posteriori* probability density functions. Unless there is information on the density available, each m must be considered equally likely prior to the assignment of p in which case the previous equation simplifies to

$$P(m_k \mid p) = \frac{P(p \mid m_k)}{\sum_i P(p \mid m_i)}. \tag{1.34}$$

Both equations as presented will provide, via repeated application over every possible value for m, $(p \leq m < \infty)$, a discrete distribution function for m. If one wishes to only compare a few relative probabilities with no interest in the entire distribution, the denominators may be set to unity to expedite the calculation. In general the distribution for m as described by equation (1.34) will have two local maxima, just as equation (1.3) in the form $p - me^{-mx_o}$ has two roots. It is therefore useful to posses some knowledge of the original system in question so that one may rule out one of these roots as an unfeasible solution.

1.2.3 Two-Dimensional Separations

An equation put forth by Roach:

$$\overline{p} = \frac{4\alpha\overline{m}}{e^{4\alpha} - 1}; \quad \alpha = \frac{\overline{m}A_o}{A}, \tag{1.35}$$

Figure 1.6: The two possible triplet combinations. There is one possible doublet, four quartets, and so on.

was originally derived to model the overlap of coal particulates on precipitators [10]. Here A is the area on which the equation is considered and $A_o = \pi d_o^2/4$ is the effective area of a single component spot having an effective diameter of d_o. The mean number of components is identified with $\overline{m}$ and the average number of distinguishable coal spots with $\overline{p}$. The parameter A_o can be identified with its one-dimensional counterpart, x_o, in the sense that the saturation, α, which specifies the component density of a separation, may always be thought of as the quotient of either with the span of the separation, all scaled by the mean number of components.

In this form the Roach equation is not very useful in interpreting real two-dimensional separations. This again is due to the fact that real separations rarely contain components that are distributed in a simple, random nature but rather display some type of spatial ordering. The Roach equation may be adapted in a straightforward way to address this issue and leave us with a modified equation that is both theoretically more powerful and of greater practical utility than its predecessor [4]. By letting $\overline{m} = \lambda A$ we may express equation (1.35) as a differential:

$$\mathrm{d}\overline{p} = \frac{4A_o\lambda(a)^2}{e^{4A_o\lambda(a)} - 1}\,\mathrm{d}a, \tag{1.36}$$

where we take the lower case a to represent a local area coordinate and identify λ with component density. By summing all such contributions over the area, A, on which a separation takes place, we are left with our modified equation for the expected number of spots in a two-dimensional separation with variable component density. Expressing this in terms of the effective diameter, d_o, we write

$$\overline{p} = \pi d_o^2 \overline{m}^2 \iint\limits_A \frac{f(a)^2}{e^{\pi d_o^2 \overline{m} f(a)} - 1}\,\mathrm{d}a, \tag{1.37}$$

where the density is now a variable function, $\lambda(a) = \overline{m}f(a)$, represented by the product of a frequency, normalized on A, with the mean number of components. It is this statistic for the mean number of components, $\overline{m}$, that we will again identify with the best guess at the number of components, m, when analyzing any one separation.

Each group of overlapping single component profiles, or spots, may be classified by how many individual components, ν, constitute its makeup. If one had an interest in triplets, $\nu = 3$, then two types of spots would have to be considered. The first kind of triplet would have profile circles interlocking in such a manner as to form a chain where the two end circles intersect only the central one and not each other. The second type would then be the more ‘compact’ orientation where each circle intersected the other two as seen in Fig. 1.6. As ν increases more possible spot combinations are possible. In accordance with Roach's theory, the expected number of spots, p_ν, containing ν components is [11]

$$p_\nu = \frac{\overline{m}(1 - e^{-4\alpha})^{\nu-1}}{\nu e^{4\alpha}}. \tag{1.38}$$

A modification paralleling the one outlined above leads to a formulation:

$$p_\nu = \frac{\overline{m}}{\nu} \iint\limits_A \frac{f(a)(1 - e^{-\pi d_o^2 \overline{m} f(a)})^{\nu-1}}{e^{\pi d_o^2 \overline{m} f(a)}}\,\mathrm{d}a, \tag{1.39}$$

that is again capable of handling situations where spot distributions are not homogeneous.

As in the one-dimensional case, estimation of component number is carried out through minimizing the sum of squares between a set of data, $\{p_i\}$, and the theoretical curve given by equation (1.37) by letting the effective single component spot diameter, d_o, be the independent variable:

$$\sigma^2 = \sum_i \left[p_i - \pi d_{oi}^2 \overline{m}^2 \iint_A \frac{f(a)^2}{e^{\pi d_{oi}^2 \overline{m} f(a)} - 1} \, \mathrm{d}a \right]^2 . \tag{1.40}$$

The square of sigma here, representing the sum of squares, will be henceforth labeled 'SSQ' so as not to cause ambiguity. As before, each p_i enumerates the number of clusters formed from overlapping single component spots while the effective diameter takes on an arbitrary value d_{oi} and $f(a)$ is the frequency that governs the location of those individual components. Note also that the weighting factors have been dropped for lack of a theory able to deal with the required variances of thinned two-dimensional Processes.

With the aid of Fig. 1.2 we review the typical structure of a 'p-plot'. Here p_i verses d_{oi} is plotted. What becomes apparent is the existence of a plateau at the left end of the plot identified with unfilled graphic symbols. This plateau can be understood by associating the d_o, corresponding to its right edge, with the natural bandwidth of the single component profiles in the separation. Single component profiles, whose centers lie closer together than the natural bandwidth, will be fused into only a single observable peak or spot as they are called in two-dimensional separations. We take the positions of such spots to be represented by single points whose location in some sense corresponds to the centroid of area of that spot projected onto the separation plane. Spot locations then reside on the separation plane directly below the apex or maximum of the profile contours. As d_o takes on values between zero and this natural bandwidth, fusing of spots does not take place. This is because all single component profiles that reside within the natural bandwidth of one another have already been fused and are represented by a new set of points placed at the effective centers of those fused profiles. As a result these effective centers generally cannot lie within the natural bandwidth of one another. As d_o grows larger, however, these effective centers, representing single or multicomponent spots, may overlap each other generating yet a different set of effective spots and hence a different value for p_i thus giving more data to fit when minimizing expression (1.40). For each d_{oi} the number of clusters, or spots, p_i, is determined from the set of spot locations that exist for a value of d_o equal to the natural bandwidth, which is the original form of the separation. This way the values obtained are as close as is possible to those that would have resulted had the set of actual single component profile locations been known. Effectively we are just re-evaluating the separation itself for the number of clusters at each d_{oi}. The number of clusters will drop off with increasing effective diameter where in limit it will take on the value of 1.

If we had the ability to shrink the natural bandwidth we would see the plateau in the p-plot shrink and rise with it until finally, at zero, the data would assume a vertical axis intercept of m, the number of components in a particular separation. It is, in fact, the minimum least squares theoretical fit that will 'overshoot' the plateau, as if the bandwidth were zero, and intercept an ordinate as close to m as the fit will allow. For this reason the p coordinates that make up the plateau are not included in the fit. Also excluded from the fit are those p coordinates that fall at the far right. These coordinates do not hold as much information as they always approach the same asymptotic value regardless of the separation.

The independent variable, d_o, is free to vary continuously but the number of visible spots, however, is limited to integer values. As d_o increases, p will remain constant for a duration before it drops abruptly by a unit increment. The cycling of these drops will reveal a pattern of many 'shelves', one for every integer value of the ordinate below the plateau. In general, shelf length will increase with growing effective diameter as enjoined profile centers lay farther and farther apart. The data used in fits are the points that replace these shelves and reside at their centers.

1.2.4 Frequency Estimation

Because of the limited amount of statistical information at hand, namely the coordinates of spot centers, one will be unable to determine the exact frequency that governs the placement of components and appears in equation (1.40). It is therefore necessary to approximate the frequency from these spot locations. This task is made difficult for two major reasons. First, one is often dealing with a relatively small number of components which by themselves will not carry enough information to determine a frequency even if overlap were nonexistent. As an exaggerated example consider the placement of only a single point within a region described by some structured distribution function. It would be impossible to deduce the nature of the distribution from only the location of that single event. Second, in trying to determine the distribution of single component centers, we do not even have knowledge as to where all the centers lay. This is because a definite fraction of those single component centers have been obliterated and replaced by fused component spot centers. These difficulties will intensify as one proceeds to work at higher levels of saturation.

Representing the experimentally estimated frequency, $f_e(\xi,\eta)$, as a function of two spatial coordinates, (ξ,η), that identify a particular location within our separation space, we now discuss two procedures used in its determination. The first of these involves differentiating a cumulative distribution, which is in some sense the default method for handling such distributions. The second, an adaptation, involves associating an uncertainty with each spot location to help deal with errors that present themselves when sampling a finite number of discrete random variables from a continuous distribution.

A two dimensional cumulative distribution function, F(x,h), of single component locations is defined through the frequency by

$$F(\xi,\eta) = \int\limits_{-\infty}^{\xi} \int\limits_{-\infty}^{\eta} f(\xi,\eta)\,\mathrm{d}x\mathrm{d}y. \tag{1.41}$$

Because $f(\xi,\eta)$ is normalized, $F(\xi,\eta)$ represents the fraction of single component spots whose abscissas and ordinates are less than or equal to ξ and η respectively. Equation (1.41) allows us to write $f(\xi,\eta)$ as a mixed partial:

$$f(\xi,\eta) = \frac{\partial^2 F(\xi,\eta)}{\partial\xi\partial\eta}. \tag{1.42}$$

In practice this distribution function will also not be known so we proceed by numerically differentiating an estimate obtained from spot coordinates in the following manner. The separation space is divided into a rectangular grid. The defined areas serve a purpose analogous to class intervals. The largest coordinate corner (ξ,η) of each area is considered in turn. There the estimated distribution function $F_e(\xi,\eta)$ takes on the value of the fraction of the number of spot locations whose abscissas and ordinates are simultaneously less than or equal to (ξ,η). Substituting $F_e(\xi,\eta)$ into equation (1.42) leaves one with only numerical differentiation to carry out prior to obtaining $f_e(\xi,\eta)$. Multi-term central difference approximations may be used for the majority of coordinates; however, as one draws close to the various edges forward and backward difference approximations must be considered.

In the second procedure, a bi-Gaussian is associated with every spot coordinate. These bi-Gaussians are both centered at these coordinates and further described as having the same standard deviation, σ_g, in each dimension, which is equivalent to rotating a simple Gaussians about a third dimensional axis perpendicular to the separation plane. The Gaussian profile was chosen on the basis that random errors, associated with spot positions, are well known to be distributed normally (in accordance with the normal distribution $N(\mu,\sigma)$). By building up such profiles for each spot location and allowing them all to superimpose we build up a surface that begins to resemble the frequency. The portion of each bi-Gaussian that falls within the boundaries of the separation is given the same integrated volume, of the reciprocal of the number of spots, p, for normalization purposes, so that each spot is weighted equally. By using these contours, representing component probability distributions, each spot location has an adjustable amount of uncertainty in its position factored into the frequency determination. The idea here is that if an ensemble of identically prepared separations were created from an original, and the spot locations of each were recorded on a single plot as a point, then each spot on the original separation would be represented by a 'cloud' of points (Fig. 1.7). The distribution of these points would begin to define a Gaussian, by the central limit theorem, as the number of members in the ensemble increased without bound. In effect then the entire ensemble is being used, aside from locating the exact mean of each component position, to estimate the frequency. The ambiguity of mean component positions stems from the fact that spot locations in the separation being analyzed will in general not coincide exactly with the cloud centers that would be produced from the collective ensemble. Glancing at Fig. 1.8 we see another advantage of this procedure over differentiating the cummulative distribution function, in that since Gaussians are continuously differentiable the resulting estimated frequency will be smooth and independent of the placement of any class intervals. We write then the estimated probability that a particular single component spot falls within a square of edge length 2δ centered about an arbitrary coordinate (ξ,η)

$$\int\limits_{\eta-\delta}^{\eta+\delta} \int\limits_{\xi-\delta}^{\xi+\delta} f_e(\xi,\eta)\,\mathrm{d}\xi\mathrm{d}\eta = V_p^{-1} \int\limits_{\eta-\delta}^{\eta+\delta} \int\limits_{\xi-\delta}^{\xi+\delta} \sum_{i=1}^{p} e^{-(\xi-\xi_{oi})^2/2\sigma_g^2} e^{-(\eta-\eta_{oi})^2/2\sigma_g^2}\,\mathrm{d}\xi\mathrm{d}\eta, \tag{1.43}$$

Figure 1.7: The hypothetical placement of a particular peak for different members of an ensemble that are all governed by the same distribution function.

where (ξ_{oi}, η_{oi}) is the $i^{\text{th}}$of p visible peak centers and the volume under the described surface that falls within the separation area, V_p, is included for normalization.

A situation exists where the estimated frequency may be calculated in a much simpler manner thus speeding up evaluation of equation (1.40) for the sum of squares, saving computer time. The function $f(\xi, \eta)$ is called a joint probability density of two random variables. If we stipulate that these random variables be independent of one another, which is to say that the way in which a separation is carried out in the second dimension has no commonalities with the first, then we may write the joint probability density function as a product of two marginal probability density functions: [9]

$$f(\xi, \eta) = g(\xi)h(\eta). \tag{1.44}$$

This fact enables us to write, if the abscissas and ordinates of all spot positions are mutually independent, our estimated marginal probability density functions for the second procedure as

$$g_e(\xi) = A_{p\xi}^{-1} \sum_{i=1}^{p} e^{-(\xi-\xi_{oi})^2/2\sigma_g^2}; \quad h_e(\eta) = A_{p\eta}^{-1} \sum_{i=1}^{p} e^{-(\eta-\eta_{oi})^2/2\sigma_g^2}, \tag{1.45}$$

where A_{pi}, included for normalization purposes, represents the projected area of all the superimposed Gaussians onto the plane determined by the i^{th} axis with the axis normal to the separation plane. Now the probability that a single component spot falls within a square of edge length 2δ centered about an arbitrary coordinate (ξ_o, η_o) can be expressed as

$$\int\limits_{\xi_o-\delta}^{\xi_o+\delta} g_e(\xi)\, \mathrm{d}\xi \int\limits_{\eta_o-\delta}^{\eta_o+\delta} h_e(\eta)\, \mathrm{d}\eta. \tag{1.46}$$

This fact serves to reduce a volume integration to two area integrations and hence the frequency must now only be evaluated along the two dimensional axes reducing our computational work load. A frequency so obtained is said to have been estimated by ‘independent means’, otherwise, if the volume integrations were carried out over the entire two-dimensional separation area, the frequency is said to have been estimated by ‘dependent means’. This same idea of independent estimation is also applicable in an analogous fashion when estimating the frequency via the first procedure discussed.

Finally, values must be chosen for the smoothing parameter, σ_g or the standard deviations (common in each dimension) on the bi-Gaussians which is free to vary from separation to separation. Control of this task is turned

Figure 1.8: A typical estimated frequency generated under the second procedure. The contour plot of this same frequency shows the boundaries of one mesh square of edge 2δ.

over to the computer which, through a nested search algorithm, is given instructions on finding that σ_g which yields an absolute minimum for the sum of squares in equation (1.40). Experimentation was also done in allowing σ_g to have independent values for each of the two coordinate dimensions as well as independent values for each spot that was inversely proportional to the local spot density. The accuracy gains were deemed as not worth the additional computational resources needed.

An indeterminate result (0/0) ensues from evaluating the integrand in equation 1.40 when the frequency takes on a zero value. As a more accurate alternative to using an error trap to filter out dips in the density, numerical errors can be eliminated by replacing the integrand with a Maclaurin series expansion for small values of the frequency:

$$\begin{aligned} u(x) &= \frac{mbx^2}{e^{bx}-1}; \quad b = \pi d^2 m; \quad x = f(\xi,\eta), \\ u(x) &= u(0) + mu'(0). \end{aligned} \tag{1.47}$$

The limit of $u(x)$ as $x \to 0$ vanishes with an application of L'Hôpital's rule and after considering the limit of the derivative we arrive at

$$u(x) \approx mf(\xi,\eta), \tag{1.48}$$

which is defined and valid at all values of $f(\xi,\eta)$ too low to evaluate the integrand of equation 1.40 in its original form because of machine epsilon.

1.3 The Code

The main codes presented for this tutorial are `spotone.f` and `spottwo.f`. These codes process the coordinates of multicomponent spot locations, for one and two-dimensional separations respectively, and yield an estimation as to the original number of single components resident, employing the techniques previously discussed. Since the workings of these two codes are fairly similar, the definitions and discussion that follow apply to both unless it is explicitly stated otherwise.

1.3.1 Some Coded Parameters

- `KC`: Array that holds the coordinates of the chromatogram peaks.
- `Q`: The number of peaks in a given chromatogram.
- `CL`: Array that holds the p-plot (p vs. x_o or d_o) coordinates.
- `P`: The total number of spots included in the fit after the plateau and trailing end of the p-plot have been truncated.
- `EE`: Marker for where to truncate the trailing end of the p-plot. `EE` $= 11$ for example means that points $p = 1, \ldots, 10$ will be excluded from equation (1.5,1.40).
- `N`: "Fit degree" parameter. Determines the degree of fit used within a sub-algorithm that calculates the most likely borders of the chromatogram from the peak locations. This is not discussed. A value of `N`=3 is a good choice.
- `FQ`: Array that holds the estimated smoothed frequency information.
- `BIN`: The number of elements in the array `FQ` or the number of sample points of the frequency for minimization of equation (1.5,1.40).
- `DIV`: The number of subdivisions of the individual `BIN` providing the mesh on which the individual Gaussians are integrated to provide the estimated frequencies.
- `TOL`: Parameter that controls the tolerance of the search routines. Appears in the main body, and again independently for the search within the subroutine `MINIMIZER`.

- SSQ: A running tally for the sum of squares (1.5,1.40) which is to be minimized.
- PRD: The prediction of m the number of individual components of the chromatogram.
- ERR: The calculated error on PRD is held within this variable. Refer to equation (1.30).

1.3.2 Design of Flow

Coordinates for peak locations on a chromatogram are read into the array KC from a user specified data file. The number of entries is identified with the number Q. This array is then sent to the subroutine NORMALIZER which sorts the coordinates, calculates likely boundaries for the separation, and then normalizes the coordinates.

The normalized coordinate array is now sent to the subroutine CLUSTER which will calculate the p-plot coordinates and save them to the array CL. This task is accomplished by successively drawing imaginary circles of increasing diameter (x_o or d_o) centered around each peak location and identifying the number of groups of interlocking circles with the parameter p. At a critical radius the number p will drop by one and then remain fixed until the next critical radius is reached; thus forming a series of 'shelves'. The value of x_o or d_o paired with a particular value of p is that of the center of the particular shelf.

Control of the program is now turned over to two nested search routines. One is in the main body of the program and the other resides in the subroutine MINIMIZER. The first searches for the smooothing parameter σ_g used in the construction of the estimated frequency (1.45) that minimizes (1.5,1.40). The next searches for the intermediate minimum sum of squares estimate for $\overline{m}$ for each frequency calculated. In this way we arrive at the lowest global minimum for the sum of squares achievable and this, ultimately, will be our final estimate for $\overline{m}$, which we assosiate with m the original number of individual components on the chromatogram.

1.4 Execution

1.4.1 Compilation

All codes provided were written in the now defunct Microsoft PowerStation Fortran 4.0. Because use is made of dynamic array allocation, the codes must be compiled in fortran 90. Modifications have been made so that they will conform to IBM's XLFortran and will compile issuing the command `xlf90 list.f -o prog`. Here the "o" switch allows for the user to define the name of the executable. For detailed information use the command `man xlf90`.

The codes may also be easily modified to compile with Sun Microsystems `f90` compiler. Only two modifications are necessary. First, wherever there is a continuation character in column six, the "&" at the end of the previous line must be removed. Second, compile with the option `-ext_names=plain` so that the fortran compiler will recognize the c++ externals `seed48` and `drand48`.

1.4.2 Generating artificial chromatograms

Since it is reasonable to suppose that the users of this tutorial will not have access to actual separations, the codes `genone.f` and `gentwo.f` have been included to generate artificial one and two-dimensional chromatograms respectively.

Given as typed input for these programs is a filename read in as a string. This file, in turn, contains the user created filenames, listed one at a time, that will be used to store the coordinates of the final peak positions of each chromatogram generated. In the subroutine GENERATOR are some predescribed marginal frequency choices that can be used or commented out if the user prefers another. Computer generated random numbers will then be mapped to single component locations governed by the chosen distribution. In Fig. 1.9 for example are single component positions for two-dimensional quadratic-linear and quadratic-sine distributions which are provided. For more details on the specifics of mapping, see the sections on "Distribution Functions" and "Dependent Frequencies" below.

Once a set of single component positions are identified, circularly contoured bi-Gaussians with constant standard deviations and amplitudes are centered about these positions. It is clear that some relation between d_o and σ, the standard deviation of a circular bi-Gaussian profile, is needed. To this end an argument is presented

Figure 1.9: On the top is a frequency that is linear across one dimension and quadratic across the other, and the resulting single component positions. The bottom frequency results when the linear portion from above is replaced by a sine function.

which is well known within the separations community. We take a superposition of two such profiles whose centers are separated from one another by an arbitrary distance of twice k standard deviations:

$$\frac{1}{2\pi\sigma^2}\left[e^{-(r-k\sigma)^2/2\sigma^2}+e^{-(r+k\sigma)^2/2\sigma^2}\right]. \tag{1.49}$$

Here the radial coordinate, r, is measured on an axis that joins the two centers. The derivative of this expression is set to zero yielding three critical points, neglecting $\pm\infty$. They represent the apex of each bi-Gaussian and the local minimum that exists at the midpoint between them. As k is reduced the critical points will approach each other until an essential value of $k = \sigma$ is reached where, and below which, the three critical points become degenerate. Thus we may say that only when two circular bi-Gaussians, of equal amplitude and variance, reside further apart than twice their mutual standard deviation are they distinguishable from one other in the sense that there will be a dual apex. This same result is valid for one-dimensional contours allowing us to write $x_o = d_o = 2\sigma$, visualized in Fig. 1.10, which provides the means to tailor our set of m single component profiles to a desired working saturation for one and two-dimensional separations of extent X and A:

$$\begin{aligned}\sigma &= \frac{\alpha X}{2m}\\ \sigma &= \sqrt{\frac{\alpha A}{\pi m}}.\end{aligned} \tag{1.50}$$

In actuality the ratio of d_o/σ is known to vary slightly as a function of the ratio of amplitudes [12], but we will assume a constant value of 2 for the provided codes for brevity.

The number of grid elements that is input describes the mesh on which data is held. This number should be large enough to guarantee smoothness but not so large as to burden computational resources. Maybe 10,000 for the one-dimensional and a few hundred for the two-dimensional case. Saturation will generally range between 0 and 1. Chose a saturation and compare how many peak locations are found in relation to how many single components were specified. As a rule of thumb for gauging component overlap, a saturation of 2 will generally translate into a p/m ratio of approximately 0.13 for a one-dimensional separation of constant density, and less otherwise. Modify the code to dump the chromatogram stored in memory to a file for visual inspection.

The search algorithm for the peak coordinates, that are stored as output in the specified filenames, is a simple nearest neighbor search. For these artificial chromatograms nothing more complicated is needed do to the the absence of noise that would be present for an actual separation. Real chromatograms then require a more complicated algorithm for seeking out peak locations that involve taking numerical derivatives as well as additional tools.

1.4.3 Input File

The files read in as input by `spotone.f` and `spottwo.f` are the very same files that were created for `genone.f` and `gentwo.f` in the "Generating Artificial Chromatograms" section. The first file read in is the list of names of those files that now contain the peak coordinates of the computer generated chromatograms. These files are subsequently read in one at a time and processed.

1.4.4 Interpretation of Results

Output files for the one and two-dimensional codes are named `outone` and `outtwo` respectively. Blocks are amended to the files as each chromatogram is processed. The first line of each block consists of four entries: σ_g (smooting factor), $\overline{m}$ (single component prediction), `SSQ` (equations 1.5 and 1.40), and `Q` (number of observable peaks). For `outone` only, there are two additional lines per block. The middle two entries of the second line give the coordinates of the first and last peak. The outer two give the calculated values for the most likely boundaries of the separation that were used for processing. Remember that the actual values for the boundaries are 0 and the value entered for `BIN`, but for a real chromatogram, where only the peak locations are given, the code has no information on what the boundaries should be. Finally, the last line gives the error on the prediction as the square root of the variance (1.27).

It is curious to note that the output files are continuously opened and closed immediately prior to and after the write statements. This was done in order to force output from the buffer to the files so that data was not

Figure 1.10: Fusion of similar Gaussians as a function of center separation.

lost during the frequent power outages and thunder storms that plagued southern Illinois where the codes were developed.

1.5 Exploration

1.5.1 Generation of a Thinned Poisson Process of varying density

To test the validity of equation (1.20) through computer simulation it becomes necessary to be able to generate a Poisson point process of variable density. To accomplish this we may turn to the definition [9] of an approximate Poison process of density λ which stipulates that:

- The number of changes occurring in nonoverlaping intervals are independent.
- The probability of exactly one change in a sufficiently short interval of length h is approximately λh.
- The probability of two or more changes in a sufficiently short interval is essentially zero.

Let frequencies span a unit domain [0,1], and take sufficiently short to be h=0.00005 for those frequencies that have associated densities, $\lambda(x) = \overline{m}f(x)$, that are either rapidly changing or exceeded 100 at some point on there domain. All others processes may be calculated on a h=0.0001 mesh. Using excessively high freqencies is largely redundant as this effect can be duplicated by merely rescaling the resolution parameter x_o.

For the center of each mesh interval in the domain, generate a uniformly random number R on the range [0,1) and use the chosen function, $\lambda(x)$, to calculate a density constant, λ, there. If then, for any mesh interval, $R < \lambda h$, assume that an event resides at its midpoint as illustrated in Fig. 1.11. For densities that do not fall to zero at the boundaries of the domain, [0,1], the generation process should be extended periodically (mirrored) to a domain of [-.5, 1.5]. The reason for this mirroring will become evident from the discussion below.

Given a group of n events constituting a cluster, let the representation of that cluster be a point placed at the position

$$x = \frac{1}{n}\sum_{k=1}^{n} x_k. \tag{1.51}$$

The set of all so placed clusters becomes the thinned Poison process of variable density. The interval extension, mentioned above, is done to make these distributions appear as if they had been sampled from larger continuous distributions. In so doing one does not get the biased build-up of clusters at the ends of the intervals due to an imposed artificial distance of greater than x_o that comes about from ending the process there. To clarify let us consider a constant density of infinite extent. If we generate the distribution only on the domain of interest, say [0,1], we have then clipped the original density function at $x = 0$ and $x = 1$. In so doing the leftmost point, for example, will belong to a cluster that always falls inside the domain and thus increments the count for p. If, on the other hand, we had generated the complete infinite distribution, then that same leftmost point just inside the domain of interest may be a member of a cluster whose weighted position, described by equation (1.51), falls to the left of zero because of exterior points. As a result, the count for p is not incremented in this situation. Fig. 1.12 shows this for a half interval extension. Even having considered this issue, it will still contribute to an increasing percent error as the resolution parameter grows and the number of clusters drops. Of course the overlap can be extended, but at the trade off of computational time.

Generate an ensemble, of say 50,000 members, of such thinned Poisson processes using your favorite frequency function. Find the variance of the observed number of clusters (The section on "Cluster Search" may help with this) and compare with equation (1.20). Rigorous testing results on this equation may be obtained from [13].

1.5.2 Distribution Functions

Distribution or frequency functions are those functions that govern the placement, and in turn the density, of the single component profiles of a separation. Let us say that we are interested in laying down single component profiles on a one-dimensional chromatogram in a quadratic fashion over the range of $0 \leq x < 1$. First we must

Figure 1.11: A schematic for the generation of a Poisson point process of variable density, given a density $\lambda(x)$.

Figure 1.12: The bottom diagram initially shows a sequence of random events on an unextended interval. The same sequence of events is then shown with an extension to the left. The weighted cluster positions, given by equation (1.51), are shown. We see here that when the extension is considered the weighted cluster position falls outside the original interval, changing the cluster count.

ensure normalization:

$$
\begin{aligned}
1 &= k\int_0^1 x^2\,\mathrm{d}x, \\
k &= 3, \\
f(x) &= 3x^2; \quad 0 \le x < 1.
\end{aligned} \tag{1.52}
$$

Next we may define a cumulative distribution function as the integral of the frequency function:

$$F(x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x = \int_0^x f(x)\,\mathrm{d}x. \tag{1.53}$$

In this way the uniformly random (on interval [0,1)) computer generated deviates, R, may be equated to the cumulative distribution function which fixes the independent variable as the mapped deviates governed by the new quadratic frequency:

$$R_i = \int_0^{x_i} f(x)\,\mathrm{d}x = x_i^3. \tag{1.54}$$

In words one would set the number R as the ordinate, come horizontally right to the cumulative distribution curve, and then straight down to read x off as the abscissa:

$$x_i = \sqrt[3]{R_i}. \tag{1.55}$$

It will not always be possible to solve equations like (1.55) for x_i using a programing language's intrinsic functions. For this case one would have to resort to finding the root of

$$R_i - x_i^3 = 0 \tag{1.56}$$

using bisection, Newton's method, or some other iterative technique.

Try generating some unique frequencies on your own. How would you generate these unique frequency deviates on an interval of say $-3 \le x < 5$?

1.5.3 Dependent Frequencies

The codes provided, `gentwo.f` and `spottwo.f`, are of the simpler form that make use of a joint probability density function, or frequency as we have called it, that can be written as the product of its marginal functions:

$$f(\xi,\eta) = \int_{-\infty}^{\infty} f(\xi,\eta)\,\mathrm{d}\eta \int_{-\infty}^{\infty} f(\xi,\eta)\,\mathrm{d}\xi. \tag{1.57}$$

Artificial chromatograms may be generated by mapping each pair (R_i and R_j) of computer generated uniformly random deviates on the interval [0,1) into the frequency of our choosing:

$$R_i = \int_{-\infty}^{\xi_i} f(\xi,\eta)\,\mathrm{d}\xi; \quad R_j = \int_{-\infty}^{\eta_j} f(\xi,\eta)\,\mathrm{d}\eta. \tag{1.58}$$

Convince yourself that for the correlated frequency

$$f(\xi,\eta) = \begin{cases} 0: & (0 \le \xi < .5) \cap (.5 \le \eta < 1) \quad \text{or} \quad (.5 \le \xi < 1) \cap (0 \le \eta < .5) \\ 2: & (0 \le \xi < .5) \cap (0 \le \xi < .5) \quad \text{or} \quad (.5 \le \xi < 1) \cap (.5 \le \eta < 1), \end{cases} \tag{1.59}$$

the above two equations no longer hold. Design changes and modify `gentwo.f` and `spottwo.f` so that they are able to deal with these types of correlated frequencies. A 'shortcut' is still possible for generating chromatograms where again two, as opposed to one, uniformly random deviates may be mapped into a single point governed by the characteristic frequency. To do this, one marginal frequency and all the columns perpendicular to it must be independently normalized. Explain how this trick might work and how it speeds up the computation. Fig. 1.13 shows the minimum sum of squares estimated frequency of a chromatogram generated with the above distribution function. The top is the result of having made the modification and the bottom shows how `spottwo.f`, in its simpler form, cannot deal with such a distribution. The simulated seperation itself, as well as its accompanying component locations, can be seen in Fig. 1.14.

Figure 1.13: The top shows the optimal estimated frequency as predicted by dependent means for the simulated separation shown in Fig. 1.14. The bottom shows the resulting least squares frequency when forcing an independent fit to the dependent simulation.

Figure 1.14: A distribution of single component locations along with a simulated separation generated from equation (1.59).

1.5.4 Cluster Search

Calculating 'p-plot' coordinates for the one-dimensional case is trivial since each interval greater than the resolution parameter generates an additional cluster. The coding is not so effortless in multi-dimensions, however. In two dimensions we have interlocking rings, in three dimensions, spheres, etc. In each case we must be careful to count each cluster only once and not once for each member.

Take an computer generated two-dimensional chromatogram and apply a self written routine to calculate p_i given x_{oi}. Add lines to `spottwo.f` to output the array `CL` for this same chromatogram and compare the results. Generalize your algorithm to higher dimensional space.

1.5.5 Bayesian Estimation

Via repeated application of equation (1.34) generate the discrete conditional probability distribution for $0 \leq m \leq 200$ componenets given $p = 15$ peaks have occurred. Take $x_o = 0.02$. Ordinarily the two maxima within this type of discrete distribution would reside further apart. Utilize the following Mathematica cell to perform this task and explore the effect of varying x_o and p.

```
Bay={}
x=1/50
p=15
Do[{For[s=0;z=0,(p+s)*x<1,s++,z=z+((-1)^s*(1-(p+s)*x)^(m-1))/(s!*(m-p-s)!)]
,Bay=Append[Bay,N[(z*m!)/p!,12]]},{m,1,200}]
ListPlot[Bay,AxesOrigin->{0,0},PlotRange->{{0,200},{0,Max[Bay]}},AxesLabel->{"m","P(m)"}]
```

Correct execution of this command will reproduce Fig. 1.15.

Figure 1.15: Results from repeated application of equation (1.34) giving the conditional probability distribution for m given that p visible peaks have occurred. For this figure $p = 15$ and $x_o = 0.02$. In most cases the two peaks will reside further apart.

Bibliography

[1] M. Martin and D. P. Guiochon, Anal. Chem. 58, 2200, (1985).

[2] J. M. Davis, Anal. Chem. 66, 735 (1994).

[3] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions* Chapman and Hall, London, (1981).

[4] K. D. Rowe and J. M. Davis, Anal. Chem. 67, 2981 (1995).

[5] J. M. Davis, Anal. Chem. 63, 2141 (1991).

[6] Sheldon M. Ross, *Stochastic Processes* John Wiley and Sons, New York, (1983).

[7] M. G. Kendall and P. A. P. Morgan, *Geometrical Probability* Charles Griffin, London, (1963).

[8] P. R. Bevington, *Data Reduction and Error Analysis for the Physicsl Sciences* McGraw-Hill, New York, (1969).

[9] Robert V. Hogg and Elliot A. Tanis, *Probability and Statistical Inference* Macmillan, New York, (1988).

[10] W. Shi and J. M. Davis, Anal. Chem. 65, 482, (1993).

[11] S. A. Roach, *Theory of Random Clumping* Methuen, London, (1968).

[12] J. Calvin Giddings, *Unified Separation Science* John Wiley & Sons, New York, (1991).

[13] K. D. Rowe and J. M. Davis, *Error Analysis of Parameters Determined with Statistical Models of Overlap from Nonhomogeneous Seperations* Chemometrices and Intelligent Laboratory Systems. 38, 123 (1997).