# Residual based localisation and quantification of peaks in X-ray diffractograms

P. L. Davies*
Universität Duisburg-Essen
Technical University Eindhoven

M. Meise*
Universität Duisburg-Essen

D. Mergel
Universität Duisburg-Essen

T. Mildenberger*
Universität Dortmund

**Abstract**

We consider data consisting of photon counts of diffracted X-rays as a function of the angle of diffraction. The problem is to determine the positions, powers and shapes of the relevant peaks. An additional difficulty is that the power of the peaks is to be measured from a baseline which itself must be identified. Most methods of de-noising data of this kind do not explicitly take into account the modality of the final estimate. The procedure we propose is based on the so called taut string method which minimizes the number of peaks subject to a tube constraint on the integrated data. The baseline is identified by combining the result of the taut string with an estimate of the first derivative of the baseline obtained using a weighted smoothing spline. Finally each individual peak is expressed as the finite sum of kernels chosen from a parametric family.

## 1 The data

X-ray diffraction is an important tool to analyze the morphology of thin films. When thin films are prepared on glass substrates they are usually polycrystalline and may even contain different crystalline phases. The experimental data are usually obtained in the form: Intensity versus diffraction angle $2\theta$. The physically relevant information lies in the position, the power, and the half-width of the peaks.
The peak positions are characteristic for the crystalline structures present in the sample. Small shifts of the peaks with respect to the ideal positions are related to mechanical strain in the crystalline lattice arising from lattice imperfections introduced during thin film preparation.
From the peak power the relative abundance of a specific crystalline orientation can be estimated allowing the determination of the texture of crystalline orientations. Such an analysis has been performed e.g. in the case of thin films of $In_2O_3$:Sn prepared by various deposition techniques (Mergel et. al, 2005).

*Research supported in part by Sonderforschungsbereich 475, University of Dortmund

The half-width of the peak is related to the crystallite size and to inhomogeneous strain within the crystallites. These parameters are strongly influenced by the preparation conditions and determine to a large degree the optical and electrical properties of the thin films.
In our laboratory practice, we have so far used an ad-hoc method to evaluate the X-ray diffractograms that proved to be adequate when the potential peak positions were a-priori known, i.e. in cases where the produced material was already identified (Mergel et. al, 2005). With this method, the baseline of the data, arising from the noise level of the signal channel, was taken as a piecewise linear interpolation between the intensity values at positions in the middle between two neighbouring theoretical positions. Denoising was done by averaging the data in a pre-defined abszissa interval. The peak position was then looked for in the vicinity of the theoretical positions and the shape of the peak was fitted with a Gaussian.
This method uses optimization criteria for noise that are statistically not well founded and shape functions that are often inadequate for X-ray peaks. Furthermore, in the general case, the crystalline structures in the films are not known and the search procedure as a whole is not applicable. Therefore, we looked for an alternative model that does not rely on a-priori knowledge of peak positions, applies a statistically better founded noise model and a more general function for the peak shape.
The new method presented in this paper is based upon five steps:

1. The data are approximated by the right-hand derivative of the taut string. This yields a first estimate of the number, the position and the height of the peaks.

2. An estimate of the first derivative is obtained from a weighted smoothing spline fitted to the original data.

3. The peak intervals are determined from (a) the positions of peaks according to (1) and (b) a threshold for the derivative obtained in (2).

4. The baseline is obtained by fitting a spline to the remaining data set after removing the peak intervals.

5. The peaks (baseline subtracted) are fitted within their respective intervals by a sum of Pearson Type VII curves. A sum is necessary because the actual peak may be the result of different overlapping X-ray reflexes (subpeaks). The subpeaks may have different widths. In many cases, different solutions may explain the experimental data equally well. Then, alternative solutions have to be delivered as a result and the user has to choose the best one based on additional physical knowledge.

A typical data set is shown in Figure 1. Although it is not obvious from the figure, the data, being counts of photons, are integers. A simple stochastic model for the counts at an angle of diffraction $2\theta$ is the Poisson distribution with mean $f(2\theta)$ for an appropriate function $f$. The noise present in the data does not exhibit any obvious dependencies so that a model is completely specified by fixing $f$ and then taking the observations at each $2\theta$ to be independently distributed. Throughout this paper, we use some different notations in place of $2\theta$, depending on which is most convenient.

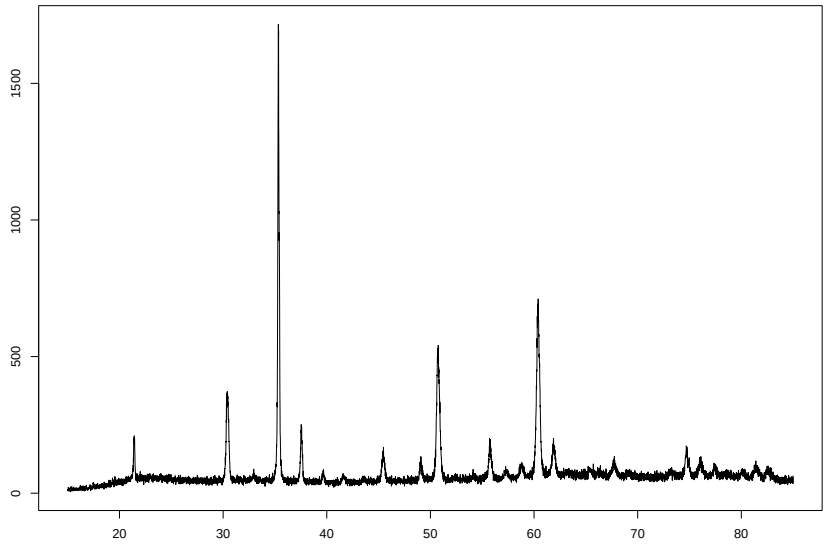


Figure 1: The intensity of diffracted X-rays as a function of the angle of diffraction ($2\theta$).

Since the angles themselves are only of interest when either deciding whether some values are physically meaningful or when calculating physical parameters after the curve has been fitted, we usually use generic indices $t_i$, $i = 1, \ldots, n$ with $t_i \in [0, 1]$ when describing the method from a more mathematical point of view. Since in thin-film diffractometry measurements are usually taken for equidistant angles, we will keep notation simple by considering only the equidistant case. Of course, with appropriate modifications, the method should also work for non-equispaced measurements, if needed.

All estimates $f_n$ of the function $f$ given in this paper are based on the idea of obtaining a simplest function which could in principle have generated the data. This concept is described in detail in section 2. It is based on the considerations in Davies (1995) and Davies, Kovac and Meise (2007) . The procedure we propose determines the number, positions, powers and shapes of the relevant peaks and their components in a fully automatic manner. As mentioned above it consists of five different steps. In sections 3 and 4 we explain respectively the taut string, which was established in Davies and Kovac (2001), and the weighted smoothing spline methods. For more details about weighted smoothing splines see Davies and Meise (2005). An informative introduction to smoothing splines in general was for example given by Green and Silverman (1994). In section 5 the results of the two methods are combined to isolate the peaks and to provide an estimate of the baseline. Finally in section 6 we describe how the individual peaks are expressed as a finite sum of kernels chosen from a parametric family of kernels.

## 2 The confidence region

The canonical model for the data $\mathbf{y}_n = \{(t_i, y(t_i)), i = 1, \ldots, n\}$ is the Poisson model with mean $f(t_i)$. For large $y(t_i)$ this accurately describes the noise level but for small $y(t_i)$ this model underestimates the noise level as there is also a ground noise due to the electronics. In practice the standard deviation of the noise is at least 7 which in the Poisson model corresponds to a mean of about 50. For such large parameter values the Poisson distribution can be modelled by a normal distribution with mean and variance 50. This leads to the model

$$Y(t) = f(t) + \sigma(t)Z(t), \qquad 0 \le t \le 1 \tag{1}$$

where $f : [0, 1] \to \mathbb{R}$ and $Z(t)$ is standard Gaussian white noise. Initially we consider a constant noise level $\sigma(t) = \sigma_n$ which is estimated from the data. This gives us an initial estimate $f_n$ of $f$ and in a second stage we put $\sigma(t) = \max(\sqrt{f_n(t)}, \sigma_n)$ and re-estimate $f$. This removes unwanted side lobes on the peaks due to the initial underestimate of the noise level by $\sigma_n$.
We now explain the construction of the confidence region which provides the basis of our concept of approximation. Suppose we have data $\mathbf{Y}_n = \{(t_i, Y(t_i)), i = 1, \ldots, n\}$ with $0 \le t_1 < \cdots < t_n \le 1$ which are generated under the model (1). For any function $g : [0, 1] \to \mathbb{R}$ we define the residuals by

$$r(\mathbf{Y}_n, t_i, g) = Y(t_i) - g(t_i) \tag{2}$$

and the standardized sums of the residuals over intervals $I \subset \{1, \ldots, n\}$ by

$$w(\mathbf{Y}_n, I, g) = \frac{1}{\sqrt{|I|}} \sum_{i \in I} r(\mathbf{Y}_n, t_i, g) \tag{3}$$

where $|I|$ denotes the number of points in $I$. For a given family $\mathcal{I}_n$ of intervals of $\{1, \ldots, n\}$ an $\alpha$-confidence region for $f$ is given by

$$\mathcal{A}_n = \mathcal{A}(\mathbf{Y}_n, \sigma, \mathcal{I}_n, \tau_n) = \Big\{ g : \max_{I \in \mathcal{I}_n} |w(\mathbf{Y}_n, I, g)| \le \sigma \sqrt{\tau_n \log n} \Big\}, \tag{4}$$

where $\tau_n = \tau_n(\alpha)$ is chosen such that

$$P\Big( \max_{I \in \mathcal{I}_n} \frac{1}{\sqrt{|I|}} \Big| \sum_{t_i \in I} Z(t_i) \Big| \le \sigma \sqrt{\tau_n \log n} \Big) = \alpha. \tag{5}$$

To see this we note that if the data were generated under (1) then (5) implies that $P(f \in \mathcal{A}_n) = \alpha$. The family $\mathcal{I}_n$ we use will be a dyadic multiresolution scheme as for wavelets. It will consist of all single points $\{i\}$, the pairs $\{1, 2\}, \{3, 4\}, \ldots$, the sets of four $\{1, 2, 3, 4\}, \{5, 6, 7, 8\}$ etc. and including all final intervals whether or not they are of this form. The procedure is therefore not restricted to sample sizes $n$ which are a power of 2. The number of such intervals is at most $2n$ and this collection has proved sufficiently fine for X-ray diffractograms. The use of such a scheme $\mathcal{I}_n$ forces any function $g$ in $\mathcal{A}_n$ to adapt to the data at all resolution levels from single points

to the whole interval. Since the noise level $\sigma$ of the data usually is not known in advance we derive it from the data by using

$$\sigma_n = 1.4826\,\text{Median}\Big\{|Y(t_i) - Y(t_{i-1})|, 2 \le i \le n-1\Big\}/\sqrt{2}. \tag{6}$$

Now $\mathcal{A}_n = \mathcal{A}(\mathbf{Y}_n, \sigma_n, \mathcal{I}_n, \tau_n)$ is no longer exact but it is honest in that the coverage probability is now at least $\alpha$. Any function $g_n \in \mathcal{A}(\mathbf{Y}_n, \sigma_n, \mathcal{I}_n, \tau_n)$ will be regarded as an adequate approximation to the data $\mathbf{Y}_n$. Depending on the context in the following we also use $\mathcal{A}_n = \mathcal{A}(\mathbf{y}_n, \sigma_n, \mathcal{I}_n, \tau_n)$ to specify the analogue confidence region for the data $\mathbf{y}_n$ of an X-ray diffractogram.
As mentioned above the estimate $\sigma_n$ underestimates the noise level for large values of $y(t_i)$ which is of the order $\sqrt{y(t_i)}$. At the same time for small values of $y(t_i)$ the noise is underestimated by $\sqrt{y(t_i)}$ but correctly estimated by $\sigma_n$. We overcome both problems by obtaining a first estimate $f_n$ of $f$ using the constant noise $\sigma_n$ of (6) and then taking the noise level at the angle of diffraction $t_i$ to be

$$\Sigma_n(t_i) = \max\left(\sigma_n, \sqrt{f_n(t_i)}\right). \tag{7}$$

Now we replace (2) by

$$\tilde{r}(\mathbf{y}_n, t_i, g, \mathbf{\Sigma}_n) = \frac{y(t_i) - g(t_i)}{\Sigma_n(t_i)} \tag{8}$$

and (3) by

$$\tilde{w}(\mathbf{y}_n, I, g, \mathbf{\Sigma}_n) = \frac{1}{\sqrt{|I|}} \sum_{i \in I} \tilde{r}(\mathbf{y}_n, t_i, g, \mathbf{\Sigma}_n). \tag{9}$$

The resulting confidence region $\tilde{\mathcal{A}}_n$ is then given by

$$\tilde{\mathcal{A}}_n = \tilde{\mathcal{A}}(\mathbf{y}_n, \mathbf{\Sigma}_n, \mathcal{I}_n, \tau_n) = \Big\{g : \max_{I \in \mathcal{I}_n} |\tilde{w}(\mathbf{y}_n, I, g, \mathbf{\Sigma}_n)| \le \sqrt{\tau_n \log n}\Big\}. \tag{10}$$

The value of $\tau_n$ for any $n$ and $\alpha$ can always be determined by simulations. It follows however from a result of Dümbgen and Spokoiny (2001) on the uniform modulus of continuity of the Brownian motion that $\lim_{n\to\infty} \tau_n = 2$ whatever $\alpha$. In practice we use the default value $\tau_n = 2.5$ The confidence regions $\mathcal{A}_n$ and $\tilde{\mathcal{A}}_n$ include many functions which are of no interest. For example all functions $g$ which interpolate the data belong to both. Interest always centres on the simplest functions where the definition of simplicity depends on the problem at hand. To detect peaks we are interested in minimizing the number of peaks subject to the function lying in $\mathcal{A}_n$ or $\tilde{\mathcal{A}}_n$. We accomplish this by using the taut string method which is described in the next section. The taut string estimate is a piecewise constant function which is not suitable for identifying the baseline. The baseline is a slowly varying function which can be associated with a small first derivative. The second concept of simplicity we use is therefore based on smoothness and is defined by

$$\int_0^1 g''(t)^2\, dt. \tag{11}$$

To minimize (11) we use an approximate procedure based on a weighted smoothing spline. The solution is a cubic spline and we use its first derivative to identify the baseline.

## 3 The taut string method

In this section we give a short description of the taut string method based on a small artificial data set. Panel 1 of Figure 2 shows data generated under (1) with $f(t) = 2.5\sin(4\pi t)$ evaluated at the points $t_i = i/32, i = 1, \ldots, 32$ and with $\sigma = 1$. The first step is to calculate the partial sums of the observations $Y(t_i)$

$$S_Y(t_i) = \frac{1}{n}\sum_{j=1}^{i} Y(t_i),\ i = 1, \ldots, n, \quad S_Y(0) = 0. \tag{12}$$

These are shown in Panel 2 of Figure 2. We now form a tube centered on the cumulative sums with an upper bound $U$ and a lower bound $L$ defined by

$$\begin{aligned} U(t_i) &= S_Y(t_i) + \epsilon,\ 1 \le i \le n-1, \quad U(0) = 0,\ U(1) = S_Y(1) && (13)\\ L(t_i) &= S_Y(t_i) - \epsilon,\ 1 \le i \le n-1, \quad L(0) = 0,\ L(1) = S_Y(1). && (14) \end{aligned}$$

The boundary conditions $U(0) = L(0) = 0$ and $U(1) = L(1) = S_Y(1)$ are chosen to reduce edge effects. The resulting tube is shown in Panel 3 of Figure 2. The taut string function $TS$ is best understood by imagining a string constrained to lie within the tube and tied down at $(0, 0)$ and $(1, S_Y(1))$ which is then pulled until it is taut (cf. Panel 4 of Figure 2). There are several equivalent analytic ways of defining this. The taut string is a linear spline with automatic choice of knots. Panel 5 of Figure 2 shows the knot locations. As an estimate $f_{ts,n}$ of $f$ we take the right derivative of the taut string, except at the last point where we take the left derivative. Closer consideration shows that this can be improved. The derivative of the taut string has a local maximum when the taut string switches from the upper to the lower boundary. The value of the derivative on this section is therefore less than the mean of the $Y$-values. Thus if we define the estimate at cross-over intervals as the mean of the $Y$-values between the knots we obtain a better approximation without altering the number of local extremes. The same reasoning applies to local minima. The function $f_{ts,n}$ obtained in this manner is shown in Panel 6 of Figure 2. The connection with the number of local extremes which explains the efficacy of the method is the following. Consider all absolutely continuous functions $H$ which are constrained to lie within the tube. Then the derivative of the taut string $f_{ts,n}$ has the smallest number of local extreme values and in particular it has the smallest number of peaks.

This still leaves open the question of the diameter of the tube. Since $\epsilon$ controls the closeness to the data the basic idea is to start with a very large $\epsilon$ which contains the integral of the mean of the data which, in this case, is the taut string solution $f^1_{ts,n}$. Using the confidence region $\mathcal{A}_n$ we determine those intervals $I \in \mathcal{I}_n$ for which $|w(\mathbf{Y}_n, I, f^1_{ts,n})| > \sigma_n\sqrt{\tau_n \log n}$. For any point $i$ which lies in any such interval we reduce the diameter of the taut string by a fixed factor $q < 1$ at the points $i$ and $i + 1$. The default value of $q$ which we use is 0.9. A new taut string estimate $f^2_{ts,n}$ is calculated and the procedure repeated in the obvious manner until the estimate lies in $\mathcal{A}_n$. As $\sigma_n$ is specified by (6), $\tau_n = 2.5$ and $\mathcal{I}_n$ is the dyadic multiresolution scheme defined above. The method is fully automatic and does not require the choice of a tuning parameter. As $\mathcal{I}_n$ contains at most $2n$ intervals and the taut

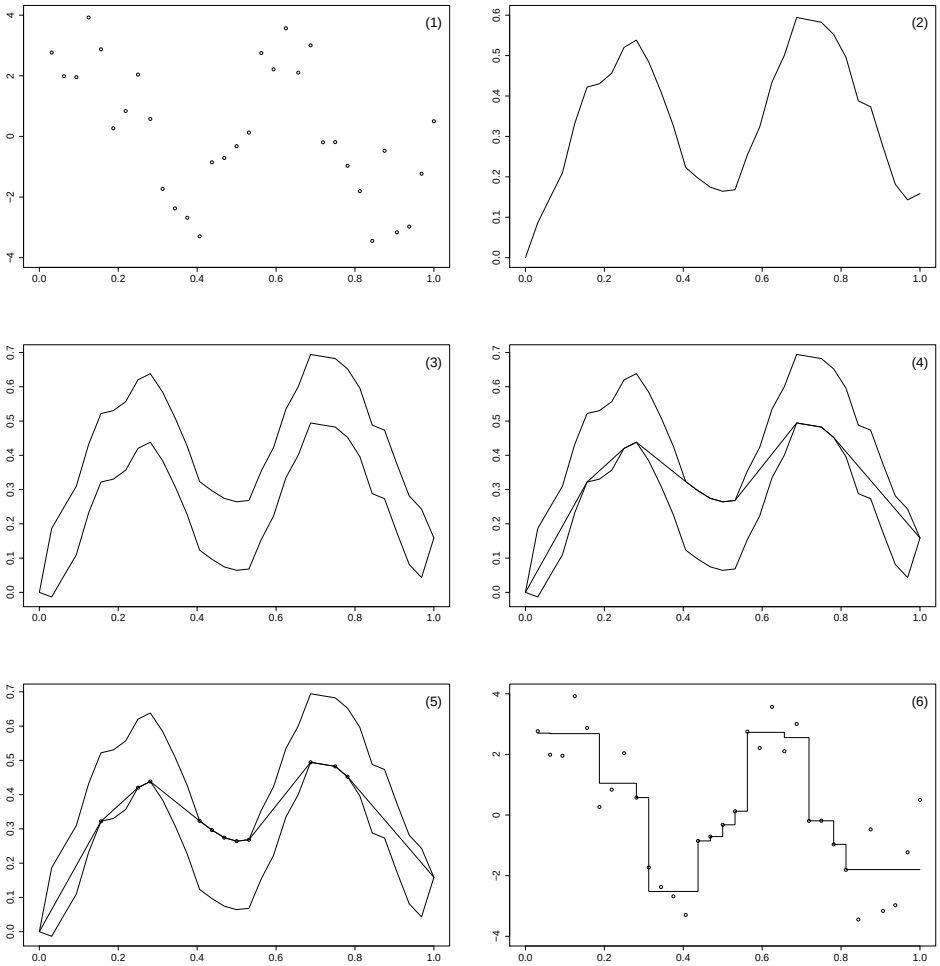


Figure 2: Panel (1) shows some noisy sine data. Panel (2) shows the cumulative sums of the data. Panel (3) shows the tube derived from the cumulative sums. Panel (4) shows the taut string through the tube. Panel (5) shows the taut string through the tube with marked knots. Panel (6) shows the data with the right-hand derivative of the taut string.

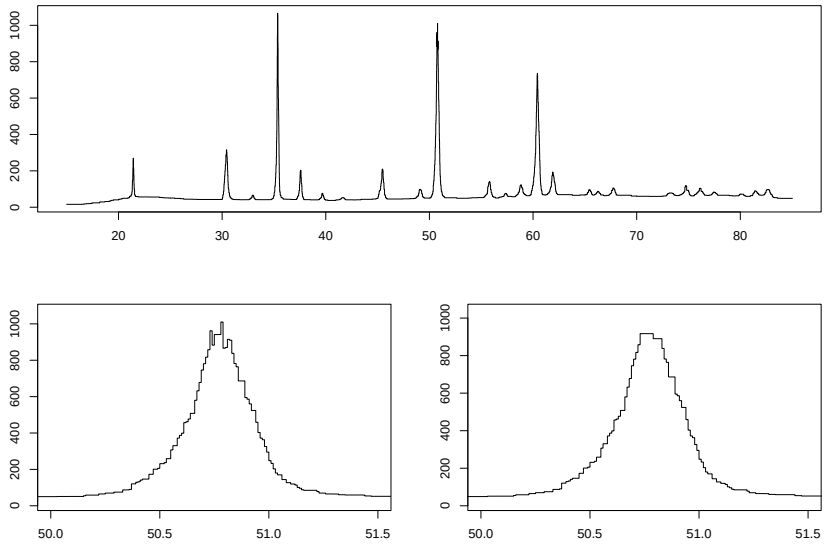


Figure 3: Top: One de-noised X-ray diffractogram, using the constant noise estimate given in (6) and a section (bottom left). Bottom right: The same section for the approximation using the local noise estimate given in (7).

string has an algorithmic complexity of $O(n)$ it follows that the whole procedure has an algorithmic complexity of order $O(n \log n)$ when the squeezing of the tube is taken into account. Large data sets with $n = 10^6$ and more can be calculated processed in less than one minute.
Panel 1 of Figure 3 shows the result of applying this procedure to an X-ray diffractogram with $\sigma_n$ given by (6). As mentioned above this underestimates the noise level for large values of $y(t)$ and this results in side lobes on the large peaks as shown in Panel 2 of Figure 3. We denote this initial estimate by $\tilde{f}_{ts,n}$ and use it in the definition of $\mathbf{\Sigma}_n$ of (7). This gives rise to the confidence interval $\tilde{\mathcal{A}}_n$ of (10) and we can now repeat the taut string procedure. The result is denoted by $f^\star_{ts,n}$. Figure 4 shows $f^\star_{ts,n}$ for the data set of Figure 1. As it can be seen in Panel 3 of Figure 3, which shows $f^\star_{ts,n}$ of the same section as $\tilde{f}_{ts,n}$ in Panel 2, the side lobes have been removed whilst leaving the rest of the initial estimate unaltered. It is clear that the automatic taut string method as just described has produced very good resolution of the peaks and it has not created peaks where none should be.

# 4 Weighted Smoothing Splines

After having determined the number and locations of the peaks the next step is to identify the baseline. We do this by fitting a smooth function to the data and then identifying the baseline by the size of the first derivative. As mentioned above ideally we would like to minimize (11) subject to $g \in \tilde{\mathcal{A}}_n$, using $f^\star_{ts,n}$ in (7). As $\tilde{\mathcal{A}}_n$ is defined

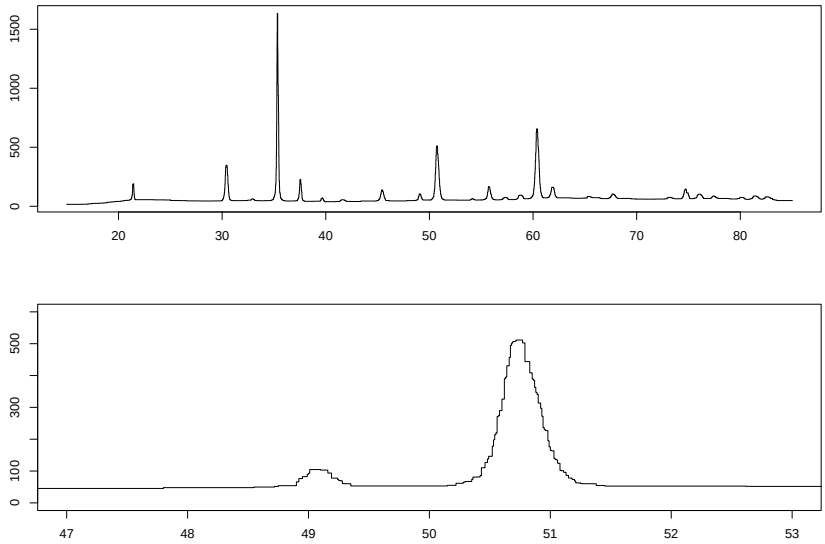


Figure 4: The de-noised data of Figure 1.

by a series of linear inequalities this, after discretization, leads to a quadratic programming problem. This is in principle solvable but for large data sets and/or data with large variations in local smoothness there are considerable numerical problems. Because of this we take an approach based on weighted smoothing splines which is as follows. For given weights $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ we consider the solution of the following minimization problem

$$S_{\boldsymbol{\lambda}}(g) := \sum_{i=1}^{n} \lambda_i (Y(t_i) - g(t_i))^2 + \int_0^1 \left( g^{(2)}(t) \right)^2 \, dt \longrightarrow \text{min!} \tag{15}$$

The solution is a natural cubic spline which we denote by $f_{wss,n}$. The weights $\boldsymbol{\lambda}$ are data dependent and chosen to ensure that $f_{wss,n} \in \tilde{\mathcal{A}}_n$. As the smoothness of the solution of (15) increases when the values of the $\lambda_i$ decrease we wish to choose the weights $\boldsymbol{\lambda}$ to be as small as possible subject to $f_{wss,n} \in \tilde{\mathcal{A}}_n$. We do this in a manner similar to that used in the taut string procedure. We start with very small weights $\lambda_i$ so that the solution is almost a straight line which we denote by $f^1_{wss,n}$. We determine those points $t_i$ which lie in intervals $I$ for which $|\tilde{w}(\mathbf{y}_n, I, f^1_{wss,n}, \boldsymbol{\Sigma}_n)| > \sqrt{\tau_n \log n}$. At such points we increase the $\lambda_i$ by a factor of $q > 1$. The default value we use is $q = 2$. The solution $f^2_{wss,n}$ is calculated and the procedure is continued in the obvious manner until the solution lies in $\tilde{\mathcal{A}}_n$. The first Panel of Figure 5 shows the result of the weighted smoothing spline, $f_{wss,n}$, for the data set of Figure 1. The second Panel shows the first derivative $f^{(1)}_{wss,n}$. The smoothness of the solution can be seen from the third panel of Figure 5 which shows the same section of the data as Figure 4.

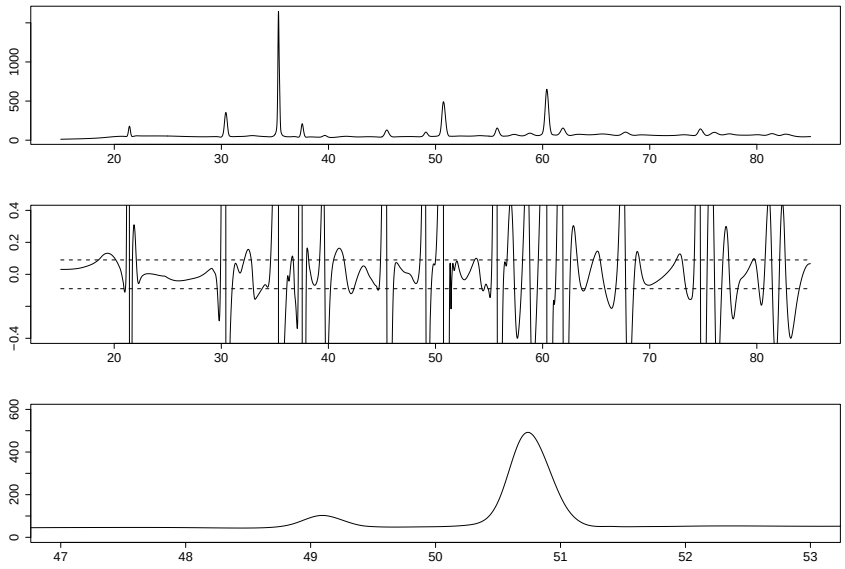


Figure 5: The upper Panel shows the weighted smoothing spline estimate $f_{wss}$ of the data of Figure 1. The middle Panel shows the first derivative $f^{(1)}_{wss}$ together with the used threshold (dotted line) and the bottom Panel shows a section of $f_{wss}$.

# 5 Identifying the baseline

To identify the baseline we combine the results of the taut string, $f^{\star}_{ts,n}$, and the weighted smoothing spline approximation, $f_{wss,n}$. The baseline is a slowly varying function so we identify it by the size of the derivative $f^{(1)}_{wss,n}$ of $f_{wss,n}$. The taut string estimate is piecewise constant so firstly we identify those intervals which correspond to the local maxima of $f^{\star}_{ts,n}$. For each specified interval we find $t_0$ with $f^{(1)}_{wss,n}(t_0) \approx 0$ and $t_0$ inside or close to the actual interval. Afterwards we determine $t_{l_2} \leq t_{l_1} \leq t_0 \leq t_{r_1} \leq t_{r_2}$ with

$$|f^{(1)}_{wss,n}(t_{l_i})| \approx |f^{(1)}_{wss,n}(t_{r_i})| \approx \text{Median}(|f^{(1)}_{wss,n}|), \quad \text{for } i = 1, 2 \tag{16}$$

and $f^{(1)}_{wss,n}(t) \geq 0$ for $t \in [t_{l_2}, t_{l_1}]$ and $f^{(1)}_{wss,n}(t) \leq 0$ for $t \in [t_{r_1}, t_{r_2}]$. The initial interval is then extended to $[t_{l_2}, t_{r_2}]$. The final intervals are taken as delimiting the peak. Peaks for which $t_{l_i}, t_{r_i}$ do not exist are ignored. The increased intervals delimit the peaks and are removed from the data. The remaining data set is approximated again using a weighted smoothing spline and the result $f_{bl,n}$ is the estimate of the baseline. The upper panel of Figure 6 shows the dataset of Figure 1 with automatically fitted baseline. The lower panel shows the data after the baseline has been subtracted.

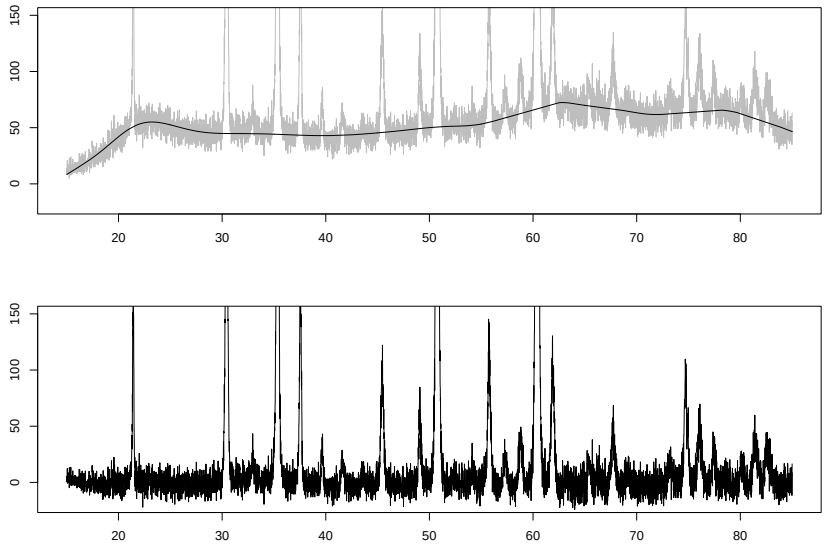


Figure 6: Baseline approximation and data with removed baseline.

# 6 The decomposition of the peaks

We now address the third problem which is to decompose each peak into a finite sum of kernels chosen from a parametric family. This is shown in Figure 7 where the peaks in the left column have been decomposed into one, two and two components as shown in the right column. The decomposition is an ill–posed problem which we regularize by looking for a solution with the smallest number of components. The exact mathemtical formulation leads to a non–convex minimization problem and we describe our algorithm below. We treat the intervals defined by (16) separately. Let $\{t_l, t_{l+1}, \ldots, t_m\} \subseteq \{t_1, \ldots, t_n\} \subseteq [0,1]$ be the segment under consideration and $L := m - l + 1$ its length. We denote by

$$\tilde{y}(t_i) = y(t_i) - f_{bl,n}(t_i) \text{ for } i = l, \ldots, m$$

the measurements in the interval, where the baseline $f_{bl,n}$ has been subtracted and will subsequently only be used for standardization of residuals. We now construct an approximation to the data $\tilde{y}(t_l), \ldots, \tilde{y}(t_m)$ which we will denote by $f_{pk,n}(t)$. Note that $f_{pk,n}$ is only defined on the peak intervals.
The main difficulty now is that not only location, power and shape of the components are unknown, but also their number within an interval, as they may be strongly overlapping. The individual components are chosen from a parametric family of kernel functions with one kernel for each component. In much the same way as in the previous sections, we start with the simplest model (1 kernel) and then check whether an adequate approximation $f_{pk,n}(t_l), \ldots, f_{pk,n}(t_m)$ to the data exists, i.e.

whether the appropriately standardized residuals satisfy

$$|\tilde{w}(\tilde{\mathbf{y}}, I, f_{pk,n}, \tilde{\mathbf{\Sigma}}_{\mathbf{n}})| \leq C_L \tag{17}$$

for all intervals $I \subseteq \{t_l, \ldots, t_m\}$. We give details on the choice of the set of intervals, the noise level $\tilde{\mathbf{\Sigma}}_{\mathbf{n}}$ and the threshold $C_L$ below, after the description of the procedure. Model complexity (number of kernels) is increased until the criterion is satisfied. Physical characteristics of interest like power, full width at half maximum (FWHM) and exact location of the peak components can be calculated from the obtained estimated components.
Each decomposition is of the form

$$f(t) = \sum_{i=1}^{k} \gamma_i p(t; \beta_i) \tag{18}$$

where $k$ denotes the number of kernels (starting with $k = 1$) and $\gamma_i$ are nonnegative weights. The kernels $p$ depend on a vector of parameters $\beta_i$ including location and shape parameters. Depending on the parametrization, the weights $\gamma_i$ correspond either to the maximum height or to the power of the peak component. The number and interpretation of the parameters as well as the range of admissible values depend on the family of curves used. Several choices of kernels are possible, but the most widely used families all include densities of the Gaussian and Cauchy (also known as Lorentz) distributions as, in some sense, extreme cases cf. [12]. Among these families are *Voigt functions*, which are convolutions of Gaussian and Cauchy densities, so-called *Pseudo Voigt functions*, which are convex combinations of Gaussian and Cauchy densities, and *Pearson Type VII* curves. However, the approach presented here is not limited to these families of curves, and should work for any suitably chosen parametric family of kernels including asymmetric ones.
In the following, we will only consider the *Pearson Type VII* family, since it works well for our data and prevents some numerical difficulties that occur when using Voigt or pseudo-Voigt functions. The curves have the form

$$p(t; \beta) = p(t; \mu_i, m_i, a_i) = \left(1 + \frac{(t - \mu_i)^2}{a_i^2 m_i}\right)^{-m_i} \tag{19}$$

where $\mu_i$ is the location parameter, $a_i$ measures the width, and $m_i \geq 1$ determines the shape of the curve. For $m_i = 1$, $p$ is the Cauchy density, and since $(1 + \frac{x^2}{m})^{-m} \overset{m\to\infty}{\longrightarrow} \exp(-x^2)$, the shape becomes finally Gaussian for large $m_i$. The kernel is not normalized, so it is not necessarily a probability density. We have $p(\mu_i; \mu_i, m_i, a_i) = 1$, so the weight $\gamma_i$ is the height at the maximum. For each Pearson VII kernel, we have to estimate four parameters: $\mu_i$, $m_i$, $a_i$, and the weight $\gamma_i$.
For fixed $k$ (starting with $k = 1$) we consider signals of the form

$$f_{pk,n}(t) = \beta_0 + \beta_1 t + \sum_{i=1}^{k} \gamma_i p(t; m_i, \mu_i, a_i) \tag{20}$$

where $p(t; m_i, \mu_i, a_i)$ is the Pearson VII function with parameters as described above. The parameters $\beta_0$ and $\theta_1$ are added to allow small changes in the baseline estimate

and should only have small values, e.g. values between $\pm d_0 := \pm 5\%$ of the height of the initial baseline estimate for $\beta_0$ and between $-d_1$ and $d_1$ for $\beta_1$. We choose $d_1$ so that the slope of the baseline estimate can change by at most 5 counts per $2\theta$. The parameter vector

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \gamma_1, \mu_1, m_1, a_1, \ldots, \gamma_k, \mu_k, m_k, a_k)$$

now completely determines the shape. Since it is not possible to check directly whether an adequate approximation of given complexity $k$ exists which satisfies our criterion, we have to focus on one or several promising candidates. Since the estimate should be "close" to the data, a natural choice is the nonlinear weighted least squares estimate, which leads to the following optimization problem:

$$R(\boldsymbol{\beta}) = \sum_{j=l}^{m} \left( \frac{f_{pk,n}(t_j; \boldsymbol{\beta}) - \tilde{y}(t_j)}{\Sigma_n(t_j)} \right)^2 \longrightarrow \min! \tag{21}$$

with

$$f_{pk,n}(t; \boldsymbol{\beta}) = \beta_0 + \beta_1 t + \sum_{i=1}^{k} \gamma_i p(t; m_i, \mu_i, a_i)$$

subject to

$$\begin{aligned}
&-d_j < \beta_j < d_j && (j = 0, 1)\\
&\gamma_i, a_i > 0 && (i = 1, \ldots, k)\\
&t_l < \mu_1 < \cdots < \mu_k < t_m && (i = 1, \ldots, k)\\
&m_i \geq 1 && (i = 1, \ldots, k)
\end{aligned}$$

Simple re-parametrizations can be used to eliminate the interval constraints, for example logarithms and affine transformations of the logit-function. For $k > 1$ every signal has $k!$ different parameterizations because of interchangability of the kernels, and a reduction of the search space is achieved by enforcing an ordering in the location parameters $\mu_1 < \cdots < \mu_k$. An appropriate transformation is given by Jupp [10].
Since (21) generally has a large number of local minima, we proceed iteratively in the following manner. We choose a starting value at random from a uniform distribution over a suitably chosen rectangular set which contains all reasonable parameter values. This is followed by a Newton-type procedure to find the nearest local minimum of $R$. We use the so-called BFGS-Method as described in chapter 3.2 of [7], but any similar algorithm should suffice. The local minimum of $R$ is then compared to the lowest value previously found. If it is lower, we check the conditions (17) (see below), and stop if they are fulfilled. In this case, an adequate approximation with given complexity has been found. Otherwise, we draw a new starting value and repeat these steps. If no adequate approximation is found within a specified number of iterations (e.g. 7500), the output is the best local minimum of $R$ that has been found. The number of kernels is then increased by one, and the procedure is started anew.

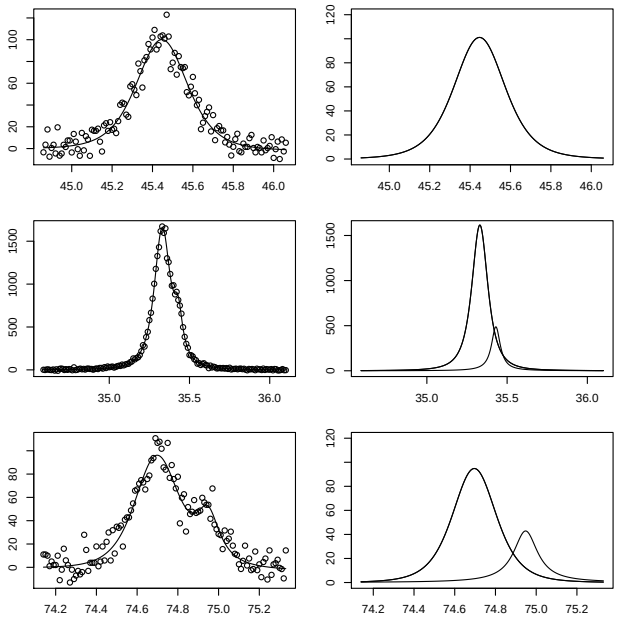


Figure 7: Some intervals of the data and fitted curves (left column). The right column displays the single peaks.

Note that this optimization algorithm does not directly aim at obtaining an adequate approximation in the sense of the criterion (17) but tries to find local optima of the weighted least squares residual function (which is less difficult, since $R$ is infinitely differentiable). In general, a solution of complexity $k$ that fulfills (17) is not necessarily a local or global optimum of the least squares function. However, our experience indicates that this heuristic works sufficiently well. In addition, the algorithm need not find the best solution (or something reasonably close) in a given number of iterations. For these reasons, the number of necessary kernels may be overestimated in some cases. Since the search algorithm is in part stochastic, another run might produce a better result.
When checking (17), we standardize the residuals using

$$\tilde{\mathbf{\Sigma}}_{\mathbf{n}}(t) = \sqrt{f_{bl,n}(t) + f_{pk,n}(t)}.$$

This is similar to (7), but now we have a much smaller number of observations. This allows us to use all subintervals of $t_l, \ldots, t_m$ in (17). We use an efficient algorithm given by Bernholt and Hofmeister [1] for this. However, special care has to be taken in choosing the critical value $C_L$, since the asymptotic choice given in section 2 is not valid here. We estimate a suitable threshold by means of simulation: We draw $L = m - l + 1$ independent and identically distributed random variables $Z_i$ from a standard normal distribution and calculate

$$\max_{1 \le q \le r \le L} \frac{1}{\sqrt{|I|}} \Big| \sum_{i=q}^{r} Z_i \Big|. \tag{22}$$

We then use the upper 0.95-quantile from 10000 replications. Note that $C_L$ depends on the interval only through its length $L$.
Figure 7 shows some examples of intervals where a successful decomposition is obtained with only one or two components.
Once a solution is found the characteristics of the peak components can be estimated by calculating the values for the fitted curves. For Pearson VII curves as used here, the corresponding weight parameter $\gamma_i$ equals the maximum height. The integrated intensity $I_i$ of the $i$-th component is obtained by

$$I_i = \frac{\Gamma(m_i - 1/2)\sqrt{\pi m_i} a_i}{\Gamma(m_i)} \gamma_i,$$

cf. [9]. The full width at half maximum of the $i$-th kernel depends only on the shape and scale parameters $m_i$ and $a_i$, and can be calculated explicitly by

$$FWHM_i = 2a_i \sqrt{m_i(\sqrt[m_i]{2} - 1)},$$

cf. [9]. Of course, $I_i$ and $FWHM_i$ must be scaled appropriately according to the grid width. If an interval contains two or more strongly overlapping peak components, or if the components have very low intensities, the values calculated may not be reliable.

# 7 Discussion

In this article, we propose a fully automatic five-step procedure that determines the number, positions, powers and shapes of the relevant peaks and their components in X-ray diffractograms. It can be applied when little or no prior knowledge of approximate peak positions is available, as is often the case in the analysis of the morphology of thin films. The whole procedure is based on the principle of choosing the simplest possible fit that is an adequate explanation of the data. We employ a criterion based on residuals to formalize the latter, while simplicity is measured by the number of peaks or in terms of smoothness. This ensures that all relevant aspects of the data are captured while at the same time overfitting is avoided.
The procedure is based on recent advances in nonparameteric regression and denoising techniques like the taut string method and weighted smoothing splines. These two different data approximation methods are normally used in Gaussian white noise settings. Here they are modified in order to deal with Poisson-noise. Both yield approximations of data with different properties, and they are combined to take advantage of this. The taut string method is very successful in producing approximations with a small number of local extremes and is therefore used in step one to determine the positions of the peaks. However, the approximation is piecewise constant and thus cannot be used for the detection of the boundaries of the peaks, which are necessary for a separate fit of the baseline. On the other hand the weighted smoothing splines - used in step two - give an approximation to the data which is twice continuously differentiable but tends to overfit the data. Hence a restriction to smoothing splines would possibly cause a large number of spurious peaks. Therefore, based on the peak locations from the taut string result the first

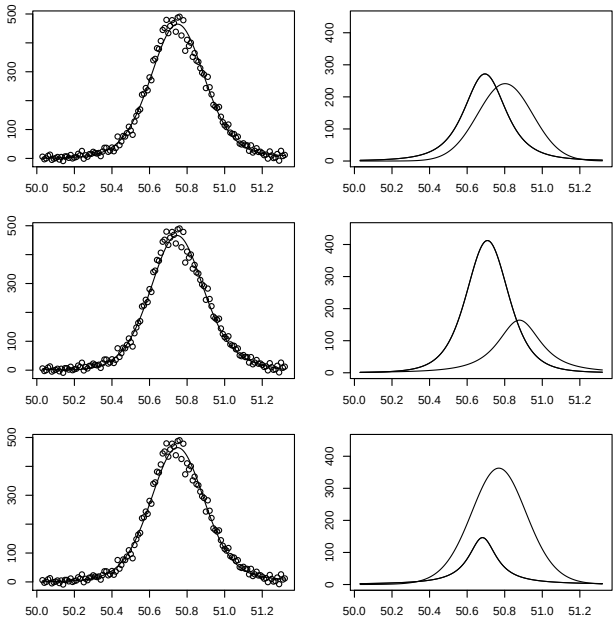


Figure 8: Three different approximations with two kernels to the same data. The resulting curves (left column) are very similar, but the seperated peaks are very different (right column)

derivative of the fitted spline is used in step three to determine baseline and peak regions of the data set. For this purpose a threshold is needed which possibly would have to be varied in a different setting. The choice might also depend on additional knowledge about the data. In step four of the procedure weighted smoothing splines are used to estimate the baseline which is necessary to determine the power of the peaks.

The last step of the procedure - decomposition of the peaks - requires the solution of a nonlinear least-squares problem. It is in the nature of this problem that multiple solutions may exist, especially when fitting two or more kernels. Figure 8 shows such a problematic case: No adequate approximation with one kernel is found, but there exist different combinations of two kernels which give an adequate approximation to the data. The method picks just one of them, possibly different ones in different runs. This could in part be remedied by proceeding as follows: After an approximation that satiesfies (17) is found, try again to find a solution with the same number of kernels several times. This will provide some idea of the variability of possible solutions for this particular segment of the data. In some cases these will be very similar, but they might also differ strongly. The experimenter may then either choose the solution that is the most meaningful, based on partial prior knowledge about possible components of the material under consideration or on the results for the other peak intervals, or decide that no physically meaningful, unambigous interpretation of this part of the data is possible.

This makes the fifth part of the procedure somewhat less automatic than the others,

as the judgement of the experimenter is still required to select one of the explanations offered by the procedure. However, this is done based on the output of the procedure, and no choices or interactions of the experimenter are required while it is running. Without utilizing prior knowledge about possible crystal structures of the material, no automatic procedure can fully avoid these problems, since the data by itself may not contain enough information to decide between several possible explanations.

## Acknowledgements

This work has been supported by the Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475) of the German Research Foundation (DFG). The authors also want to thank Ursula Gather for helpful discussions and suggestions.

## References

[1] T. Bernholt and T. Hofmeister. An algorithm for a generalized maximum subsequence problem. In J.R. Correa, A. Hevia, and M. Kiwi, editors, *Latin 2006: Theoretical Informatics*, volume 3887 of *Lecture notes in Computer Science*, pages 178–189, Berlin, Heidelberg, 2006. Springer Verlag.

[2] P. L. Davies. Data features. *Statistica Neerlandica*, 49:185–245, 1995.

[3] P.L. Davies and A. Kovac. Local extremes, runs, strings and multiresolution (with discussion). *Annals of Statistics*, 29(1):1–65, 2001.

[4] P.L. Davies, A. Kovac, and M. Meise. Asymptotics, local adaptivity, shape regularization and confidence bounds. Technical Report 13-07, Sonderforschungsbereich 475, Fachbereich Statistik, University of Dortmund, Germany, 2007.

[5] P.L. Davies and M. Meise. Approximating data with weighted smoothing splines. Technical Report 48/05, Sonderforschungsbereich 475, Fachbereich Statistik, University of Dortmund, Germany, 2005.

[6] L. Dümbgen and V.G. Spokoiny. Multiscale testing of qualitative hypotheses. *Annals of Statistics*, 29(1):124–152, 2001.

[7] R. Fletcher. *Practical Methods of Optimization. Second Edition.* John Wiley, Chichester, 2000.

[8] P.J. Green and B.W. Silverman. *Nonparametric regression and Generalized Linear Models: a roughness penalty approach.* Number 58 in Monographs on Statistics and Applied Probabality. Chapman and Hall, London, 1994.

[9] Hall, Jr., M.M., V. G. Veeraraghavan, Herman Rubin, and P. G. Winchell. The approximation of symmetric x-ray peaks by pearson type vii distributions. *Journal of Applied Crystallography*, 10:66–68, 1977.

[10] D.L.B. Jupp. Approximation to data by splines with free knots. *SIAM Journal of Numerical Analysis*, 15(2):328–343, 1978.

[11] D. Mergel, T. Thiele, and Z.H. Qiao. Texture analysis of thin $In_2O_3$:Sn films prepared by direct-current and radio-frequency magnetron-sputtering. *Journal of Materials Research*, 20(9):2503–2509, 2005.

[12] L.S. Zevin and G. Kimmel. *Quantitative X-ray diffractometry*. Springer, New York, 1995.