Patterns in Calabi–Yau Distributions

Yang-Hui He[a, *], Vishnu Jejjala[b, †], Luca Pontiggia[b, ‡]

[a] *School of Physics, NanKai University, Tianjin, 300071, P.R. China and*
Department of Mathematics, City University, London, EC1V 0HB, UK and
Merton College, University of Oxford, OX1 4JD, UK

[b] *NITheP, School of Physics, and Mandelstam Institute for Theoretical Physics,*
University of the Witwatersrand, Johannesburg, WITS 2050, South Africa

Abstract

We explore the distribution of topological numbers in Calabi–Yau manifolds, using the Kreuzer–Skarke dataset of hypersurfaces in toric varieties as a testing ground. While the Hodge numbers are well-known to exhibit mirror symmetry, patterns in frequencies of combination thereof exhibit striking new patterns. We find pseudo-Voigt and Planckian distributions with high confidence and exact fit for many substructures. The patterns indicate typicality within the landscape of Calabi–Yau manifolds of various dimension.

[*] hey@maths.ox.ac.uk

[†] vishnu@neo.phys.wits.ac.za

[‡] lucatpontiggia@gmail.com

Contents

1 Introduction

A Calabi–Yau n-fold is a Kähler manifold of n complex dimensions with a trivial canonical bundle. In superstring theory, it serves as a compactification manifold wherein a ten dimensional theory at high energies reduces to an effective theory in four spacetime dimensions. In particular, global $SU(n)$ holonomy ensures that 2^{1-n} of the original supersymmetry is preserved. Thus, confronted by the vacuum selection problem, Calabi–Yau compactifications present an avenue for Standard Model building especially in the context of the heterotic string [1,2]. Indeed, the basis of the landscape is to consider flux compactifications on these geometries [3,4].

To facilitate this approach to a low-energy phenomenology derived from string theory, mathematicians and physicists have constructed large datasets of Calabi–Yau threefolds [5,7–19] as well as various refined analyses of properties thereof [24–31]. By far the largest database was constructed in a *tour de force* of algebraic geometry, combinatorics, physics, and computer algorithms by Kreuzer and Skarke based on the theorems of Batyrev and Borisov [7–12,32,33]. In short, these Calabi–Yau n-manifolds X_n are realized as a smooth hypersurface embedded in a toric variety A_{n+1} of complex dimension $n+1$; the Calabi–Yau condition simply translates to the requirement that the polytope defining A_{n+1} be **reflexive**. We will henceforth consider only such Calabi–Yau manifolds, of which there are a plethora.

Let us briefly recollect what all this means. The (possibly singular) toric variety A_{n+1} is specified by an integer polytope Δ in $\mathbb{R}^{n+1}$, which is a collection of vertices (dimension 0) each of which is an $(n+1)$-vector with integer entries and such that each pair of neighboring vertices defines an edge (dimension 1), each pair of edges defines a face (dimension 2), etc., all the way up to a facet (dimension n). Alternatively, Δ can be defined by a set of integer linear inequalities, each of which slices a facet. The polytope is then the convex body in $\mathbb{R}^{n+1}$ enclosed by these facets. We will always include the origin as being contained in Δ. Using the usual dot product $\langle\ ,\ \rangle$ inherited from $\mathbb{R}^{n+1}$, the dual polytope is defined by

$$\Delta^\circ := \left\{ v \in \mathbb{R}^{n+1} | \, \langle m, v \rangle \geq -1, \forall\ m \in \Delta \right\} \ . \tag{1.1}$$

The polytope Δ is *reflexive* if all the vertices of Δ° are integer vectors. In this case, we can define the Calabi–Yau hypersurface X_n explicitly as the polynomial equation

$$\sum_{m\in\Delta} c_m \prod_{r=1}^{k} x_r^{\langle m, v_r\rangle + 1} = 0 \ , \tag{1.2}$$

where $v_{r=1,\ldots,k}$ are the vertices of Δ° with k being the number of vertices of Δ° (or equivalently the number of facets of Δ), x_r are the coordinates of A_{n+1}, and c_m are numerical coefficients parameterizing the complex structure of X_n. Indeed, the reflexivity of Δ ensures that the exponents are integral whereby making the hypersurface polynomial as required.

The classification of these Calabi–Yau manifolds thus amounts to that of reflexive polytopes in various dimensions, and the intense computer work of Kreuzer and Skarke was to combinatorially find such polytopes. For $n=1$, there are 16 such polytopes in $\mathbb{R}^2$, and we have Calabi–Yau onefolds, or elliptic curves. For $n=2$, there are 4319 such polytopes in $\mathbb{R}^3$, and we have Calabi–Yau twofolds, or K3 surfaces. For $n=3$, there are $473,800,776$ such polytopes (which was a formidable computer task!), and we have the Calabi–Yau threefolds. This sequence

$$\{1, 16, 4319, 473800776, \ldots\} \tag{1.3}$$

of remarkable growth rate, can be found in the Online Encyclopedia of Integer Sequences [34]. The numbers in higher dimension are still not known, nor has there been an asymptotic analysis of their growth. It should be emphasized that generically a reflexive polytope corresponds to a *singular* toric variety even though the hypersurface is chosen (by generic coefficients c_m) to miss the singularities and hence ensuring the smoothness of the Calabi–Yau X_n. For example, of the some half-billion reflexive polytopes in $\mathbb{R}^4$, only 136 A_4 are in fact smooth [35]. As we desingularize the toric variety by various star-triangulations of Δ, we are led to potentially *inequivalent* Calabi–Yau manifolds. In principle, the *same* Calabi–Yau geometry can arise from different reflexive polytopes or triangulations of a given reflexive polytope. Whereas K3 is essentially unique, we do not know how many Calabi–Yau threefolds there are. A systematic study to classify the desingularizations, to compute the necessary topological data, and to build an interactive online database [16] is under way. The moral is that there are almost certainly far more than half a billion Calabi–Yau threefolds!

Luckily, the Hodge numbers depend only on the polytope and not on the choice of desingularization. (The intersection numbers, however, do depend on the choice.) For Calabi–Yau threefolds, the pair of Hodge numbers $(h^{1,1}, h^{1,2})$ is a famous quantity. Indeed, the plot in Part (a) of Figure 1 has become iconic. Here, the sum $h^{1,1} + h^{1,2}$ is plotted against the Euler number $\chi = 2(h^{1,1} - h^{1,2})$, and the left-right symmetry supplies "experimental evidence" for *mirror symmetry.* There is enormous redundancy in this data: of the some half a billion reflexive polytopes, there are only $30,108$ distinct pairs of Hodge numbers and the pair $(27, 27)$ dominates the multiplicity, totaling almost one million. We have attempted to visualize the distribution of the multiplicity by having a color density plot of the logarithm of the number over each pair in Part (b) of the Figure.

Understanding this multiplicity forms the inspiration for the present work. While there have been analyses on the *shape* of the funnel-like plot [24, 29, 31], there has not been much work on its *density*, *i.e.*, the distribution of the multiplicity of Hodge data for the Calabi–Yau manifolds of various dimension. Of course, fundamentally, this is entirely due to the combinatorics of reflexive polytopes and might in principle be analytically determined. However, given the complexity of the problem it is expedient to analyze the available data which have been compiled over the years, observe intriguing patterns, and draw statistical inferences before turning to analytic treatments. This is what we achieve in this work.

The organization of the paper is as follows. We perform a detailed analysis on the structure and behavior of the threefold data in Section 2. This is motivated by looking for an exact function describing the relationship of the distribution of the Hodge pairs $(h^{1,1}, h^{1,2})$ with frequency.

In Section 2.1, we study the distribution of $(h^{1,1} - h^{1,2}, f)$. We find that this distribution is composed of a family of curves, for which each curve can be described using a modified pseudo-

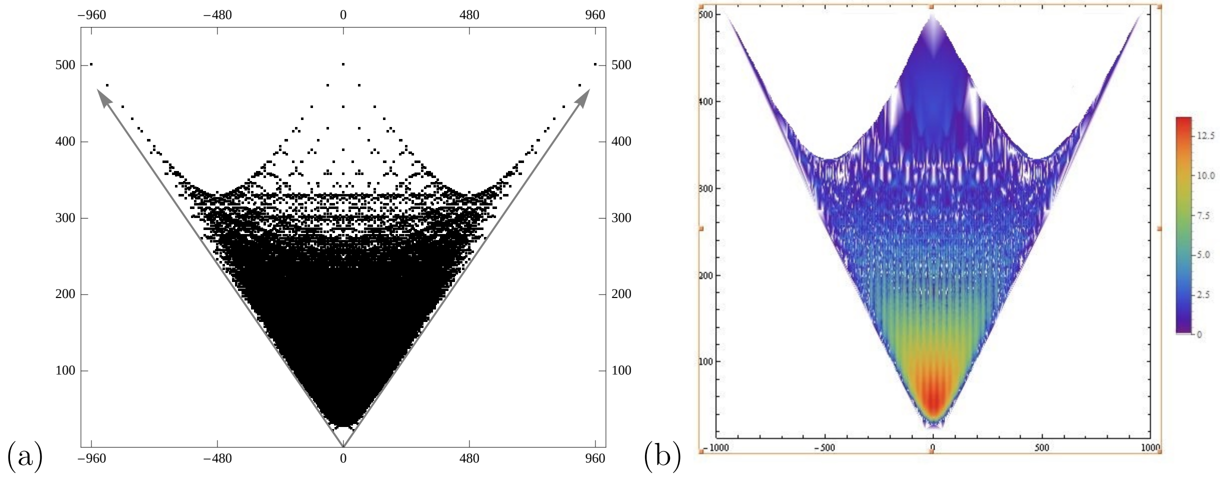


Figure 1: (a) The cumulative plot of $\chi = 2(h^{1,1} - h^{1,2})$ on the abscissa versus $h^{1,1} + h^{1,2}$ on the ordinate for Calabi–Yau threefolds as hypersurfaces in toric fourfolds; (b) marking also the natural logarithm of the multiplicity of the Hodge pair with a color grading.

Voigt model. Although an approximation, the model is able to describe the general trend of the data, as well as some additional fine structure within each individual data point. Performing an analysis on the parameter relationships shows that three out of the five parameters can be expressed as a single variable, but conclude that additional modifications need to be introduced in the model to overcome certain shortfalls.

Subsequently, Section 2.2 performs an analysis on the structure of $(h^{1,1} + h^{1,2}, f)$. Similarly, this distribution is composed of a family of curves for which each curve can be described using a Planckian profile. Combining the regression analysis for each curve within the distribution, we construct a single function able to approximately model the entire distribution of $(h^{1,1}+h^{1,2}, f)$ with only two variables. Section 2.3 uses the model developed in Section 2.1 to describe the distribution of the Euler number χ.

Section 2.4 is dedicated to the description of model validation in our context, as the usual statistical tests are inadequate. In Section 3 and Section 4, we perform primary analyses of Calabi–Yau twofolds (Picard number and multiplicity) and Calabi–Yau fourfolds. Due to the lack of a complete data set, we are unable to provide a thorough analysis of the fourfolds as with threefolds. Finally the Appendix presents many supplementary plots and figures for the various sections. We conclude with a summary and outlook in Section 5.

2 Calabi–Yau Threefolds

As advertised in the Introduction, we will begin with the analysis of threefolds and identify patterns within this rich distribution of Hodge numbers and their frequency as plotted in Figure 1. It turns out striking patterns do exist, pointing to a definite structure within the threefold data, which consists of the triple $(h^{1,1}, h^{1,2}, f)$, where f is the number of reflexive polytopes in the Kreuzer–Skarke database with the given Hodge pair. Here, $h^{1,1}$ and $h^{1,2}$ respectively count the Kähler and complex structure moduli of the Calabi–Yau obtained from the reflexive polytope. More precisely [6], we have that

$$\begin{aligned} h^{1,1}(X) &= \ell(\Delta^*) - \sum_{\text{codim}\theta^*=1} \ell^*(\theta^*) + \sum_{\text{codim}\theta^*=2} \ell^*(\theta^*)\ell^*(\theta) - 5; \\ h^{1,2}(X) &= \ell(\Delta) - \sum_{\text{codim}\theta=1} \ell^*(\theta) + \sum_{\text{codim}\theta=2} \ell^*(\theta)\ell^*(\theta^*) - 5 \ . \end{aligned} \tag{2.1}$$

In the above, Δ is the defining polytope for the Calabi–Yau threefold X and Δ^* is its dual. Moreover, θ and θ^* are the faces of specified codimension of these polytopes respectively; $\ell(\ )$ is the number of integer points of the polytope while $\ell^*(\ )$ is the number of interior integer points. Indeed, our analysis of the distribution of Hodge numbers ultimately reduces to counting these integer points.

To facilitate the analysis, we plot $(h^{1,1} - h^{1,2}, f)$ and $(h^{1,1} + h^{1,2}, f)$ as shown in (a) and (b) of Figure 2, respectively. Recall that the Euler number $\chi = 2(h^{1,1} - h^{1,2})$. We will use the difference $h^{1,1} - h^{1,2}$ rather than the Euler number. In the simplest heterotic constructions, $|h^{1,1} - h^{1,2}|$ corresponds to the index of the Dirac operator and gives the number of generations of particles in the low-energy spectrum [1].

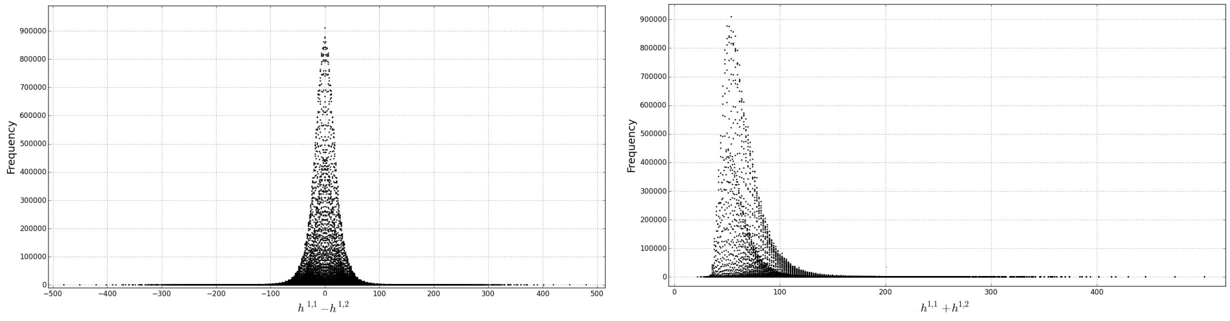
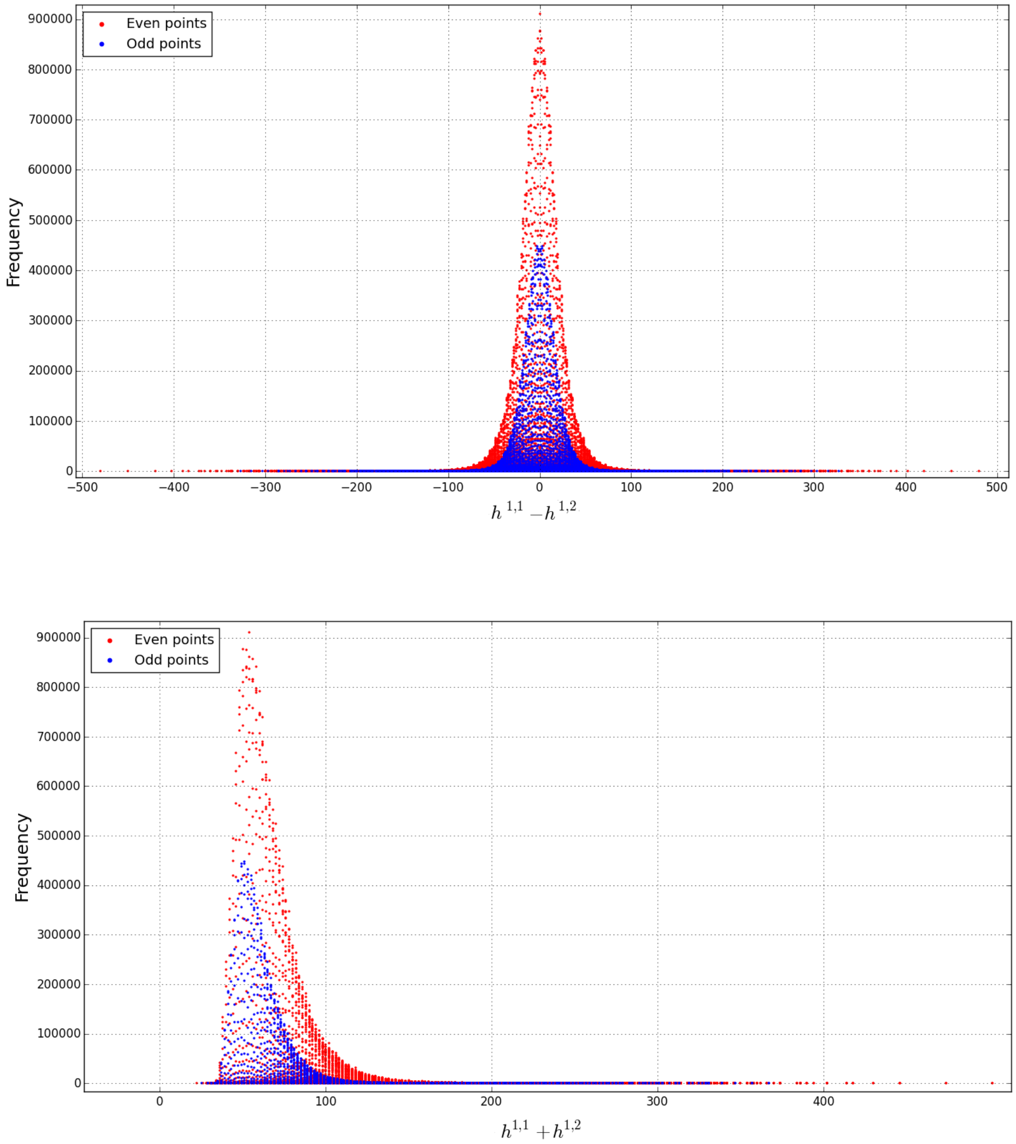


Figure 2: (a) Frequency f plotted against $(h^{1,1} - h^{1,2})$; (b) Frequency f plotted against the sum of Hodge numbers $(h^{1,1} + h^{1,2})$.

By inspection, these plots already exhibit two patterns. Firstly, in both the $h^{1,1} - h^{1,2}$ and $h^{1,1} + h^{1,2}$ plots, there appears to be an inner distribution contained within the outer distribution. We find that these inner and outer distributions are related to the parity of $h^{1,1} \pm h^{1,2}$. Figure 3 elucidates this point by having the odd and even values in different colors. Though this parity structure may be a result of the Kreuzer–Skarke algorithm, its consistent appearance means we need to treat the distributions of even and odd distinctly for now.

The second evident structure which can been seen by inspection, is that the outer edge of the



(a)

(b)

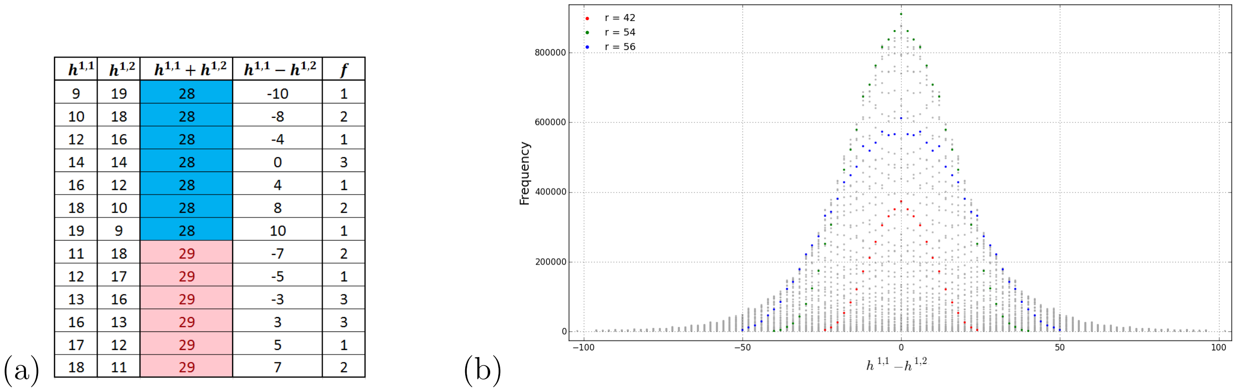
Figure 3: (a) The $h^{1,1} - h^{1,2}$ distribution for threefolds, highlighting the two sub-distributions, where red and blue data points correspond to even and odd values of $(h^{1,1} - h^{1,2})$, respectively; (b) The same, but for $h^{1,1} + h^{1,2}$.

distribution of $h^{1,1} - h^{1,2}$ (Figure 3(a)) appears to follow a normal like curve, whereas the edge of $h^{1,1} + h^{1,2}$ (Figure 3(b)) follows a Planck like curve. It is through the analysis of these distributions that we deduce their characteristic behavior and underlying structure. In the main body of this paper, we outline the results and analysis of only the even distributions for $h^{1,1} - h^{1,2}$ and $h^{1,1} + h^{1,2}$, except where it is important to present both. It turns out that any structure and patterns which are found in the even distributions for $h^{1,1} - h^{1,2}$ and $h^{1,1} + h^{1,2}$ are found identically in the odd distribution (see Appendix for various plots).

2.1 Analysis of $h^{1,1} - h^{1,2}$

Before we can present the results, it is important to explain some notation. When working with the distribution of $h^{1,1} - h^{1,2}$, we find that it is composed of many curves, whose individual structure is the same as the "edge" or boundary of the distribution mentioned earlier. As a consequence of this, we refer to $h^{1,1} - h^{1,2}$ as being composed of a "family of curves." Each curve is then classified by its **r-value**, where $r = h^{1,1} + h^{1,2}$. It is important to be clear that in this analysis, although $h^{1,1} - h^{1,2}$ is just half the Euler number, we are not summing over all the possible values of $h^{1,1} + h^{1,2}$. We are keeping these values distinct: hence, the r-curves we obtain. Later on, in Section 2.3, we sum over all possible values of $h^{1,1} + h^{1,2}$ to get two plots representing the full Euler number distribution.

Consider the example in Figure 4(a). By ordering the data in terms of $h^{1,1}+h^{1,2}$, one can classify data sets within $h^{1,1} - h^{1,2}$ by an r-value. Holding r fixed, we can plot the frequency f versus the difference $h^{1,1} - h^{1,2}$. We call each value of r a curve, which we can overlay on the same plot. In this example, we tabulate data for curves identified by $r = 28$ and $r = 29$. As a further illustration, we show explicitly the curves of the even distribution within $h^{1,1} - h^{1,2}$ for $r = 42, 54, 66$ in Figure 4(b). By mirror symmetry, the curve is symmetric about the vertical axis, where $h^{1,1} - h^{1,2} = 0$.



$h^{1,1}$	$h^{1,2}$	$h^{1,1}+h^{1,2}$	$h^{1,1}-h^{1,2}$	f
9	19	28	-10	1
10	18	28	-8	2
12	16	28	-4	1
14	14	28	0	3
16	12	28	4	1
18	10	28	8	2
19	9	28	10	1
11	18	29	-7	2
12	17	29	-5	1
13	16	29	-3	3
16	13	29	3	3
17	12	29	5	1
18	11	29	7	2

Figure 4: (a) Example of repeated values of the sum $h^{1,1} + h^{1,2}$ being 28 and 29; (b) Three highlighted curves ($r = 42, 54, 66$) within the even $h^{1,1} - h^{1,2}$ distribution. The transparent grey data dots are all the data plots for the distribution. Refer to Figure A.1 for the corresponding odd plot.

We can now perform a regression analysis for each individual curve, in the quest of obtaining a function describing the distribution. In the analysis, we indeed find an approximate function predicting the fine structure of the data. We operate with one caveat: we ignore data points which have a frequency lower than 2000. At large r, the data, whose frequency is below 2000, begins to deviate from our model. The reason for such deviations, comes down to the fact that our model, though remarkably accurate, is still an approximation. We suspect that with further modifications, such deviations can be accounted for and that consequently, it may be possible to find an exact function to map the frequency distribution of $h^{1,1} - h^{1,2}$. Such statements also apply to the distribution of $h^{1,1} + h^{1,2}$.

2.1.1 A Pseudo-Voigt Fit

Due to the normally-distributed, peak-like nature of these curves, we performed a regression analysis using the following models: Gaussian; Cauchy (Lorenztian); Pearson7; Breit–Wigner; Voigt; and Pseudo-Voigt. In the Appendix A.1.2, we perform a side by side comparison. It turns out that both the Voigt model (A.-1a) as well as the Pseudo-Voigt model (A.-2a) give excellent fits.

We focus on the **Pseudo-Voigt model** as it gives the best fits. This is a linear combination of a Gaussian and Lorentzian (Cauchy) distribution:

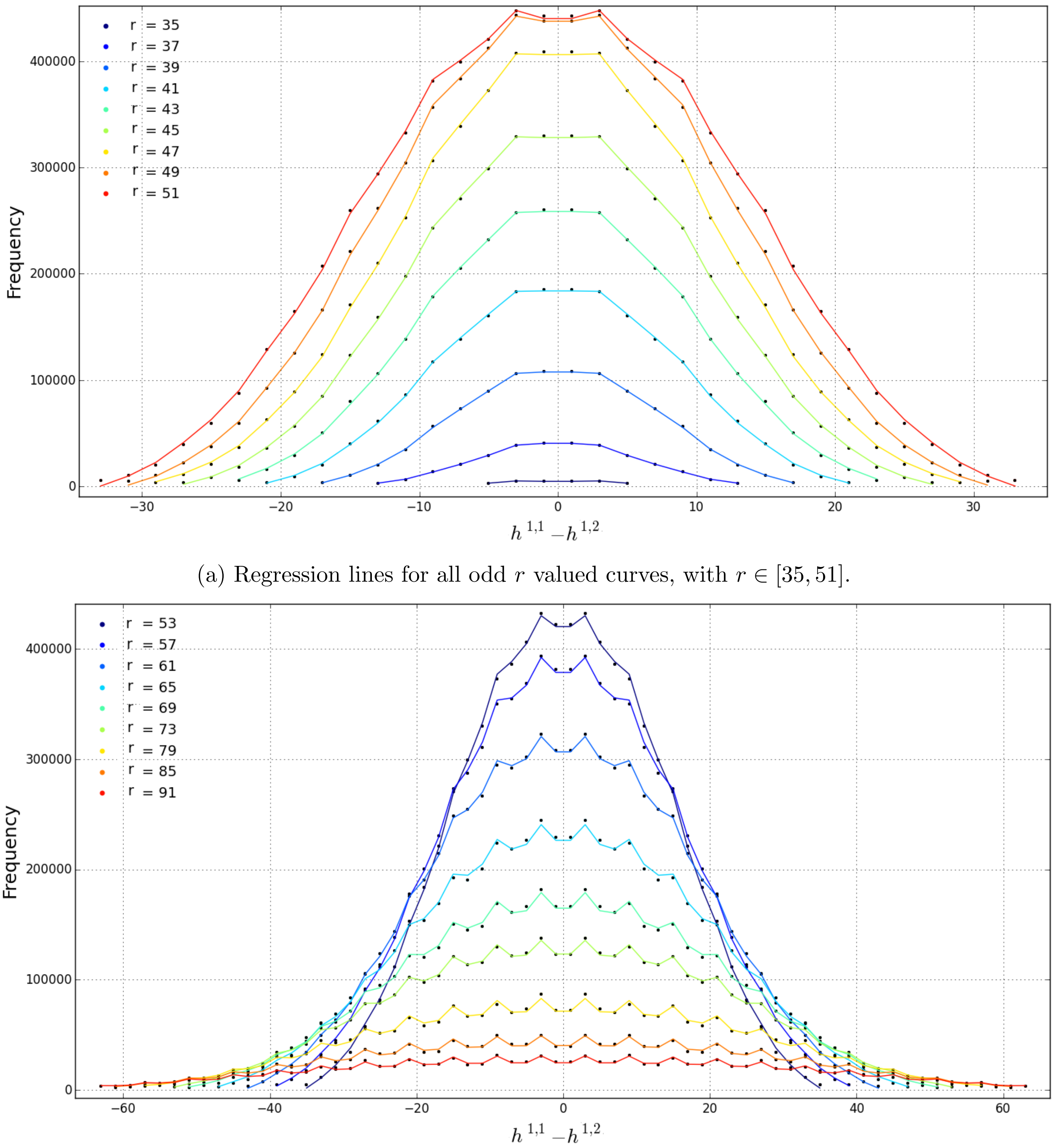
$$f(x, A, \mu, \sigma, \alpha) = (1-\alpha)\frac{A}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \alpha\frac{A}{\pi}\left[\frac{\sigma^2}{(x-\mu)^2+\sigma^2}\right] \ , \tag{2.2}$$

with amplitude (A), center (μ), Gaussian width (σ), and fractional parameter alpha (α). However, we can modify the above distribution slightly so that the amplitude A of the distribution has an oscillating component

$$A(x, A_0, a, b) = A_0 + a\cos(2\pi b \cdot x) \ , \tag{2.3}$$

where A_0 is the original amplitude of a particular curve described by the Pseudo-Voigt distribution, a is the amplitude of oscillations, and b represents the period. By doing a regression analysis one curve at a time using this modified Pseudo-Voigt model, we are almost able to replicate not just the basic structure of each curve, but even the individual behavior of each data point in the entire distribution. (See Appendix A.1.3 for a comparative plot of the all the regression curves using the standard, unmodified, Pseudo-Voigt model.)

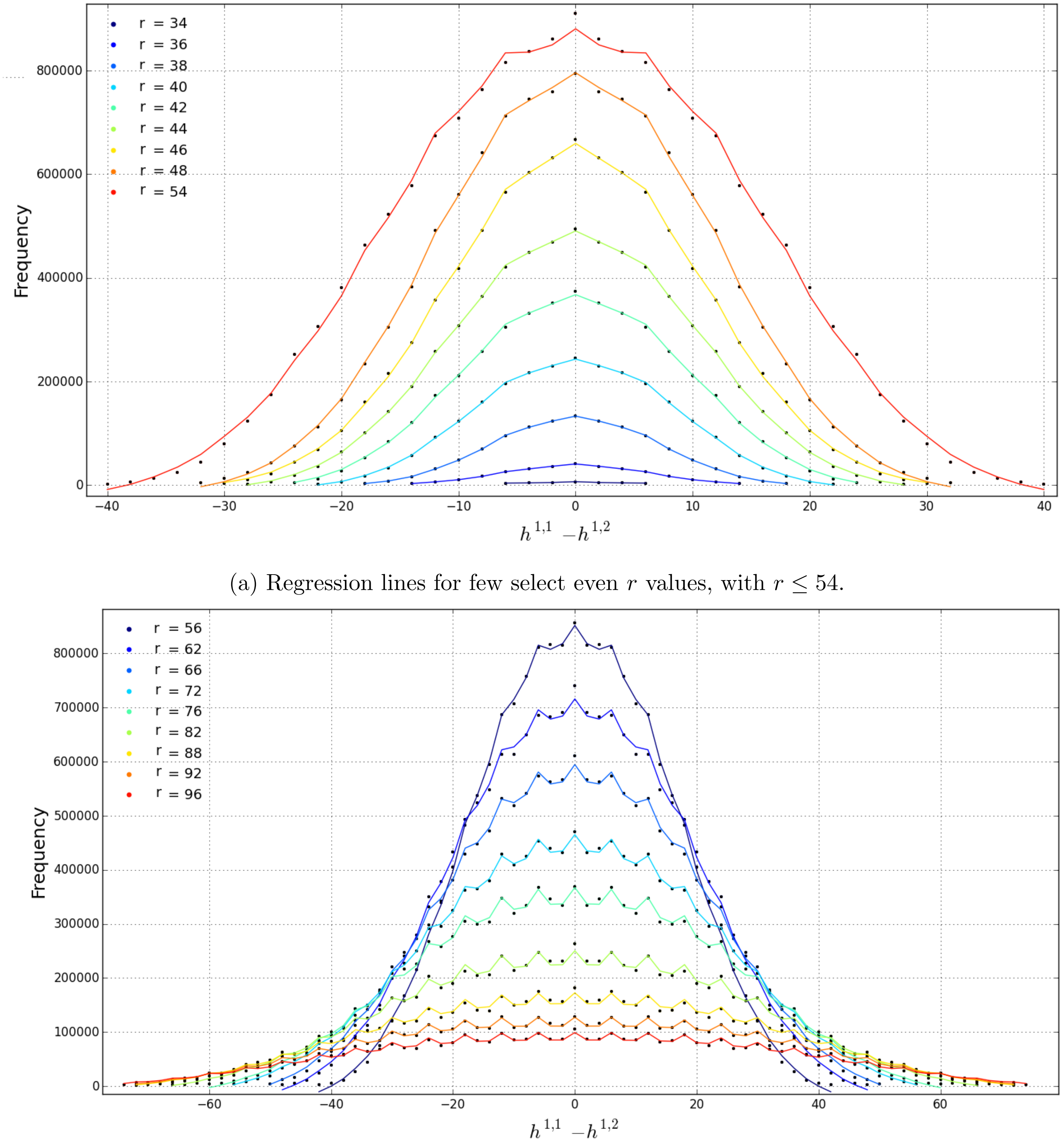
We plot the frequency against $h^{1,1} - h^{1,2}$ for various values of r (odd and even). Figures 5 and 6 are striking in their accuracy.



(a) Regression lines for all odd r valued curves, with $r \in [35, 51]$.

(b) Regression lines for few select odd r values, with $r > 51$.

Figure 5: Plots of frequency against $h^{1,1} - h^{1,2}$ for various odd values of r. Each line represent a modified pseudo-Voigt profile based on the regression analysis for each curve. See A.-1a for a plot of all even curves.



(a) Regression lines for few select even r values, with $r \leq 54$.

(b) Regression lines for few select even r values, with $r > 54$.

Figure 6: Plots of frequency against $h^{1,1} - h^{1,2}$ for various even values of r. Each line represent a modified pseudo-Voigt profile based on the regression analysis for each curve. See A.-1b for a plot of all odd curves.

As these figures illustrate, each curve follows a Pseudo-Voigt profile, however the individual data points seem to "jump" up and down, as if oscillating. It is this behavior of the data points which can be accounted for by the modified Pseudo-Voigt model. To do the regression analysis, we used Python *lmfit* with a custom model which is just the modified Pseudo-Voigt model. The parameters that were fitted are $(A_0, a, b, \sigma, \alpha)$. Due to mirror symmetry, $\mu = 0$. In Appendix A.1.4, one can find a table with the value of every parameter for every curve as well as their reduced χ^2 values.

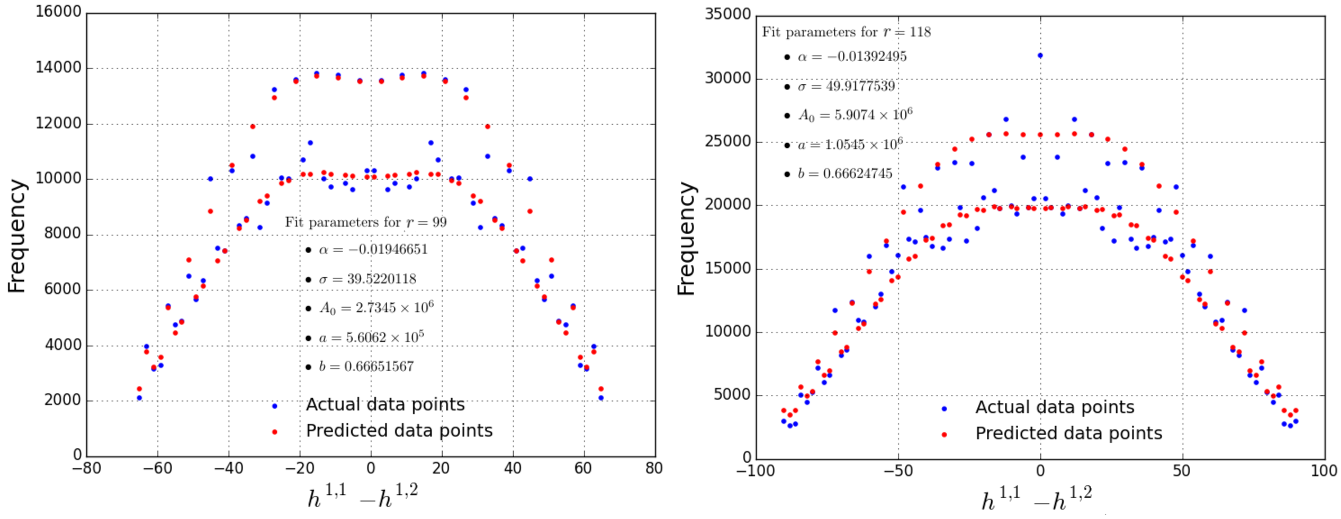


Figure 7: These two plots serve two purposes. The first is to show how the modeled data should really look by using data points (red points) instead of the (perhaps misleading) lines (refer to Comment 1 below). The second purpose is to illustrate that as r becomes large (left plot has $r = 99$, right plot has $r = 118$), the actual data points deviate more and more from the modeled data, implying that there is a missing function in the modified Pseudo-Voigt model which would allow one to describe the data at much lower frequencies.

A few comments explicate the regression lines and the behavior of the distributions.

1. When we refer to the model as being an "excellent fit," it is principally a statement made by inspection of the curves and the data. If one inspects the reduced χ^2 values (Figure A.0), the numbers are large, which statistically does not refer to a good fit. This is misleading however. Firstly, we need to consider that the number of parameters used in the model is five. This allows for a larger χ^2_R value. Secondly, the distribution is based on a discrete set of data. When doing a regression analysis using the modified Pseudo-Voigt model, one obtains an equation which describes a continuous curve. Lastly, the frequency values span over several orders of magnitude. The tiniest deviation from a parametric model — in this case, the modified pseudo-Voigt profile — will be detected in cases where there is such a huge sample size. Typically the predicted model gives data points which are in the range of 0.02% to 3% accuracy from the actual data point. The tail behavior of the model is less accurate however, here the predicted values can be off from between 60% and 80%. For severe cases,

the last data point (larges t value of $h^{1,1} - h^{1,2}$) can have an error of up to 300% — this is another example of the model being less accurate at lower frequency. When one is dealing with such sample sizes, even a 1% error can give a difference of up to a couple of thousand. This difference summed over all the data points for a particular curve result in a large χ_R^2 value. Due to the discussion in Section 2.4 we from now on ignore the χ_R^2 as a test for model validation. Instead we opt for probability plots — which can also be seen in Section 2.4.

2. Since one obtains a continuous model to describe the discrete data, in reality, we should not be plotting fitted curves, but rather fitted data points — as can be seen in Figure 7. It is just illustratively more clear to display the curves. One could in principal work out out what the discrete approximation is to our continuous model.

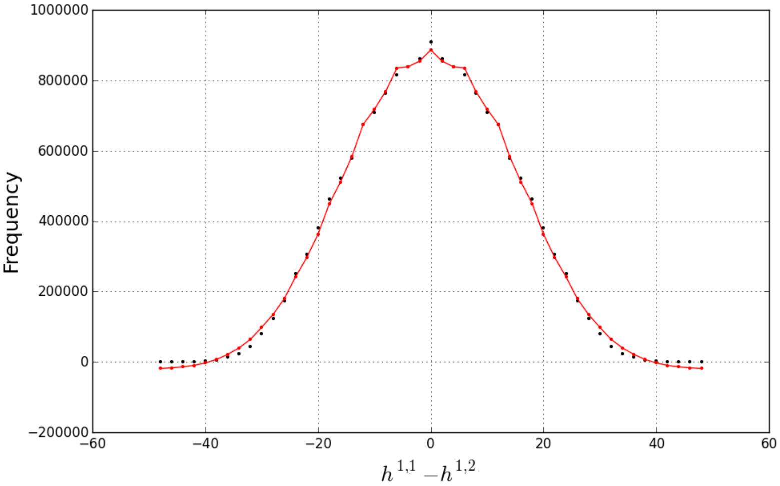


Figure 8: By considering the entire frequency range, the model is not able to adequately describe the tail behavior. The model goes into the negative frequency range instead of tapering off to 0.

3. Although the modified pseudo-Voigt distribution does a good job to model the behavior of the data, one still needs to address the problems experienced with our model at low frequency. A problem which is hidden, by virtue of our cut-off frequency, is that the tail of our models predicts negative values, Figure 8. There is a possibility that by having different variances σ_g, σ_c for the mixing of the two distributions (Gaussian, Lorentzian), one could adjust the tail behavior. Introducing more and more parameters however does not always resolve the problem, as it is possible to over-fit the data. Yes, the model may be more accurate, but one loses physical significance. In a situation like ours, where one does not have any physical backing for choice in models, this line between fitting and over fitting is not so clear.

4. The odd distribution's behavior is more regular. In comparison to the even distribution, as one increases in r value, the behavior of the individual data points remain somewhat constant

relative to the fitted curve. The even distribution becomes more and more irregular as one increases the r value. This suggests that there is an added parameter which seems as if it should be function of r. By regular and irregular we are referring to how well the data point is described by the model.

5. Both distributions become very irregular as the value of r becomes large ($r > 99$ and $r > 120$ for odd and even distributions respectively — see Figure 7). A large r value refers to curves which have a relatively low frequency. Again this suggests that the Pseudo-Voigt model needs to some how have some function of r which "distorts" the behavior of the curves as r increases (by the looks of how the real data deviates from the modeled one, it seems that the missing functions is also oscillating in nature).

There exist, however, certain cases where the model is exact. In other words predicted values are the same as the actual values. This happens when one adjusts the frequency cutoff for each r curve individually. That is to say, we only examine data points with at least f_0 reflexive polytopes with a given value of r and $h^{1,1} - h^{1,2}$. If there are fewer than f_0 cases, the data is ignored.

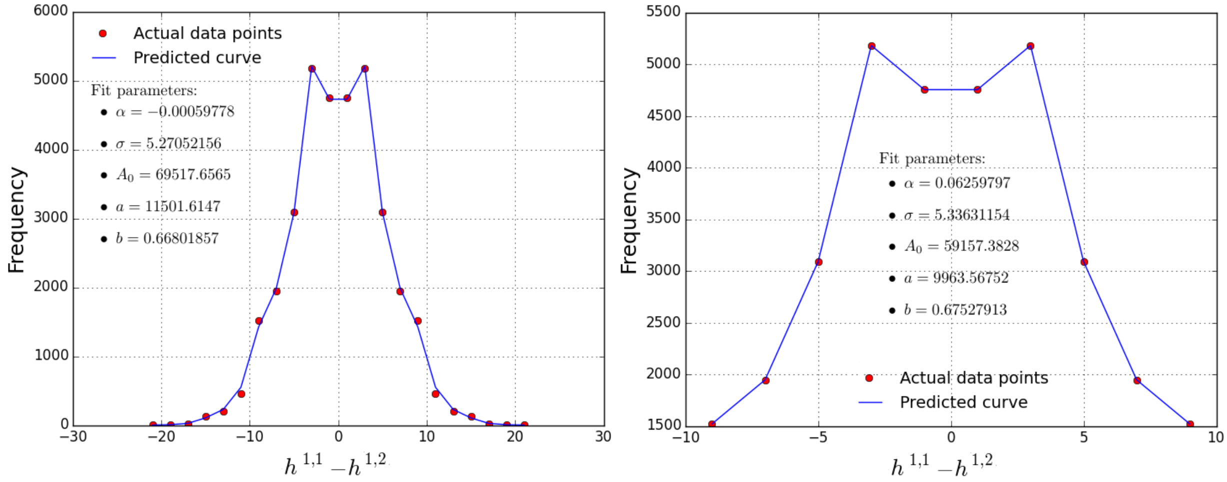


Figure 9: Left plot shows the modeled line according to the modified Pseudo-Voigt distribution with no cutoff frequency. We obtain a good fit to the data. The right plot has a cutoff frequency of 460, which is equivalent to a percentage cut off of 9.68% (calculated relative to the peak frequency for that r-curve). This curve is exact.

This trend persists for all values of r, however what becomes apparent is that it's not the percentage cutoff frequency that determines whether or not one gets an exact fit, but rather, the number of data points that remains after the percentage cut of has been effected. Figure A.1 gives a table of how many data points remain after an appropriate cut off percentage has been chosen to achieve a perfect fit. From this table we see that for even curves, one almost always requires 7 data points to achieve a perfect fit; for the odd curves, the number of data points is 10. The reason for this constant number throughout all the curves is that the centers of all the distributions for the various curves are all similar. As soon as one includes a larger number of data points we cannot achieve exact fits, and the model becomes approximate. At very low r values the number of data points remaining after cutoff are not too different to the total number of points. As r increase, the total number of points increase — the fact that we can achieve exact fits becomes less meaningful. The other models — even when including an oscillatory component were unable to give exact fits.

The model is thus much more accurate at low r values, and as r increases the actual data deviates more and more from the fit. This reinforces the statements from the comments that the Pseudo-Voigt model can be modified further with some function $f(r, x)$ such that it will greatly improve the accuracy of the fit, and perhaps even become exact.

After the above analysis, we return to our goal of finding a single function describing the distributions. It is clear from the above that the function has to be a function of at least two variables, $f = f(x, r)$. We thus continue the analysis by plotting all the parameters vs r, in search for any relationships. We find that three parameters σ,b and α can be expressed in terms of r, the other parameters, while they show trends, do not give a precise relationship with r. For the even distribution of $h^{1,1} - h^{1,2}$, the r values range from 36 to 110, whereas for the odd distribution (see Figures A.2a, A.2b) the r values range from 37 to 99. By looking at Figure 10(a), it turns out that:

$$\alpha(r) = c_\alpha \ , \quad b(r) = c_b \ , \quad \sigma(r) = c_{\sigma_1} r + c_{\sigma_2} \ . \tag{2.4}$$

Our model of $h^{1,1} - h^{1,2}$ now looks as follows:

$$\begin{aligned} f(x, r, A_0, a) = (1 - c_\alpha) & \frac{A_0(r) + a(r)\cos(2\pi c_b \cdot x)}{\sqrt{2\pi}(c_{\sigma_1} r + c_{\sigma_2})} e^{\frac{-(x)^2}{2(c_{\sigma_1} r + c_{\sigma_2})^2}} + \\ & c_\alpha \frac{A_0(r) + a(r)\cos(2\pi c_b \cdot x)}{\pi} \left[\frac{(c_{\sigma_1} r + c_{\sigma_2})^2}{x^2 + (c_{\sigma_1} r + c_{\sigma_2})^2} \right] , \end{aligned} \tag{2.5}$$

where $A_0(r)$ and $a(r)$ are two unknown functions yet to be determined (see Figure 10(b) for relationship plots). For replicating the plots as precisely as possible, one would need to keep the parameters, as they are, up to their 17 decimal values, without excluding terms as we have done. If one wants to reproduce the data from the model, one has to use the exact expressions. Making an approximation from an already approximate model leads to large errors.

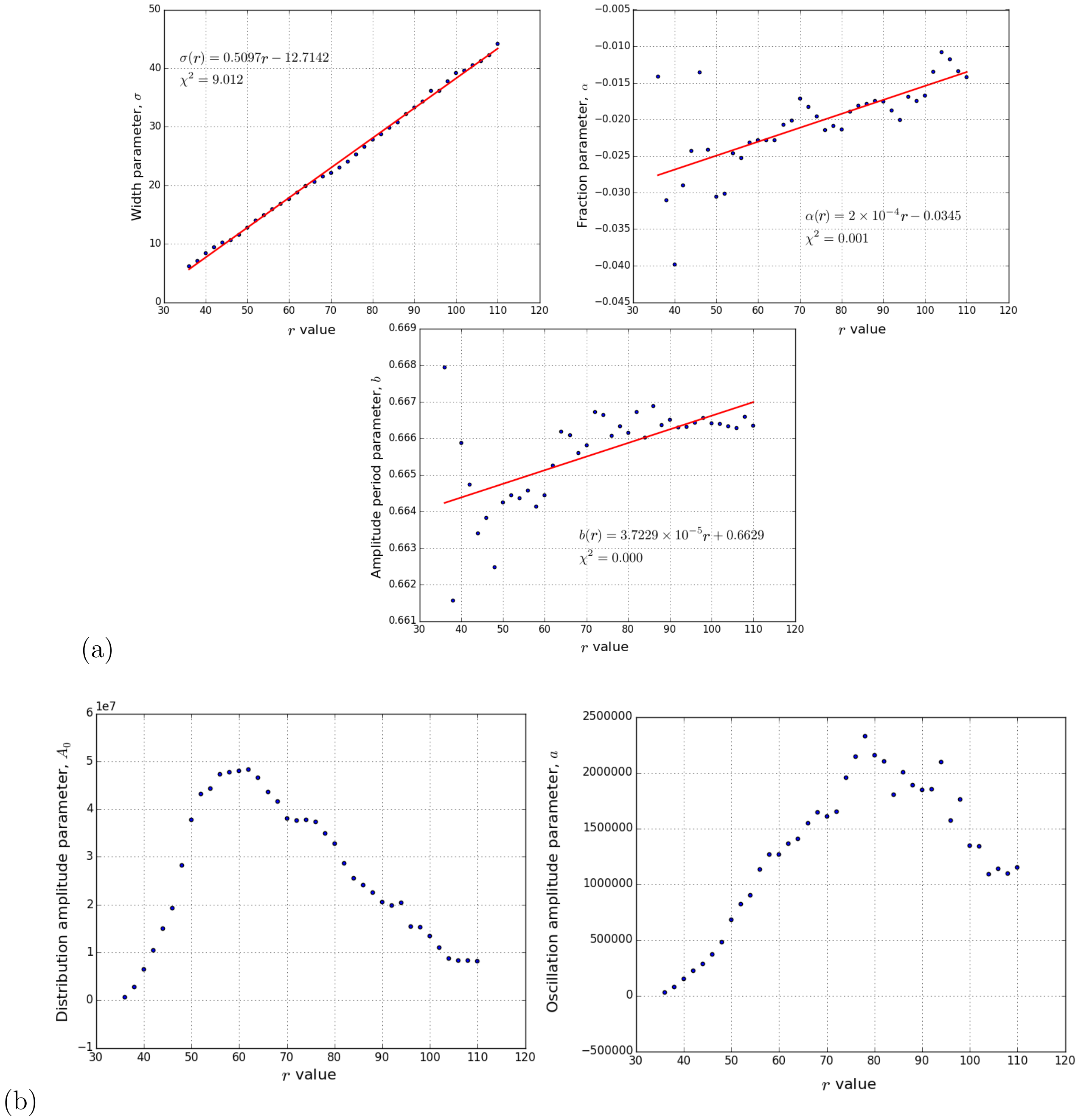


Figure 10: For the even distribution of $h^{1,1} - h^{1,2}$. (a)The width parameter σ has a linear relationship with r such that $\sigma(r) = 0.5097r - 12.7142$. The amplitude period parameter, b, also has a linear relationship, however, since r is at most order 3 in magnitude, we can regard it as a constant such that $b(r) = 0.6629 \sim 2/3$. The same goes for the fraction parameter,α; we can regard it as a constant such that $\alpha(r) = -0.0345$. For odd parameter fit statistics see Figure A.2a; (b) Plots of A_0 vs r (left) and a vs r (right). Both exhibit a similar pattern, however it is difficult to discern any nice relationships. For odd parameter plots see Figure A.2b.

The first plot in Figure 10.(a) in particular evinces a sinusoidal fluctuation about the mean.

This again indicates the possibility of refining the plots by adding an extra function.

2.2 Analysis of $h^{1,1} + h^{1,2}$

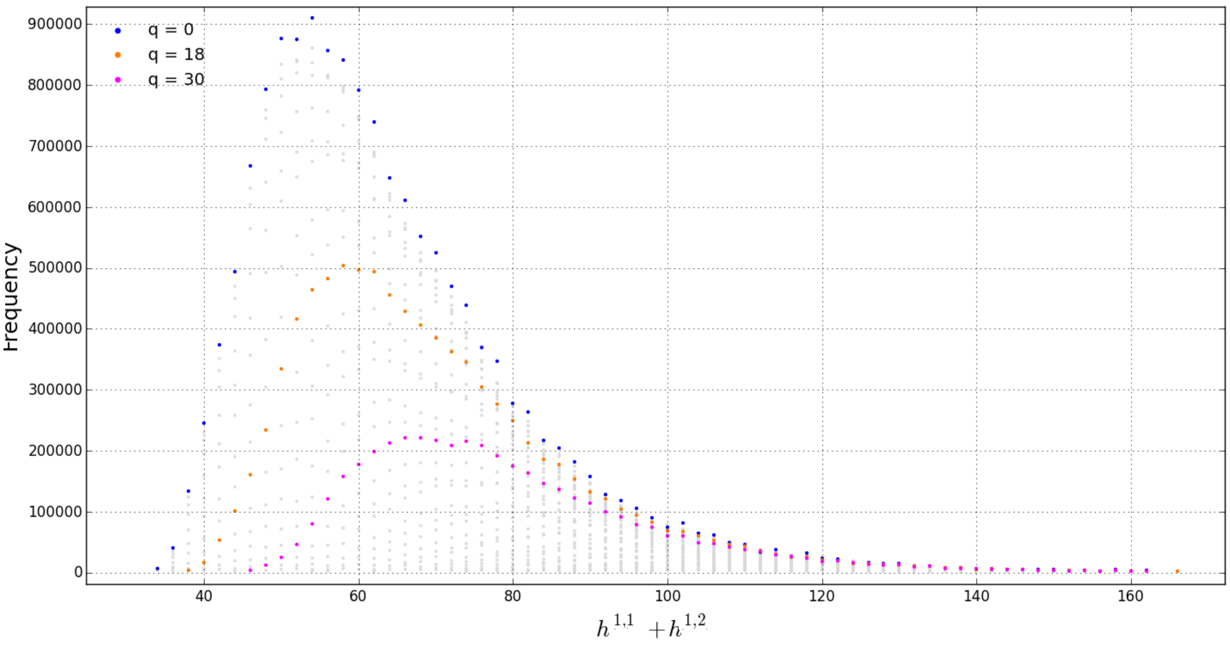


Figure 11: Three curves ($q = 0, 18, 30$) within the even $h^{1,1} + h^{1,2}$ distribution. The transparent grey data dots are all the data plots for the distribution. Refer to Figure A.2 to see the same example for the classification of odd curves within the odd distribution.

We begin by classifying the curves within the $h^{1,1} + h^{1,2}$ distribution (Figure 2) in an analogous way to how it was explained before. This time, we order the data by $h^{1,1} - h^{1,2}$ such that a single curve within $h^{1,1} + h^{1,2}$ can be identified by its q-value, where $q = h^{1,1} - h^{1,2}$. Due to mirror symmetry, the curve for $q = -a$ is the same curve as $q = a$, thus within our two-dimensional plots will only have $q > 0$. In continuation to the analysis on $h^{1,1} - h^{1,2}$, we use a cutoff frequency of 2000 and only present results from the even distribution within $h^{1,1} + h^{1,2}$, unless stated otherwise. As an example, illustrating the classification of curves within $h^{1,1} + h^{1,2}$, consider the curves $q = 0, 18, 30$ in Figure 11.

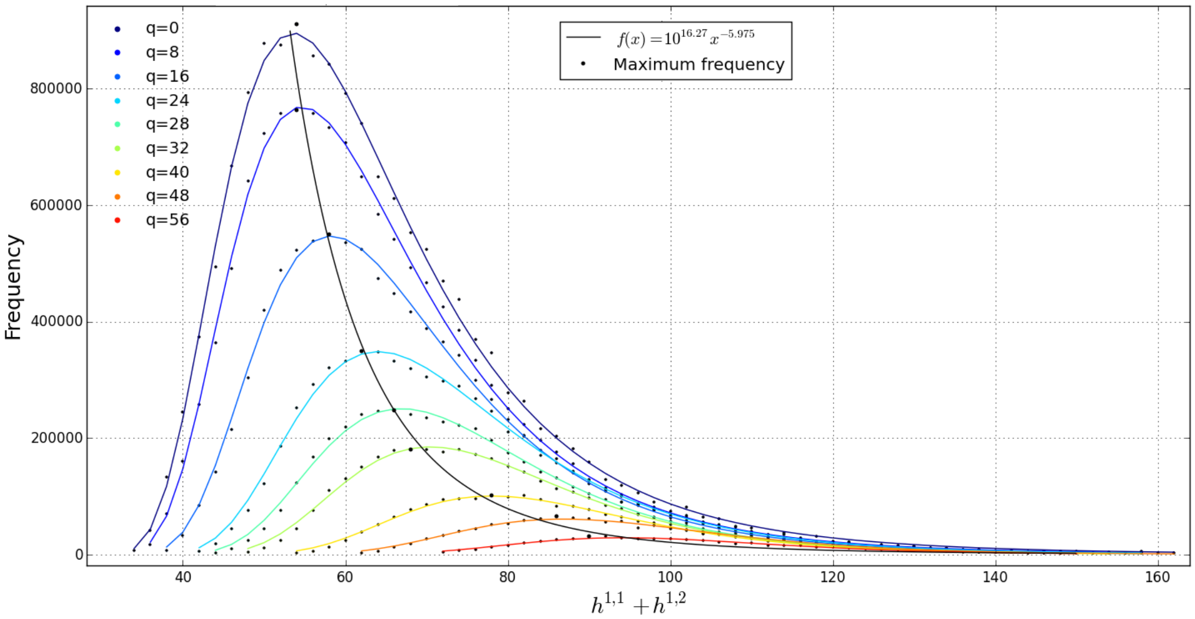
2.2.1 A Planckian Fit

Each curve within the $h^{1,1} + h^{1,2}$ distribution behaves the same. Just like in the $h^{1,1} - h^{1,2}$ distribution, we do a regression analysis for each curve within the distribution independently, in the quest to describe the entire $h^{1,1} + h^{1,2}$ with a single function. The model we chose to describe $h^{1,1} + h^{1,2}$ is the simplest possible Planckian model

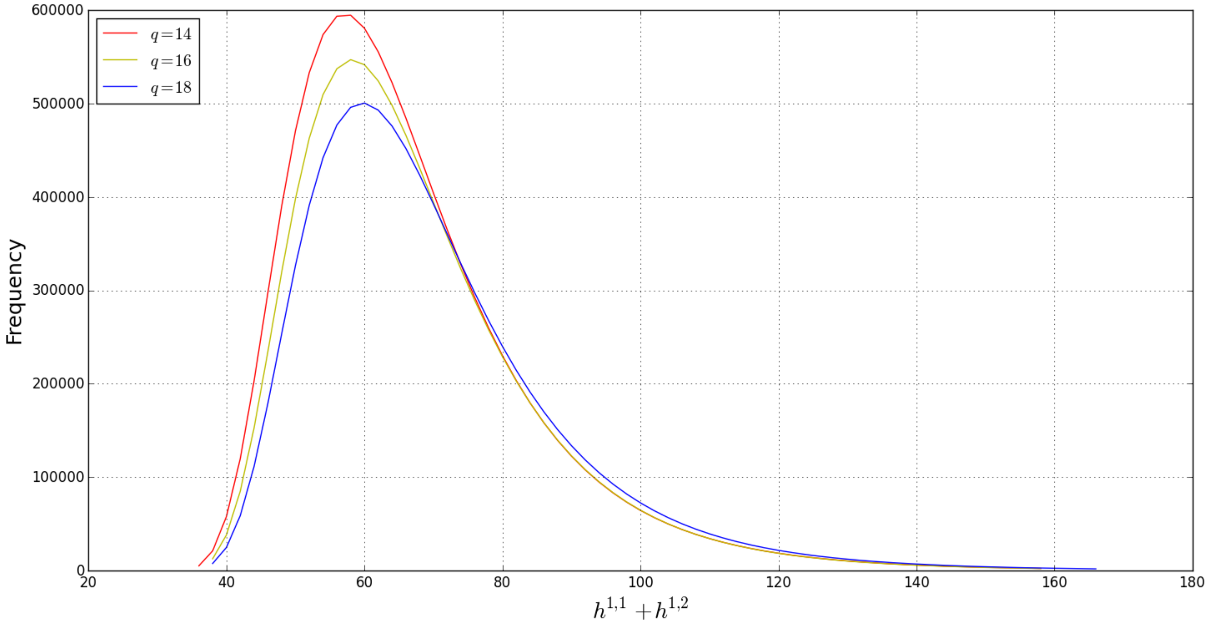
$$f(x, A, n, b) = \frac{A}{x^n} \frac{1}{e^{b/(x-22)} - 1} \tag{2.6}$$

The parameters are the amplitude A, the power n, and some real constant b. The shift in x-axis is so that the distribution begins at 0 as the smallest $h^{1,1} + h^{1,2}$ above the cutoff is 22. The choice of a Planckian model in the above form is greatly motivated by the blackbody distribution $f(T, \lambda)$. The q curves within $h^{1,1} + h^{1,2}$ appear to behave in a manner analogous to the curves of constant T within the blackbody distribution. This is an initial trial. Later, we will discover additional structure in the distribution by trying to mimic the blackbody distribution exactly. It turns out that the general behavior of the distribution is modeled very well, *cf.* Figure 12a.

Consider the maximum of each of the curves. As indicated in Figure 12a, we can fit the maxima to a curve as indicated using the data plotted for the given values of q. From the above analysis, the $h^{1,1} + h^{1,2}$ distribution behaves analogously to a blackbody spectrum — except for one small subtlety. It is in this subtlety that the added structure within $h^{1,1} + h^{1,2}$ is observed.



(a) Lines of best fit from a regression analysis for a few select curves. The black data points represent the maximum frequency for that particular q-curve. the Black line is a line of best fit to describe the points of maximum frequency — this is analogous to a blackbody spectrum. See Figure A.3a for the curves within the odd distribution.



(a) The curves segregate into three classes determined by the value of the even integer modulo 6. A similar pattern occurs in the odd distribution; see Figure A.2a.

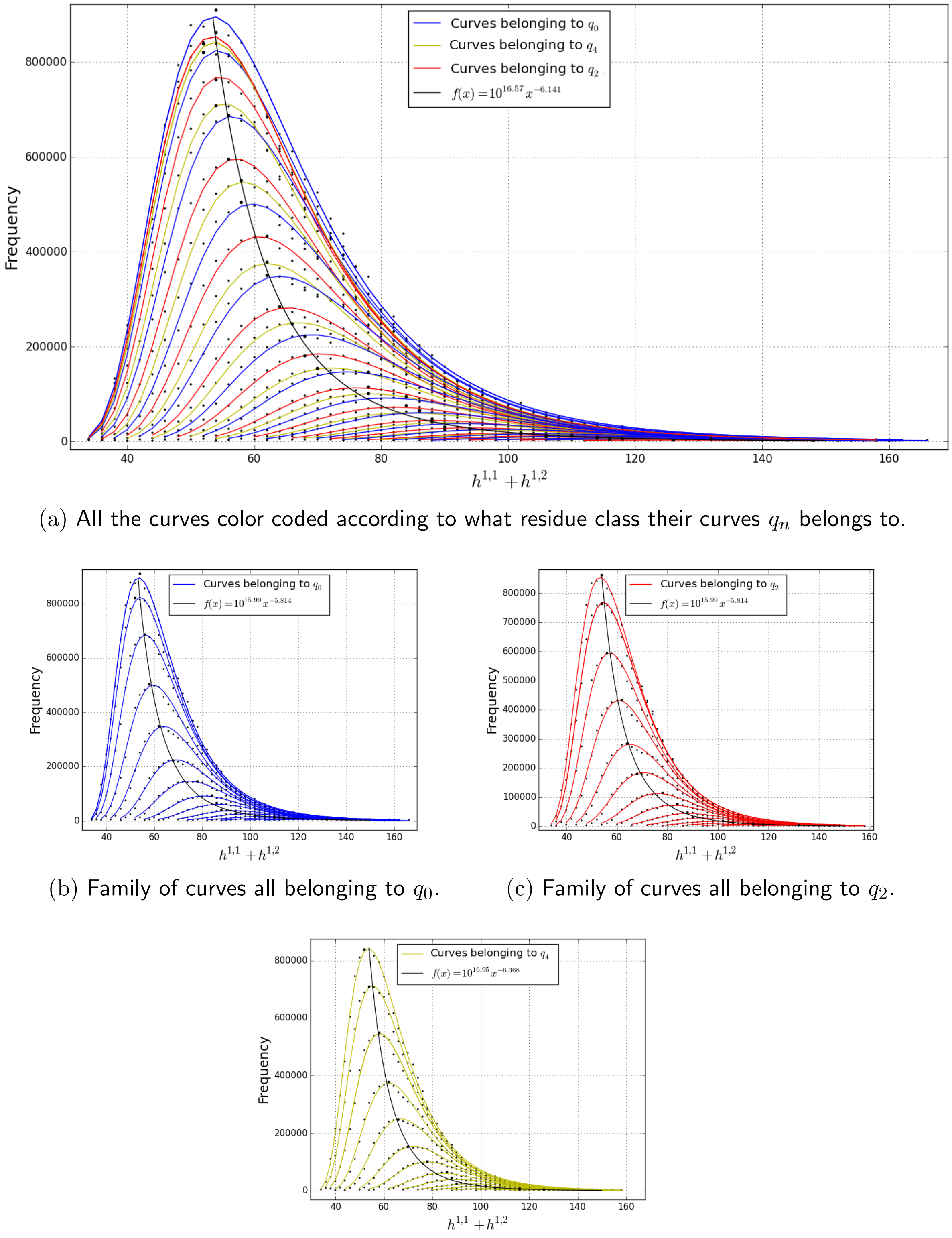
Figure 11: In the attempt to describe the data analogously to a blackbody distribution (a), we discover some subtle structure (b).

Just as was seen in Figure 2, $h^{1,1}+h^{1,2}$ appears to split up into two smaller distributions based on the parity of $h^{1,1}+h^{1,2}$. One can then further break up both the even and odd distributions into three further sets. The manner we observed this added fine structure is again motivated by a blackbody spectrum. In a true blackbody distribution, the curves of constant T never overlap. However, if you consider the lines of best fit only, when looking at our distribution one sees an overlap of certain curves. For example, observe the following plot of curves which clearly cross in Figure 11a.

It turns out that this overlapping occurs consistently to the point where one can classify the curves (defined by their q value) into residue classes q_n distinguished by n mod 6. On the left hand side of the $h^{1,1}+h^{1,2}$ axis, the curves are ordered with red (residue class q_2) above yellow (residue class q_4) above blue (residue class q_0), whereas on the right hand side of the axis, the order is reversed. Similar behavior is observed in the odd distribution of $h^{1,1}+h^{1,2}$ with the curves in the residue classes q_1, q_3, and q_5 (see Figure A.2a).

The clusters of curves constitute an entire set of mod 6 residue classes. These classes now define a set of curves which belong to a very "nice" distributions that behave exactly like a blackbody distribution.[1] Compare, for example, a plot of the all the curves for even distribution of $h^{1,1}+h^{1,2}$, separated into their residue classes, Figure 12

[1] Of course $h^{1,1}+h^{1,2}$ is not continuous. It is discrete. However, the structure of the best fit curve to the data points appears very similar to that of a continuous blackbody distribution.



(a) All the curves color coded according to what residue class their curves q_n belongs to.

(b) Family of curves all belonging to q_0.

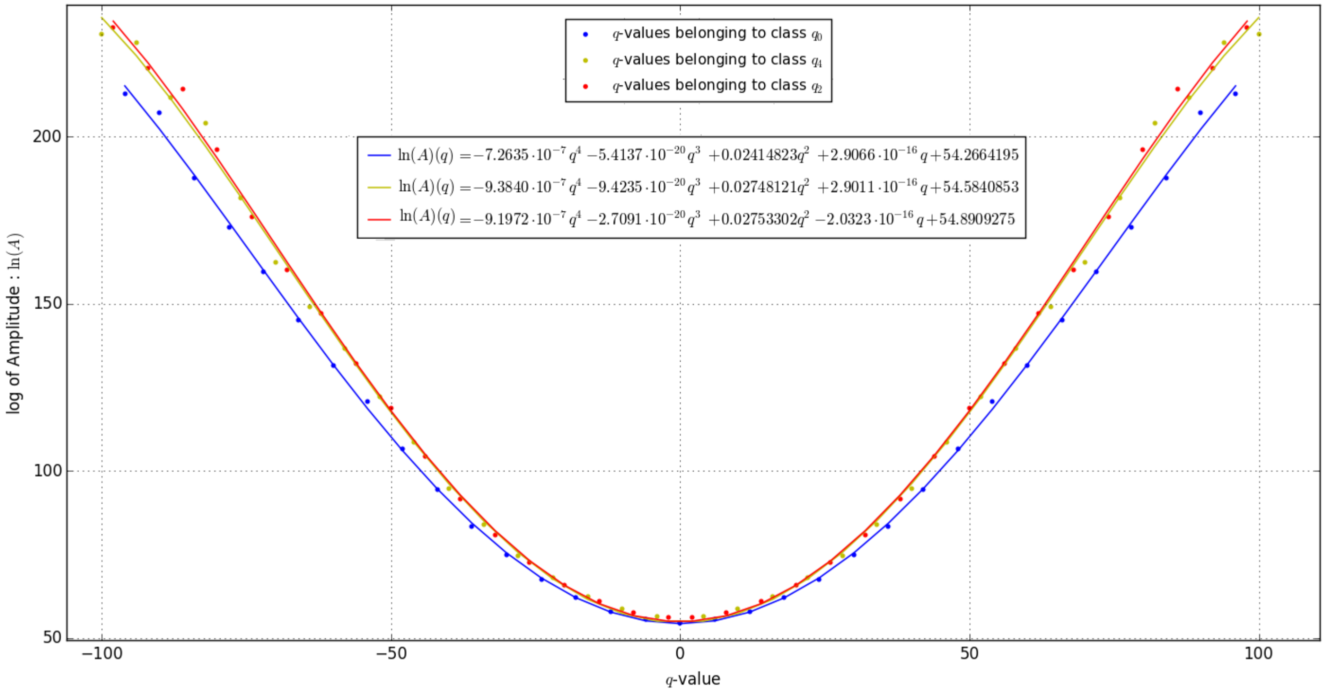
(c) Family of curves all belonging to q_2.

(d) Family of curves all belonging to q_4.

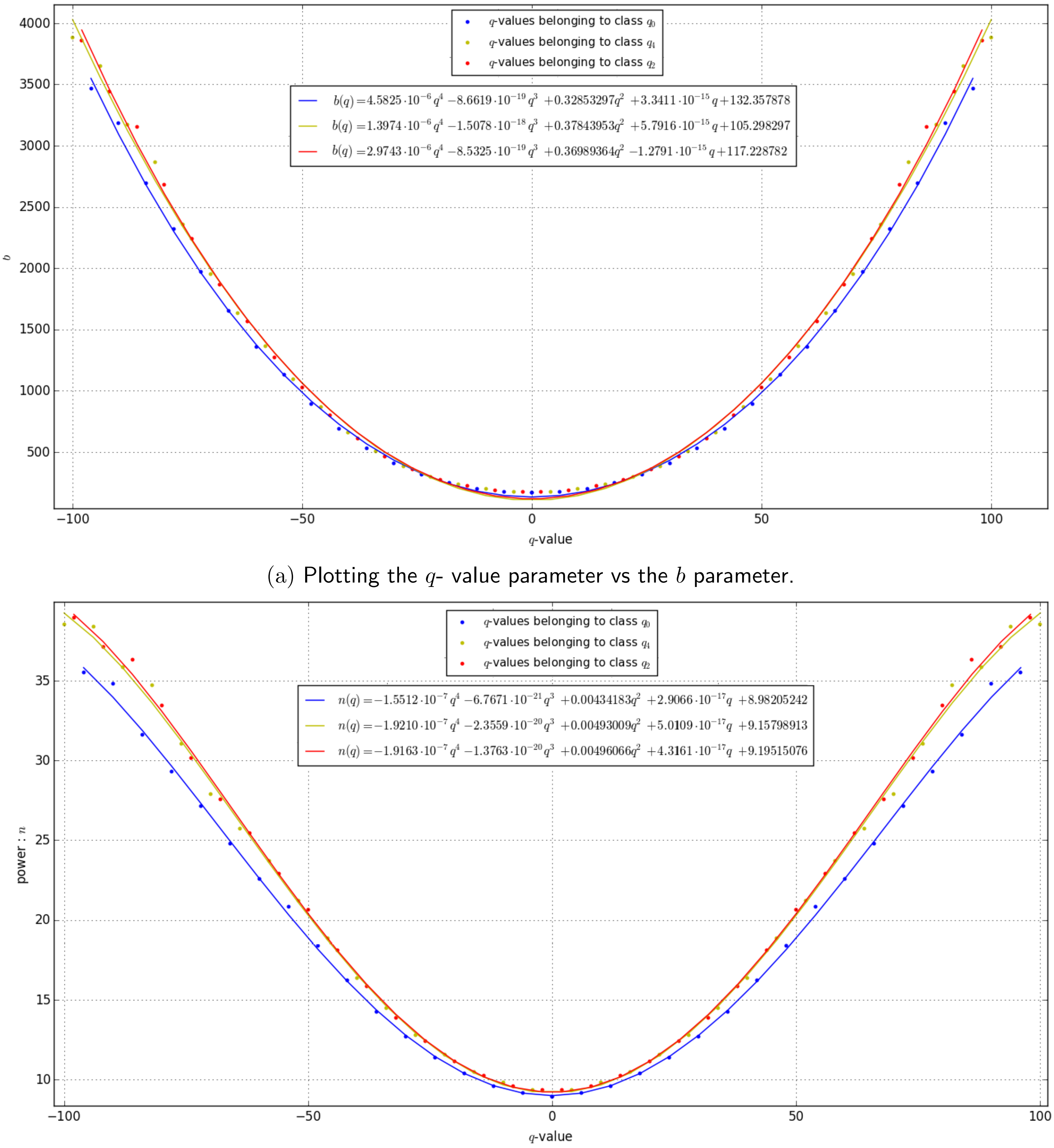
Figure 12: We illustrate the added structure for even $h^{1,1}+h^{1,2}$ data, by displaying how the regression curves can be divided into residue classes. For the list of odd curves, refer to Figure A.2.

As a first approximation we have successfully modeled the general trend of the data. There is, however, a fine structure to the individual data points that we would like to model. Introducing an oscillating term in the amplitude, as seen in the analysis of $h^{1,1} - h^{1,2}$, unfortunately did not seem to improve the fits.
Again, it appears that the least number of variables our functions can have is two, $f = f(x, q)$. This function will be slightly different in the values of coefficients, depending on which residue class one is modeling.
Just as for $h^{1,1} - h^{1,2}$, we wish to express the parameters for the $h^{1,1} + h^{1,2}$ model (2.6) in terms of q. We therefore write $A = A(q)$, $b = b(q)$, $n = n(q)$ and seek to find expressions for the coefficients.

While the x-axis of $h^{1,1} + h^{1,2}$ has only positive q values — due to the fact the data points will overlap — when plotting them against the parameter values, we also have to consider the negative values of q. We present the various relationships (see Figure A.2 for the plots for the odd distribution of $h^{1,1} + h^{1,2}$ analogous to Figure 12).



(a) Plotting the q- value parameter vs the $\log(A)$ parameter.



(a) Plotting the q- value parameter vs the b parameter.

(b) Plotting the q- value parameter vs the power n parameter.

Figure 12: The parameter plots are color coded according to what residue class their q value belong to.

Each distribution has an equation with different parameter values. However, the fact that we can express all the parameters in terms of q means we are able to get a generalized formula to describe the entire $h^{1,1}+h^{1,2}$ distribution — as long as the frequency is above 2000. For succinctness we use the following notation for the coefficients

$$A_{k,i}, \quad n_{k,i}, \quad b_{k,i} \ , \tag{2.7}$$

where the subscript $k = 0,1,2,3,4,5$ refers to residue class q_k, and $i = 0,1,2,3,4$ refers to the coefficient of the i^{th} power of q. Thus, we have:

$$A_k(q) = \exp(\sum_{i=0}^{4} A_{k,i}q^i) \ , \quad n_k(q) = \sum_{i=0}^{4} n_{k,i}q^i \ , \quad b_k(q) = \sum_{i=0}^{4} b_{k,i}q^i \ , \tag{2.8}$$

where the matrix of coefficient values for $A_{k,i}, n_{k,i}$ and $b_{k,i}$ can be found in Appendix A.2.2.[2] Our function (2.6) now is able to approximately describe the entire $h^{1,1}+h^{1,2}$ distribution:

$$f_k(x,q) = \frac{e^{\sum_{i=0}^{4} A_{k,i}q^i}}{x^{\sum_{i=0}^{4} n_{k,i}q^i}} \frac{1}{\left(e^{\frac{\sum_{i=0}^{4} b_{k,i}q^i}{(x-22)}} - 1\right)}, \tag{2.9}$$

Of course there are certain constraints on the values of q. For a given k, q has to be an integer which falls within the residue class q_k. For even values of k, $x = 2m$, and for odd k, $x = 2m+1$. We have $m > 12$.

A few comments about the analysis on the $h^{1,1}+h^{1,2}$ distribution are in order.

1. The Planckian model used in (2.6) could be modified in some manner such that there is some oscillating behavior in the amplitude. Any kind of oscillatory term we introduce, only has a mild effect on the model's behavior. As the q values exceed 100, the model is not able to describe the data very well.

2. Assuming one adds an oscillatory component to the model, the module used in python to do the regression analysis called *lmfit* is sensitive to the initial conditions set by the user. Since the model is a custom model, it is difficult to find the correct initial conditions such that the best fit line oscillates close to every point (as with $h^{1,1}-h^{1,2}$).

3. It is possible that the model used does not have the features required to describe the oscillatory "up and down" behavior of the data points. The Planckian model was chosen in that the $h^{1,1}+h^{1,2}$ distribution resembled a blackbody distribution.

4. In choosing a polynomial model for Figures 13a,12a,12b, we picked the lowest order polynomial that gave the best fit. Choosing the order to be four for all the plots appeared to be convenient.

[2] Perhaps it is important to state explicitly — due to potential confusion — that the coefficients $A_{k,i}$ refers to the natural logarithm of the amplitude values while A_k is the actual amplitude seen in the model.

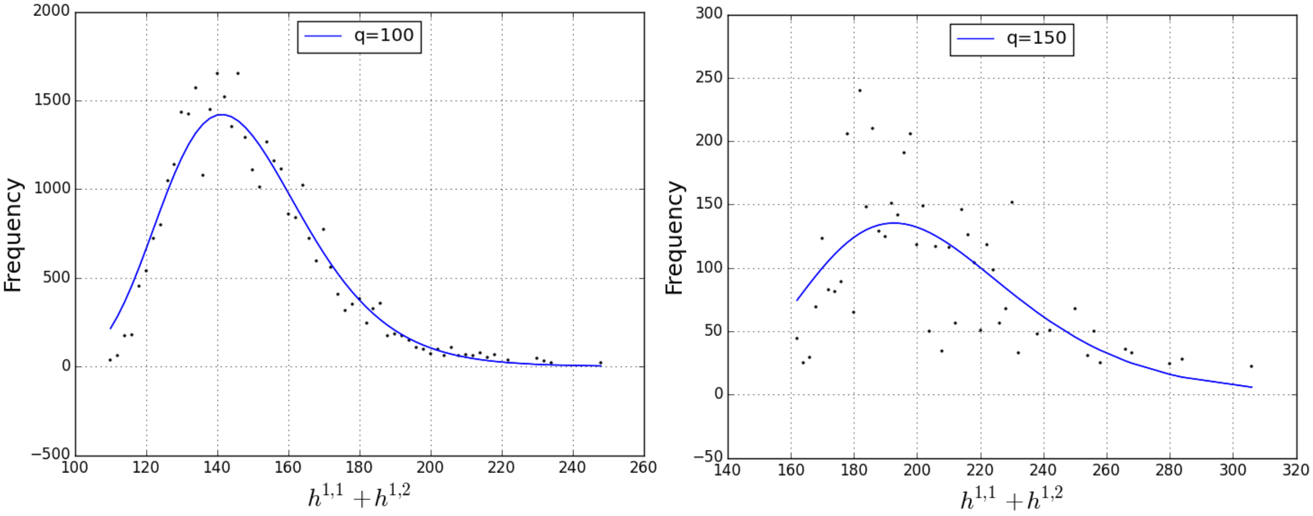


Figure 13: Left figure is the fitted model(blue line) for a q value of 100 and right has a q value of 150. As the q-value increases, the scattering of the data points within $h^{1,1}+h^{1,2}$ increases to the point where the model works no longer. For an example of how the model begins to break down at large q, see Figure A.3.

However, it is apparent that the parameter relationship plot in Figure 12a would be better described by a polynomial of order 6. One could use an order 6 polynomial for all the other relationships plots too, but doing so might not have any physical significance. One can achieve an arbitrarily good fit the larger the order of the polynomial used, but that does not necessarily mean the chosen model is the correct model.

2.3 The Distribution of the Euler Number

The Euler number for Calabi–Yau threefolds is

$$\chi = 2(h^{1,1} - h^{1,2}) \ . \tag{2.10}$$

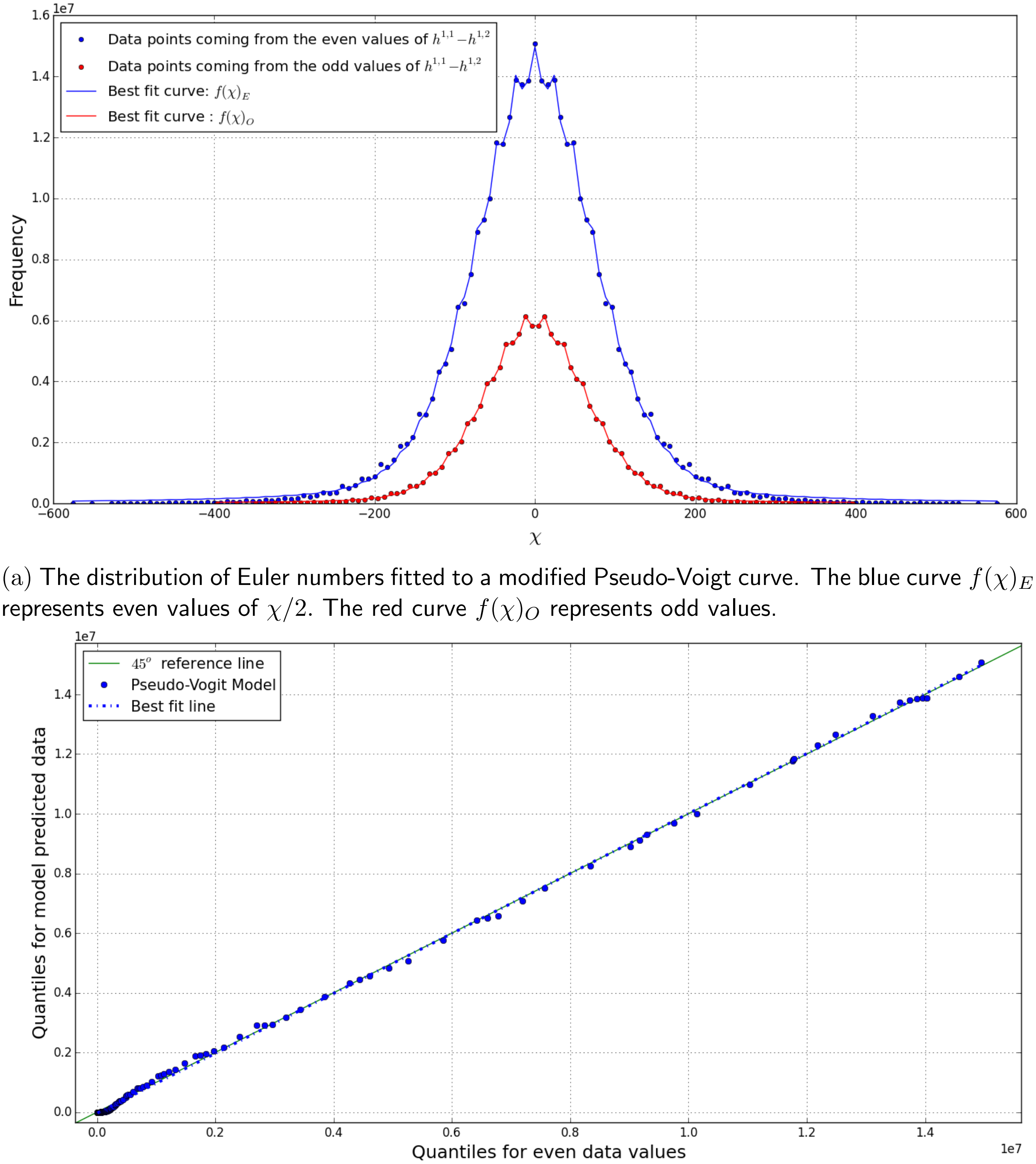
As mentioned previously, we are summing over all the various r-curves to obtain the full-Euler number distribution. A plot of χ versus frequency yields the Pseudo-Voigt distribution. In particular, we can model the behavior of the distribution almost perfectly using the modified Pseudo-Voigt curve (2.5) and (2.12), which is repeated here for convenience:

$$f(x, A, \sigma, \alpha) = (1-\alpha)\frac{A}{\sigma\sqrt{2\pi}}e^{\frac{-(x)^2}{2\sigma^2}} + \alpha\frac{A}{\pi}\left[\frac{\sigma^2}{x^2+\sigma^2}\right] \ , \tag{2.11}$$

where

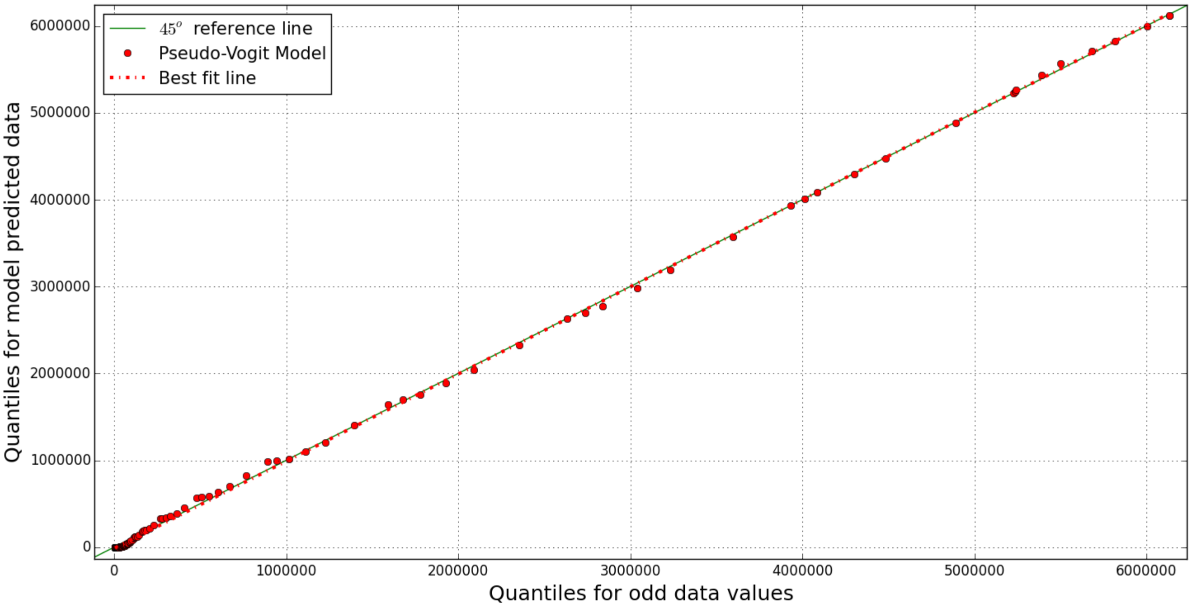
$$A(x, A_0, a, b) = A_0 + a\cos(2\pi b \cdot x)\ . \tag{2.12}$$

The results of the regression analysis for the Euler number distribution is presented in Figure 14a.



(a) The distribution of Euler numbers fitted to a modified Pseudo-Voigt curve. The blue curve $f(\chi)_E$ represents even values of $\chi/2$. The red curve $f(\chi)_O$ represents odd values.

(b) Probability plot for the even values of $\chi/2$. The model fits the data with $R^2 = 0.99944$.



(a) Probability plot for the odd values of $\chi/2$. The model fits the data with $R^2 = 0.99965$.

Figure 13: Various plots illustrating the actual fit of the modified pseudo-Voigt model. We can tell we have a good fit by looking at the probability plots for the quantiles of the standard pseudo-Voigt distribution vs quantiles for the actual data. The. The R^2 values in (b) and (c) are given relative to the line $y = x$.

The fitted parameter values for $f(\chi)_E$ corresponding to even values of $h^{1,1} - h^{1,2}$ are:

$$(A_0, \sigma, \alpha, b, a) = (1.9032 \times 10^9, 75.8305889, 0.00718459, 0.58347826, 8.7427 \times 10^7) \ . \tag{2.13}$$

Likewise, the fitted parameter values for $f(\chi)_O$ corresponding to odd values of $h^{1,1} - h^{1,2}$ are:

$$(A_0, \sigma, \alpha, b, a) = (7.6043 \times 10^8, 64.9735680, 0.00549425, 0.83357720, 3.6881 \times 10^7) \ . \tag{2.14}$$

Although χ is only even, the two curves originate from the fact that if you take $\chi/2$ you get even and odd values. The two curves arise from the parity of $\chi/2$ and are presented in Figures 14a.

2.4 Goodness-of-fit

A goodness-of-fit test is implemented as a means of testing how well a given model describes some given data. Typically the model validation process consists of only quoting a single statistically generated number like the R^2, χ^2 or p values. Based on the size of this number, one then makes inferences on how well the chosen model fits the observation. One needs to be careful however of misusing such indicators as an absolute measure for assessing goodness-of-fit.

For a structural equation model (SEM) — in our case, the modified pseudo-Voigt and Planckian models — this assessment is not so straight forward as it would be for a simple regression analysis.

To quantify the predictive power of an SEM, a single statistical test does not suffice - in fact, there is no single test. According to [37], the best one can do is assess three different aspects of what it means to have a good fit, these are: overall fit, comparative fits to a test model and model parsimony.[3] The only real test available is the chi-squared (χ^2) test, when it comes to overall fit, this χ^2 statistic is the most popular test. The χ^2 test compares observed and predicted correlation matrices with each other, and so, statistical significance is evaluated based on the value of χ^2. A large χ^2 value signifies a considerable difference between the correlation matrices. A low value indicates there is little statistical difference between matrices. Since the χ^2 test is between actual and predicted matrices only, when looking for overall fit, one searches for non-significant differences between the correlation matrices. Often, rather than presenting the χ^2 or χ^2_R (the chi-squared value relative to the degrees of freedom for the model) value, a p value is given instead. The p value, in a way, informs us whether one should reject a null hypothesis or not. A rejection of the null hypothesis (p value lower than the significance level) means that differences in the "observed vs. predicted" are too significantly large to deduce that the chosen model fits the data. The p-values can be determined by a p-value calculator by inputting the χ^2_R value. There is no standard way of choosing a significance level for the p-value, but typically $p < 0.05$ is considered statistically significant.

In general, statistical non-significance given by appropriate values of the χ^2 fit statistics is adequate. However, one must be careful of drawing similar conclusions for structural equation modeling. The fit statistic makes a statement of the correlation matrices only, not about whether or not the correct model is identified. This is largely due to the sensitivity to sample size of the χ^2 test. In our analysis, the sample size (number of reflexive polytopes) is enormous — almost one billion! For large samples (> 200) the χ^2 test will give significant differences for any model used. This sensitivity to a sample size, together with an *effect size* and *alpha value*, is related to what one calls the power of a test - the probability of not incorrectly accepting a null hypothesis that is actually false.

Without worrying too much about what an effect size and alpha value is; for any alpha value, the greater the sample size, the greater the power of the statistical test. However, increasing the sample size beyond a certain amount, can result in the test having "too much" power. Perceived effects in very large sample sizes, will always become significant[4]. Observe how in tables A.0 and A.4 the χ^2_R values for all the different curves is extremely large, naively indicating that we have a horrible fit — which would be an incorrect conclusion.

It is clear from the above discussion that we cannot use the χ^2 or p values in validating our choice in model. What is not so clear, is the additional subtlety in using purely statistical means to asses goodness-of-fit for our data. This subtlety lies at the heart of almost all statistical tests —

[3] Parsimony refers to the ability of a model to give a certain degree of fit whilst having the least required number of predictor variables.

[4] Conversely is also true, for extremely small sample sizes, any effect which should be significant, becomes insignificant

the construction of a null hypothesis. The term frequency, as used in the statistical sense, refers to the number of outcomes for a certain event. The measurement of this outcome will often have certain known or unknown factors affecting it. These tests check for the probability that the errors found are too significant to be solely do to random variations in the data. For example, assume that statistical tests give non-significant results. If the residuals are small enough to be considered random errors in the measurement of the frequency, we could say that the model is appropriate. If however, the residuals are too large or present additional structure, we could say the model is good, but not quite the correct one as the residual errors are not "random enough". In our case, there is no notion of measured frequency and error in measurement of frequencies. Our frequencies are generated as a result of a combinatoric calculation. Statistical tests assume that the input is from measurement and observations (obeying some null-hypothesis), thus they are inherently constructed with this notion in mind. By inputting our data, the tests are trying to calculate something from a data set which does not obey the very assumption they use in their calculations. We are not exactly clear how much this affects statistical outcomes, but it is important to keep in mind.

How do we validate then, that our chosen models are a good fit, or that our model is the best one at describing the data? We implement graphical methods. The first graphical method is obviously through pure inspection — this is not quite statistically quantifiable. There is a statistically based graphical method to asses goodness-of-fit called probability plots, Q-Q plots or P-P[5] plots. These plots were initially constructed to test the "normality" of a data set when the sample size is too large too depend on the χ^2 and p values. In principle, a standard probability plot tells you the likelihood that the a sample's distribution of data obeys a normal distribution — hence checking for normality. The answer to the question is not given by a statistical value, but rather by a graphical representation — from which one can extract statistical numbers. If the plotted data on this probability plot is a straight line, then we can determine that the sample set is normally distributed.

We can extend this concept further: we can take two different samples, and take a probability plot to determine if two data sets come from populations with a common distribution. Such a probability plot is referred to as a Q-Q (quantile-quantile) plot. Extending this concept one more time — as for our use — we will take the quantiles of our theoretical distribution (the modified pseudo-Voigt and Planckian profiles) as our "first sample" and plot them against the quantiles of our data as our "second sample", this will give us our probability plot. In all the probability plots, it is the quantiles of the respective data sets which are plotted against each other.

Quantiles are basically just a generalization of quartiles. For example, the k^{th} percentile of a set of values divides them, such that the number of values which lie below is $k\%$, and the number of values which lie above is $(100 - k)\%$. The 25th percentile is the lower quartile or the $\frac{1}{4}$ quantile. Quantiles are the same as percentiles, but indexed by sample fractions rather than by sample

[5] A P-P plot is the plot of the cumulative distribution frequency of the one data set against the CDF of the other. P-P plots are not as useful as Q-Q plots, thus are seldom used.

percentages. Suppose that $p \in [0, 1]$, the aim is to find the value that is the fraction p of the way through the ordered data set. As an example, if $p = \frac{1}{2} = 0.5$, we want to know what is the value that sits at $p = 0.5$ of the way through i.e. half way. The value that sits there (this value may have to be interpolated) will be called the quantile for the fraction $p = 0.5$. There are many different algorithms for generating the quantiles for a given data set, we use python to generate the quantiles in a manner similar to that discussed above. For an ordered data set, $x_1 \leq x_2 \leq x_1 \ldots \leq x_{n-1} \leq x_n$, the most common way of calculating quantiles is to first compute the empirical distribution function:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} = 1(x_i \leq x), \quad x \in \mathcal{R}, \tag{2.15}$$

and then define the quantile function to be the inverse of $F(x)$:

$$F^{-1}(p) = min\{x \in \mathcal{R} : F(x) \geq p,\ p \in (0, 1)\}. \tag{2.16}$$

By generating the quantiles of some theoretical model and comparing them to the quantiles of a given data set of equal length, one can determine if the data set belongs to the same distribution as the data set belonging to the theoretical model — *i.e.*, does the data fit the model. If the quantiles are roughly equal the plots will all be more or less on a straight line.

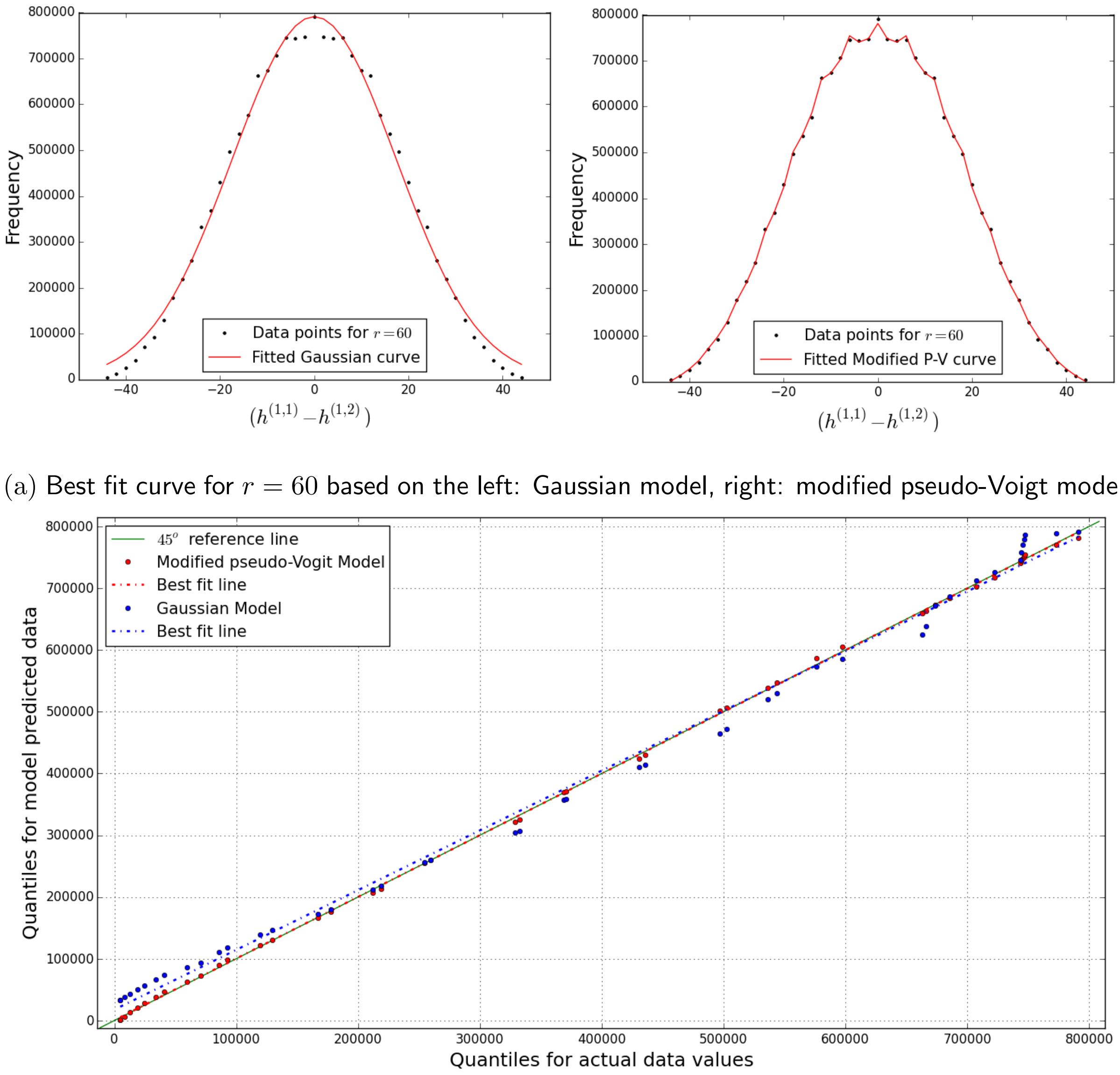
In probability plots :

1. The length of data set needs to be equal. For unequal lengths, one must perform an interpolation of data.

2. If two identical data sets were compared to one another, the points would lie exactly on a 45 degree line. Thus, for two different data sets, the deviation from this reference line determines the likelihood that the sets belong to similar distributions. To quantify this likelihood, one can calculate the R^2-value of the data, relative to the $y = x$ reference line.

3. Q-Q plots are not only limited to determining similarity in data sets. By analyzing the deviations which occur, one can determine how the scale and location of the data is shifted - the data would follow some line $y = mx + c$, where m, c would be the estimates of these shifts in scale and location. Also, from the distribution of points above or below the reference line, one can infer aspects of the tails and skewness in the data.

Consider the following curves for the $h^{1,1} - h^{1,2}$ distribution with $r = 60$ in Figures 14a and 14b.



(a) Best fit curve for $r = 60$ based on the left: Gaussian model, right: modified pseudo-Voigt model.

(b) Probability plot for Figure 14a. The x-axis represents the quantiles for the actual data, the y-axis represents the theoretically predicted quantiles — dependent on the model chosen (red: modified pseudo-Voigt model ($R^2 = 0.99974$); blue: Gaussian model ($R^2 = 0.99334$). The R^2 values are not relative to the best fit lines, but are relative to the $45°$ reference line $y = x$. The closer the R^2 value is to 1, the more similar the predicted quantiles are to the actual ones, thus, the better the model describes the data.

Figure 14: Using probability plots, we are able to statistically see which model provides the better fit. We employ such graphical methods as standard goodness-of-fit tests such as the χ^2 fail to give meaningful results.

For the $h^{1,1} + h^{1,2}$ distribution we just plot the data of $q = 2$ together with the corresponding probability plot in Figure 15.

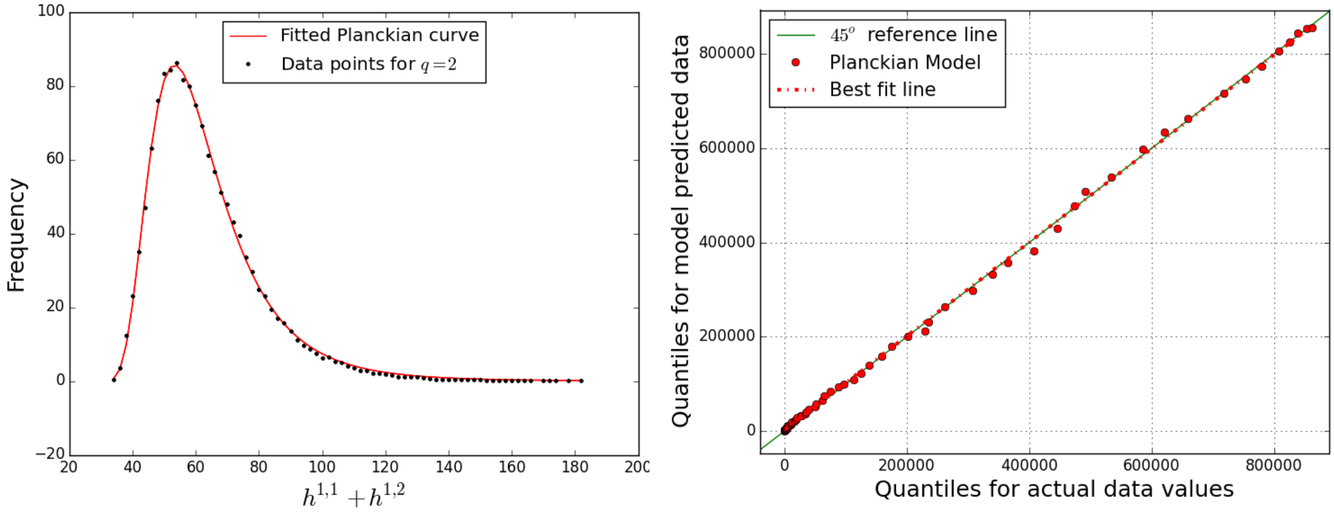


Figure 15: Left: best fit curve of $h^{1,1}+h^{1,2}$ distribution for curve $q=2$ based on the Planckian model. Right: probability plots of our fitted theoretical Planck model vs the $q=2$, $h^{1,1}+h^{1,2}$ distribution.

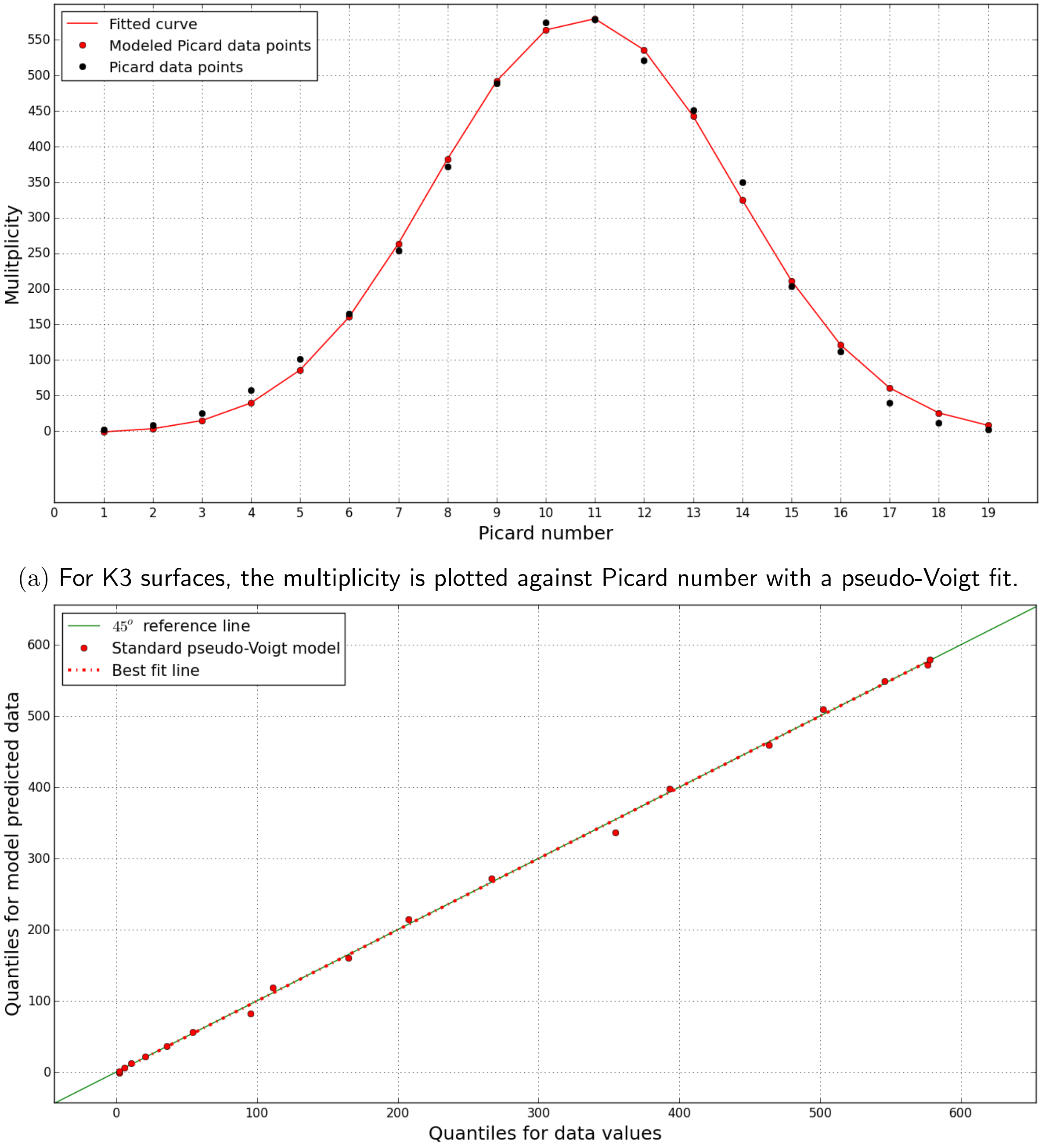
In its current form, the probability plots do not allow us to calculate p-values of the various models. This due to the same issue encountered previously. If one however standardizes the data according to the Z-standardization:

$$Z = \frac{X-\mu}{\sigma}, \tag{2.17}$$

where μ and σ are the mean and standard deviation, it is possible to calculate the p- values since the magnitude of each sample gets rescaled. The probability plot of all the models is displayed in the Appendix, with the relative p-values for each model — Figure A.-3a and Figure A.-4a. What we see is that the modified pseudo-Voigt is statistically the model which provides the best fit.

3 Calabi–Yau Twofolds: K3 Surfaces

As noted in the Introduction, there are 4319 data points, corresponding to hypersurfaces as Calabi–Yau twofolds, *i.e.*, K3 surfaces, in reflexive three dimensional polytopes. Being algebraic K3 surfaces, there is only one relevant topological invariant, the Hodge number, $h^{1,1}=19$. However, there is a further refined algebraic quantity for the K3 surface X, the rank of the Neron–Severi lattice $H^2(X;\mathbb{Z})\cap H^{1,1}(X)$, which is the **Picard Number** $\rho(X)$ and which enumerates the number of divisors on the surface up to algebraic equivalence. The Picard numbers of the 4319 K3 surfaces were computed in [10]. We present the distribution thereof in Figure 16a.



(a) For K3 surfaces, the multiplicity is plotted against Picard number with a pseudo-Voigt fit.

(b) Probability plot for the multiplicity quantiles vs the fitted standard pseudo-Voigt quantiles. The R^2 value is 0.99908.

Figure 16: Using probability plots, we are able to statistically see which model provides the better fit. We employ such graphical methods as standard goodness-of-fit tests, such as the χ^2 test, fail to give meaningful results.

We only used the standard pseudo-Voigt profile as the modified one did not change the fit significantly. Here are the fit statistics for best fit curve: $(A, \mu, \sigma, \alpha) = (4517.45, 10.76, 2.97, -0.031)$, as shown in Figure 16.

What is interesting about Figure 16a is that the "oscillations" of the actual data points above and below the modeled curve is very apparent, yet modifying the pseudo-Voigt profile is unable to give any significant improvement. This leads to two potential conclusions: (a) The pseudo-Voigt profile is not the best profile to use in combination with an oscillatory component; (b) The manner in which the oscillations occur is not so straight forward as introducing a simple cosine function. An interesting exercise would be to superimpose a cosine function along the distribution, by rotating it as one traverses the profile. As long as the wavelength, amplitude and angle of rotation are all small enough, the continuously rotated cosine function should remain a function everywhere along the profile.

4 Calabi–Yau Fourfolds

The analysis of the four fold data is performed in the same spirit as the threefold data. We aim to look for patterns in the frequency plots. Due to complex conjugation and Poincaré duality, the only topological invariants of fourfolds that vary are $h^{1,1}$, $h^{1,2}$, $h^{1,3}$, and $h^{2,2}$. Three of these are independent [13]:

$$h^{2,2} = 44 + 4h^{1,1} - 2h^{1,2} + 4h^{1,3} \ . \tag{4.1}$$

We compiled a database for the frequency of the triplets $(h^{1,1}, h^{1,2}, h^{1,3})$ to then obtain the following data structure

$$(h^{1,1}, h^{1,2}, h^{1,3}, f) \ .$$

Since one expects mirror symmetry within the invariants $(h^{1,1} \pm h^{1,3})$ [36], a plot of $h^{1,1} - h^{1,3}$ against $h^{1,1} + h^{1,3}$ (Figure 17) should be symmetric about the line $h^{1,1} - h^{1,3} = 0$.

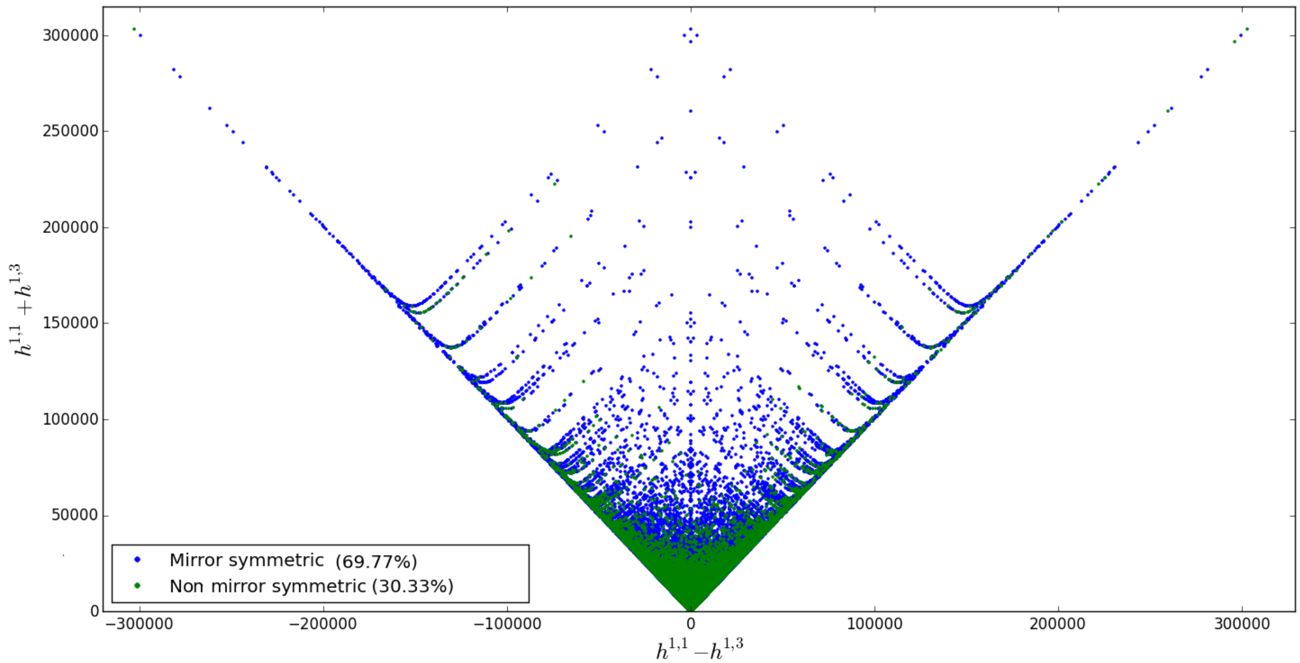


Figure 17: The blue points correspond to manifolds with a mirror symmetric counterpart in the data set.

Doing a quick analysis of the data yields the following observations: only partial mirror symmetry is found. For 69.77% of data points, the point $(h^{1,1}-h^{1,3}, h^{1,1}+h^{1,3})$ is accompanied by the point $(-h^{1,1}+h^{1,3}, h^{1,1}+h^{1,3})$. Taking frequency into account, the percentage drops to 27.35% — see Figure A.5 in the Appendix. This is most likely due to an incomplete data base.

For now, we have performed a primary analysis on the Euler distribution only. The Euler number for fourfolds is [13]:

$$\chi = 6(8 + h^{1,1} - h^{1,2} + h^{1,3}) \ . \tag{4.2}$$

Interestingly enough, the distinction between even and odd distributions persist in the fourfold data base. For illustrative purposes, we show the distribution of $\chi/6$ against frequency.

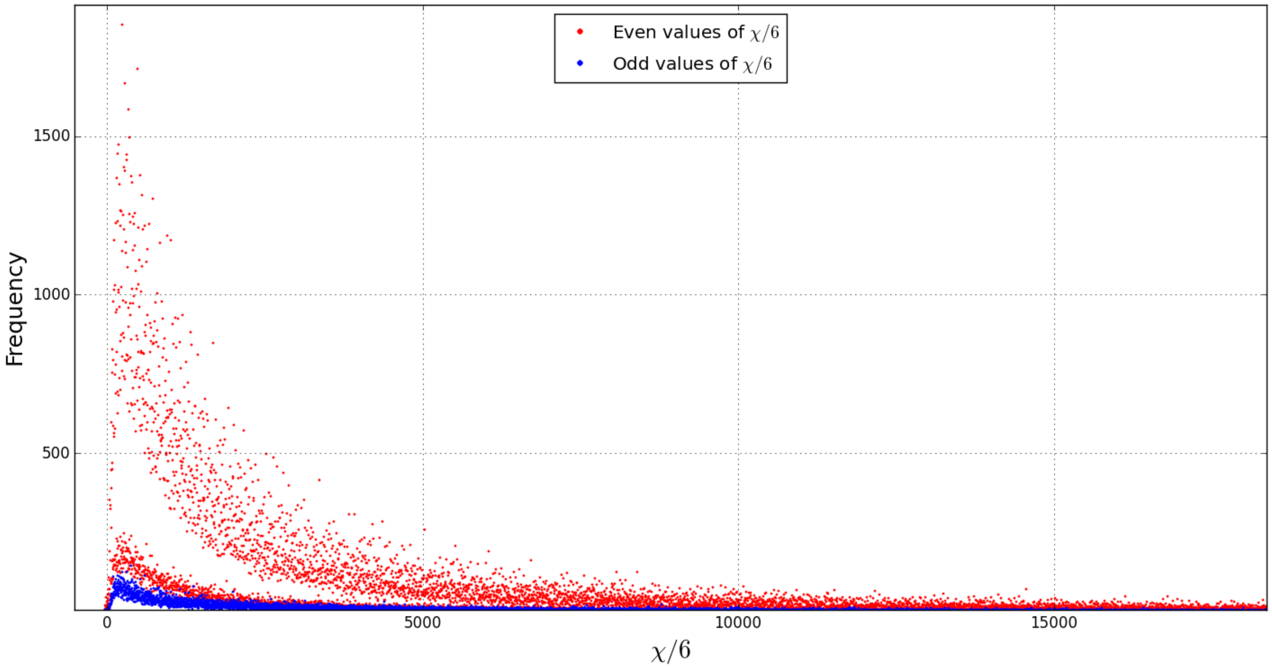
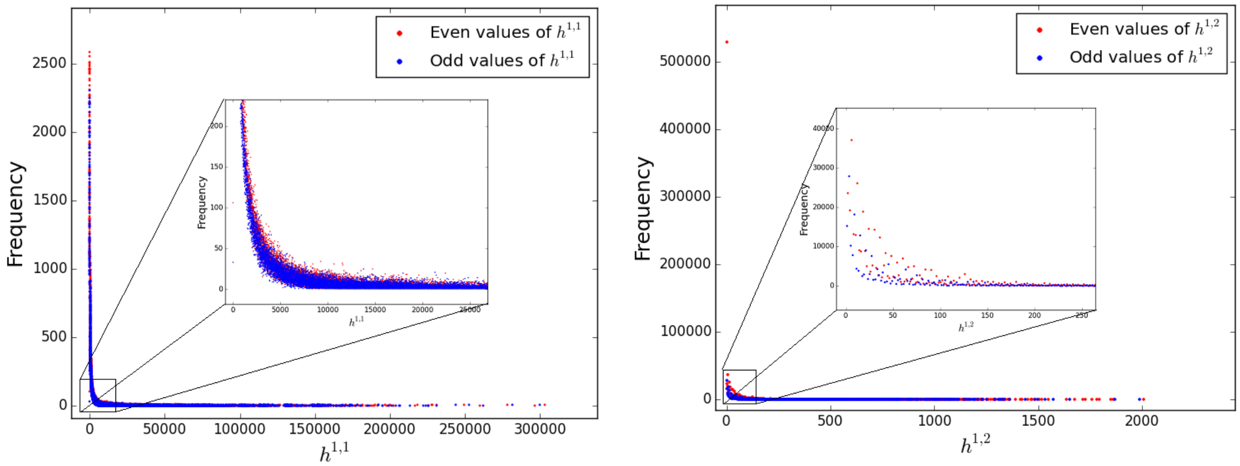


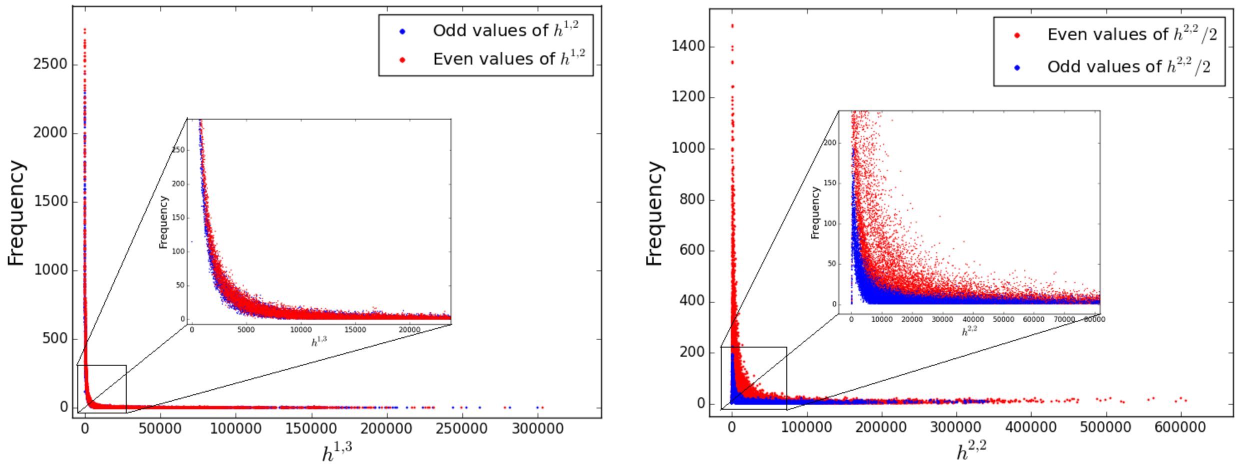
Figure 18: Frequency of Calabi–Yau fourfolds with a given Euler number.

It is not immediately clear what is the reason for the gap, presumably it could be a cluster of data points which is missing from the data base. Until one obtains the complete fourfold data base of Hodge numbers, one can't say much else. We also present plots of the individual Hodge numbers $h^{i,j}$ vs. frequency.



(a) $h^{1,1}$ vs. frequency.

(b) $h^{1,2}$ vs. frequency.



(a) $h^{1,3}$ vs. frequency.

(b) $h^{2,2}$ vs. frequency.

Figure 18: The frequency for all the Hodge $h^{i,j}$ numbers. Red points and blue are odd and even points respectively for the various Hodge numbers. The data points are very dense close to the origin making it difficult to properly illustrate the mixing of odd and even Hodge numbers. Only $h^{2,2}$ (c) has a clear separation between of an even values.

5 Conclusions and Outlook

By examining the distribution of Hodge numbers of Calabi–Yau manifolds of complex dimension two, three and four, realized as hypersurfaces in toric varieties of one higher dimension as constructed by Kreuzer and Skarke based on the results of Batyrev and Borisov, we have found many hithertofore undiscovered patterns. We summarize our key points as follows.

- For threefolds, there are 30108 distinct pairs of Hodge numbers $(h^{1,1}, h^{1,2})$ from 473800776 reflexive polytopes, the frequency of both the half-Euler number $h^{1,1} - h^{1,2}$ and the sum $h^{1,1} + h^{1,2}$ are distributed according to whether the value is odd or even;
  - The half-Euler number $h^{1,1} - h^{1,2}$ follows a modified pseudo-Voigt distribution

    $$f(x) = (1-\alpha)\frac{A'}{\sigma\sqrt{2\pi}}e^{\frac{-(x)^2}{2\sigma^2}} + \alpha\frac{A'}{\pi}\left[\frac{\sigma^2}{x^2+\sigma^2}\right] .$$

    where the modification is made in the amplitude A of the distribution, such that

    $$A' = A_0 + b\cos(2\pi \cdot b) .$$

    There is fine periodic substructure in terms of curves indexed by an integer r. Our model is accurate for low r-values ($r \in [36, 120]$ and $r \in [37, 99]$); using probability plots as test for goodness of fit, this modified pseudo-Voigt model is indeed the best one out of several standard candidates (cf. Figure A.0 for all the R^2 and p values).

Among A, σ, α, b, a, the parameters σ, b, α have a strong linear relationship with r:

$$\begin{array}{rll} & \text{Even } r & \text{Odd } r \\ \sigma(r) = & 0.5097r - 12.7142 & 0.51379r - 13.2494 \\ \alpha(r) = & 2 \times 10^{-4}r - 0.0345 & 2.25 \times 10^{-4}r - 0.0388, \\ b(r) = & 3.7299 \times 10^{-5}r + 0.6629 & 7.9101 \times 10^{-5}r + 0.65956 \end{array}$$

For a small subset of curves with a low r-value and an appropriate cut-off frequency, it is extraordinary that the model *exactly fits the data.* That is, it appears that the number of data points for each curve required, such that the model will result in a perfect fit is: 7 for even r-valued curves and 10 for the odd valued r-curves, see Figure A.1.

- The quantity $h^{1,1} + h^{1,2}$ follows a Planckian distribution

  $$f(x) = \frac{A}{x^n} \frac{1}{e^{b/(x-22)} - 1}$$

  There is a substructure of curves, indexed by an integer q, each Planckian and with some periodic behavior. The curves q_n appear clustered into groups of residue classes distinguished by n mod 6, and the parameters $\log(A), n, b$ all have extremely strong relationships with the q value.

  By substituting this relationship into the model, we have a function $f_k(x, q)$ that approximately describes the entire $h^{1,1} + h^{1,2}$ distribution up to a q value of $69, 100$:

  $$f_k(x, q) = \frac{e^{\sum_{i=0}^{4} A_{k,i} q^i}}{x^{\sum_{i=0}^{4} n_{k,i} q^i}} \frac{1}{\left(e^{\frac{\sum_{i=0}^{4} b_{k,i} q^i}{(x-22)}} - 1 \right)}, \tag{5.1}$$

  with $k = 0, 1, \ldots 5$ and the coefficients given in A.8,A.9,A.10.

- The Euler number $\chi = 2(h^{1,1} - h^{1,2})$ follows the modified Pseudo-Voigt distribution composed with a sinusoidal $A + A_0 + a\cos(2\pi b \cdot x)$ which is almost an exact fit, with the coefficients given by $(A_0, \sigma, \alpha, b, a) = (1.9032 \times 10^9, 75.8305889, 0.00718459, 0.58347826, 8.7427 \times 10^7)$, at $R^2 = 0.99944$ for even χ and
  $(1.9032 \times 10^9, 75.8305889, 0.00718459, 0.58347826, 8.7427 \times 10^7)$ at $R^2 = 0.99965$ for odd χ,

  The modified pseudo-Voigt distribution is remarkably accurate in predicting the overall and fine sub-structure of the Euler number distribution.

- For K3 surfaces, we have looked at the distribution of the multiplicity with Picard number. We find that this distribution follows a standard pseudo-Voigt profile. Adding in the sinusoidal modification does not significantly increase the overall fit. The parameters are given by $(a, \mu, \sigma, \alpha) = (4517.45, 10.76, 2.97, -0.031)$ with $R^2 = 0.99908$.

- For Calabi–Yau fourfolds, there is no exact mirror symmetry, due to incompleteness of available data. Nevertheless, by breaking up the data into three groups, we have

  - Mirror symmetric partners with the same frequency: 27.35%

  - Mirror symmetric partners without the same frequency: 42.22%

  - Non mirror symmetric partners: 30.33%

  By plotting the various $h^{i,j}$ vs frequency we see there is no distinction between even and odd data values for $h^{i,j}$, expect for $h^{2,2}/2$. This distinction is carried out further in the Euler number distribution where odd points are clustered on a band with much lower frequencies. The even values of $\chi/6$ appear to be distributed along to separate bands.

It is remarkable how well the pseudo-Voigt distribution, modified with a sinusoidal component, fits the distribution of topological numbers of toric Calabi–Yau manifolds, often giving an exact fit. Of course, what we are studying at heart is the number of integer points inside (*cf.* (2.1)) reflexive polytopes. This is a highly non-trivial counting problem whose answer will ultimately give full analytic results for our distributions and we suspect that the answer should be some generalized pseudo-Voigt function.

Now, in addition of Calabi–Yau manifolds, stable vector bundles over various such manifolds in a variety of construction beyond Kreuzer–Skarke have also been studied algorithmically over the years in the context of heterotic compactification (*cf. e.g.*, [20–23]). One can see a somewhat pseudo-Voigt profile in these as well, even though there is no underlying polytope and the counting problem is dictated by certain Diophantine system. It would be interesting to see why this shape is universal in such classifications.

Acknowledgements

We are grateful to Cyril Matti for collaboration during the early stages of this project. We thank Mark Dowdeswell for his input with regards to the goodness-of-fits for the various plots. We thank Bruno Le Floch for comments on the manuscript. YHH is indebted to the Science and Technology Facilities Council, UK, for grant ST/J00037X/1, the Chinese Ministry of Education, for a Chang-Jiang Chair Professorship at NanKai University, and the city of Tian-Jin for a Qian-Ren Award. YHH is also perpetually indebted to Merton College, Oxford for continuing to provide a quiet corner of Paradise for musing and contemplations. VJ and LP are supported by the South African

Research Chairs Initiative of the Department of Science and Technology and the National Research Foundation.

A Appendix

Here we include all additional plots to supplement the main body. This includes the relevant plots for the odd distributions — since in the main text we only presented the plots for even distributions — as well as the regression analysis statistics and parameter values for both distributions.

A.1 Supplementary plots for the $h^{1,1} - h^{1,2}$ distribution

All even plot counterparts will be referenced in the figures. The plots appear in the same order as in the main body, with descriptions only if necessary.

A.1.1 Plots for the odd distribution as counterparts to the even ones

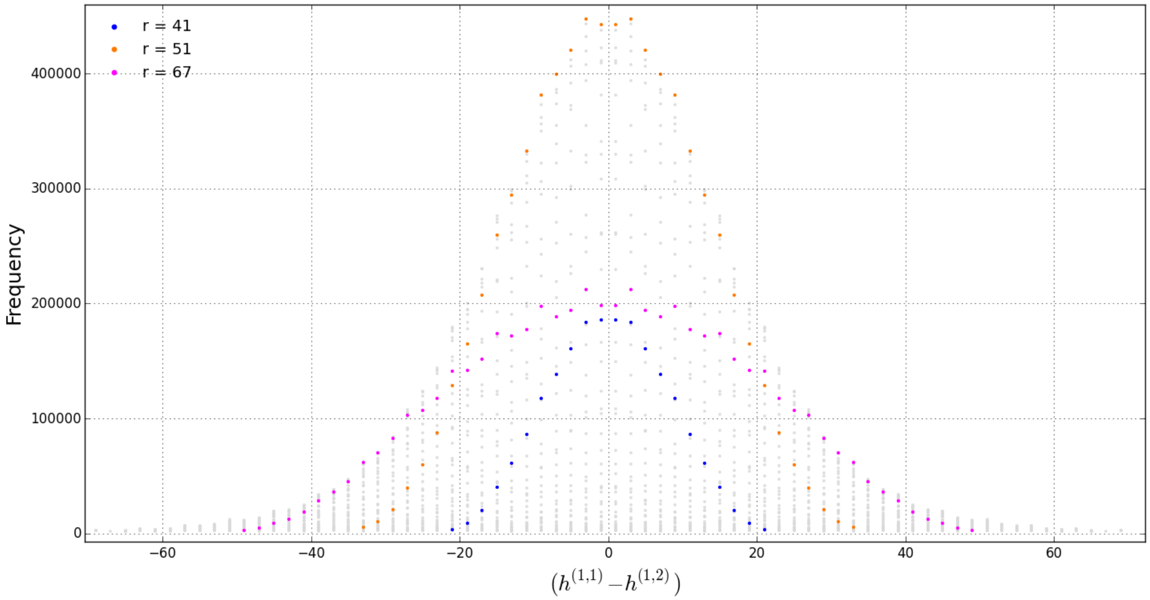
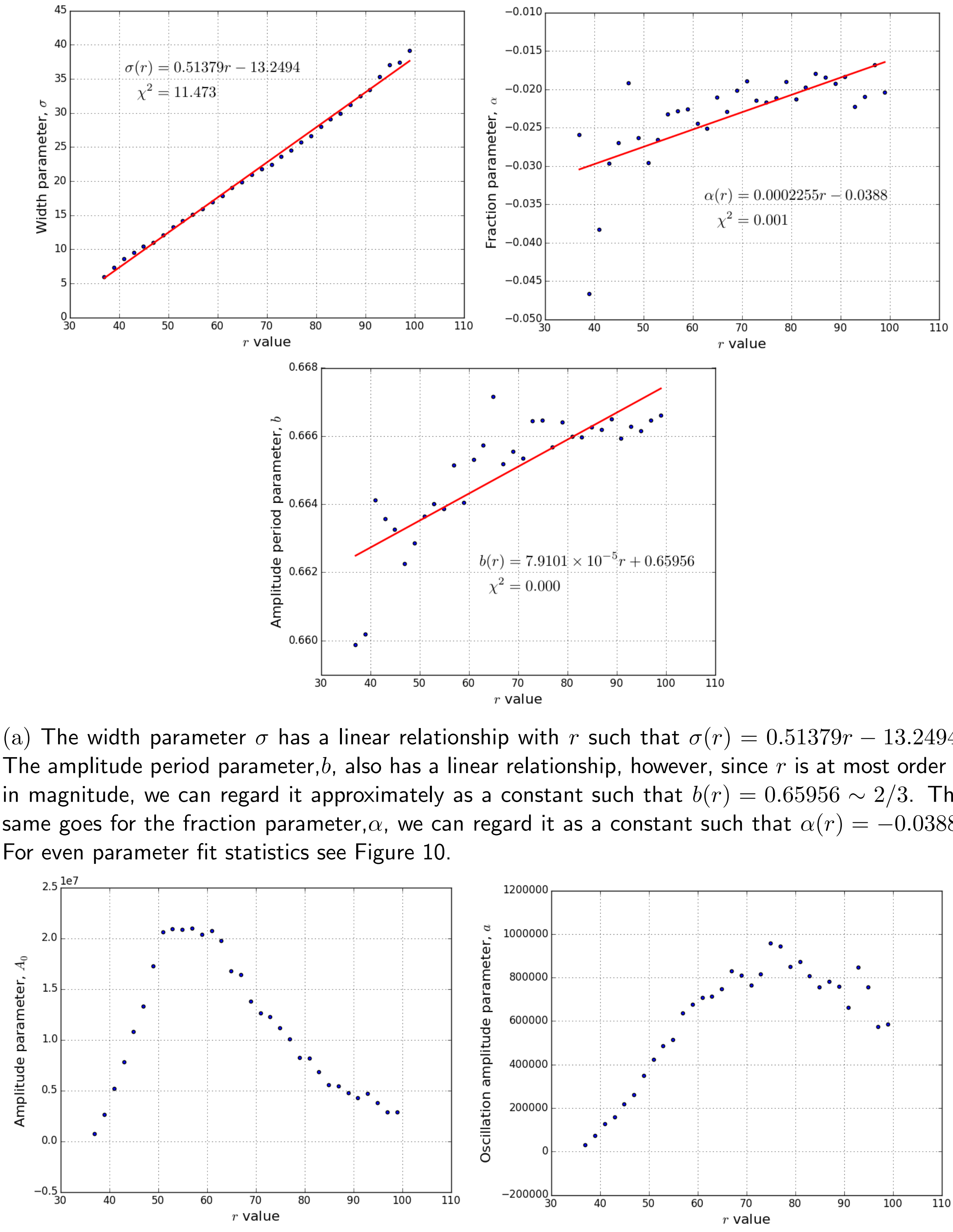


Figure A.1: Three highlighted curves ($r = 41, 51, 67$) within the odd $h^{1,1} - h^{1,2}$ distribution. The transparent grey data dots is the rest of the distribution. Refer to Figure 4 for the even plot.



(a) The width parameter σ has a linear relationship with r such that $\sigma(r) = 0.51379r - 13.2494$. The amplitude period parameter,b, also has a linear relationship, however, since r is at most order 3 in magnitude, we can regard it approximately as a constant such that $b(r) = 0.65956 \sim 2/3$. The same goes for the fraction parameter,α, we can regard it as a constant such that $\alpha(r) = -0.0388$. For even parameter fit statistics see Figure 10.

(b) Plots of A_0 vs r (left) and a vs r (right). Both exhibit a similar pattern, however it is difficult to find any nice relationships. For even parameter plots see Figure 10.

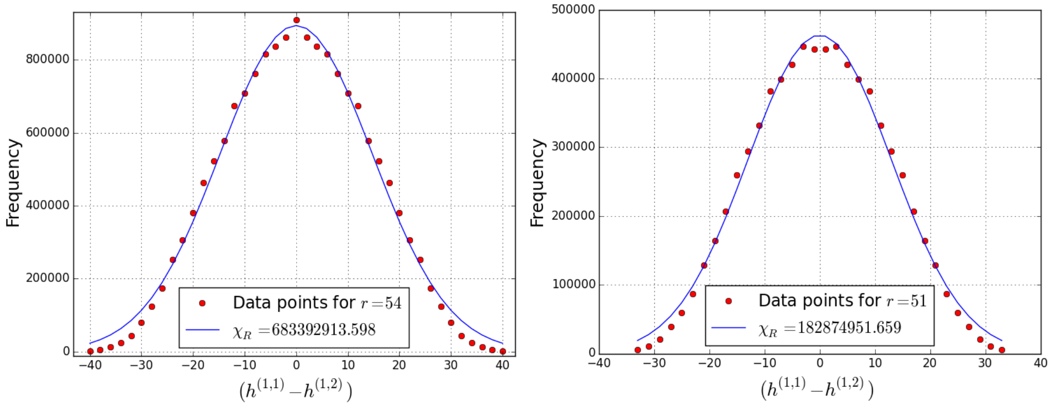
Figure A.2: The plots of the various parameters A, σ, α, b, a versus r for odd values of r.

A.1.2 Comparative plots

Here we present a comparison of various models we used, by plotting them side by side with the relevant fit-statistics. We choose a single even curve, $r = 54$, and odd curve, $r = 51$, to illustrate the difference between models.

Gaussian Model

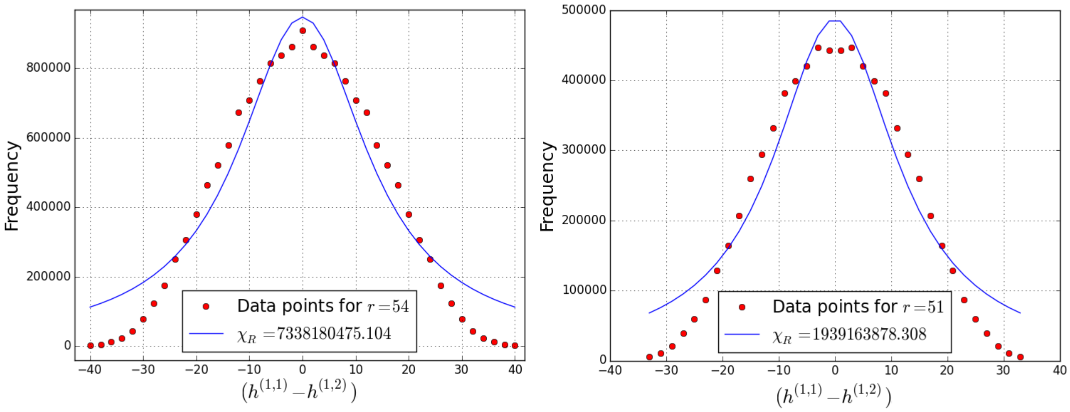
$$f(x, A, \mu, \sigma) = \frac{A}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{A.1}$$



(a) Gaussian model.

Lorentzian Model

$$f(x, A, \mu, \sigma) = \frac{A}{\pi} \left[\frac{\sigma}{(x-\mu)^2 + \sigma^2} \right] \tag{A.2}$$

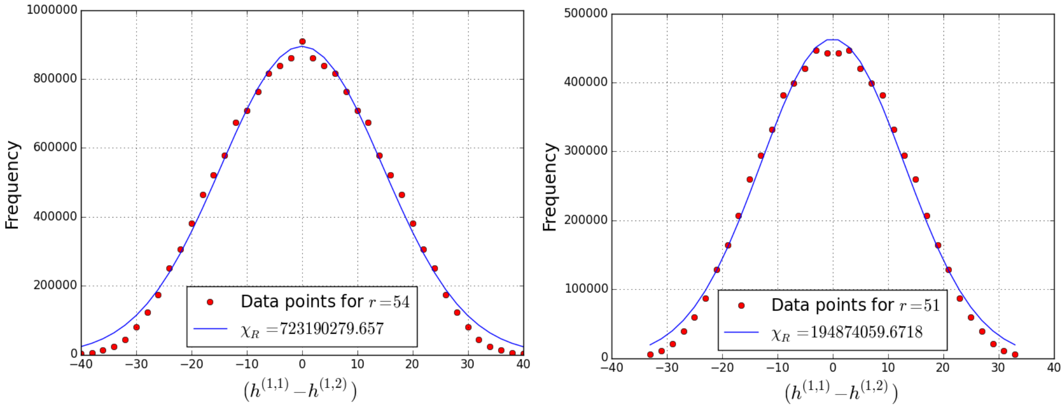


(a) Lorentzian (Cauchy) model.

Pearson7 Model

$$f(x, A, \mu, \sigma, m) = \frac{A}{\sigma\beta(m - \frac{1}{2}, \frac{1}{2})}\left[1 + \frac{(x-\mu)^2}{\sigma^2}\right]^{-m}, \tag{A.3}$$

where β is the Beta function.

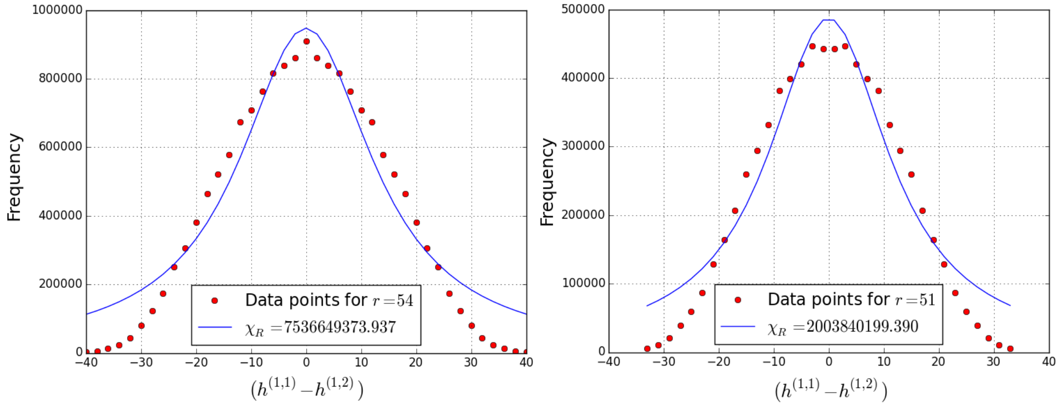


(a) Pearson7 model.

Breit-Wigner Model

This model is based on the Breit-Wigner function.

$$f(x, A, \mu, \sigma, t) = \frac{A(t\sigma/2 + x - \mu)^2}{(\sigma/2)^2 + (x-\mu)^2} \tag{A.4}$$



(a) Breit–Wigner model.

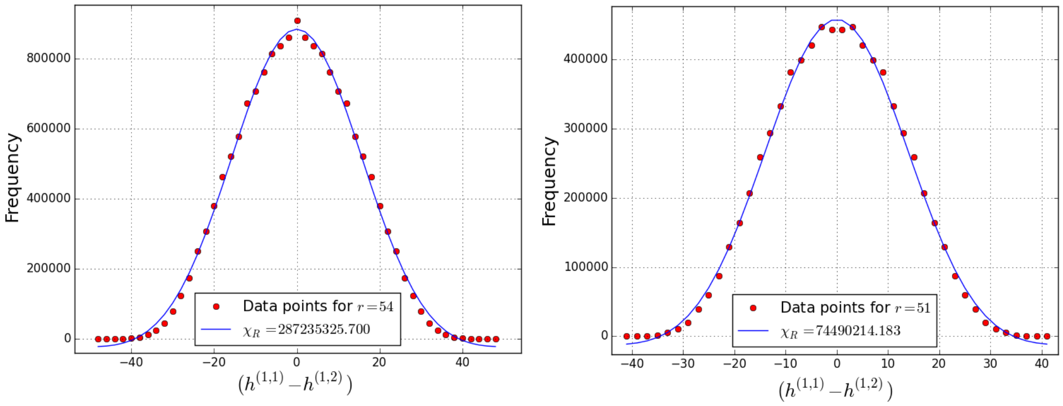
Voigt Model

$$f(x, A, \mu, \sigma, \gamma) = \frac{a\mathrm{Re}[(z)]}{\sigma\sqrt{2\pi}} \tag{A.5}$$

where

$$z = \frac{x - \mu + i\gamma}{\sigma\sqrt{2}} \;, \quad w(z) = e^{-z^2}\mathrm{erfc}(-iz) \tag{A.6}$$

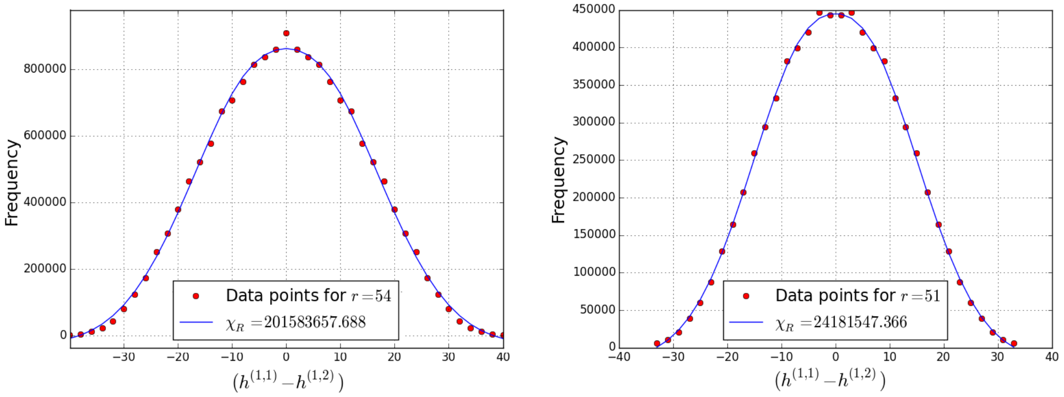
The Voigt model is a convolution of the Gaussian and Lorentzian models.



(a) Voigt model.

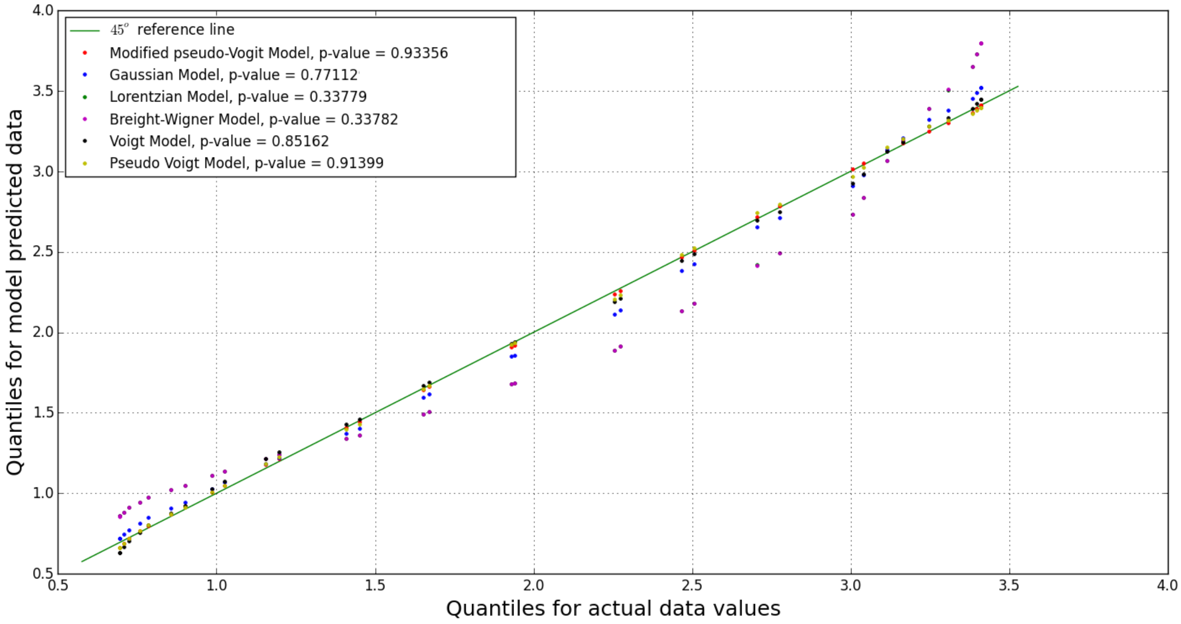
Pseudo-Voigt Model

$$f(x, A, \mu, \sigma, \alpha) = (1 - \alpha)\frac{A}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} + \alpha\frac{A}{\pi}\left[\frac{\sigma^2}{(x - \mu)^2} + \sigma^2\right] \tag{A.7}$$

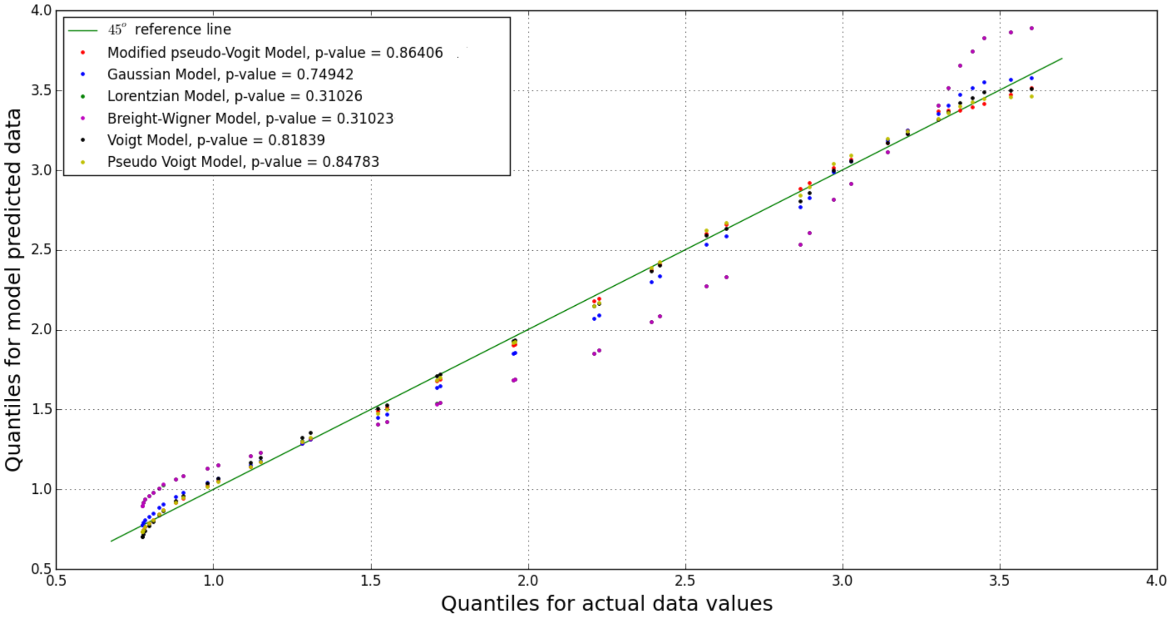


(a) Pseudo-Voigt model.

We present the standardized and shifted probability plots for the above comparisons:



(a) The probability plot for $r = 51$.

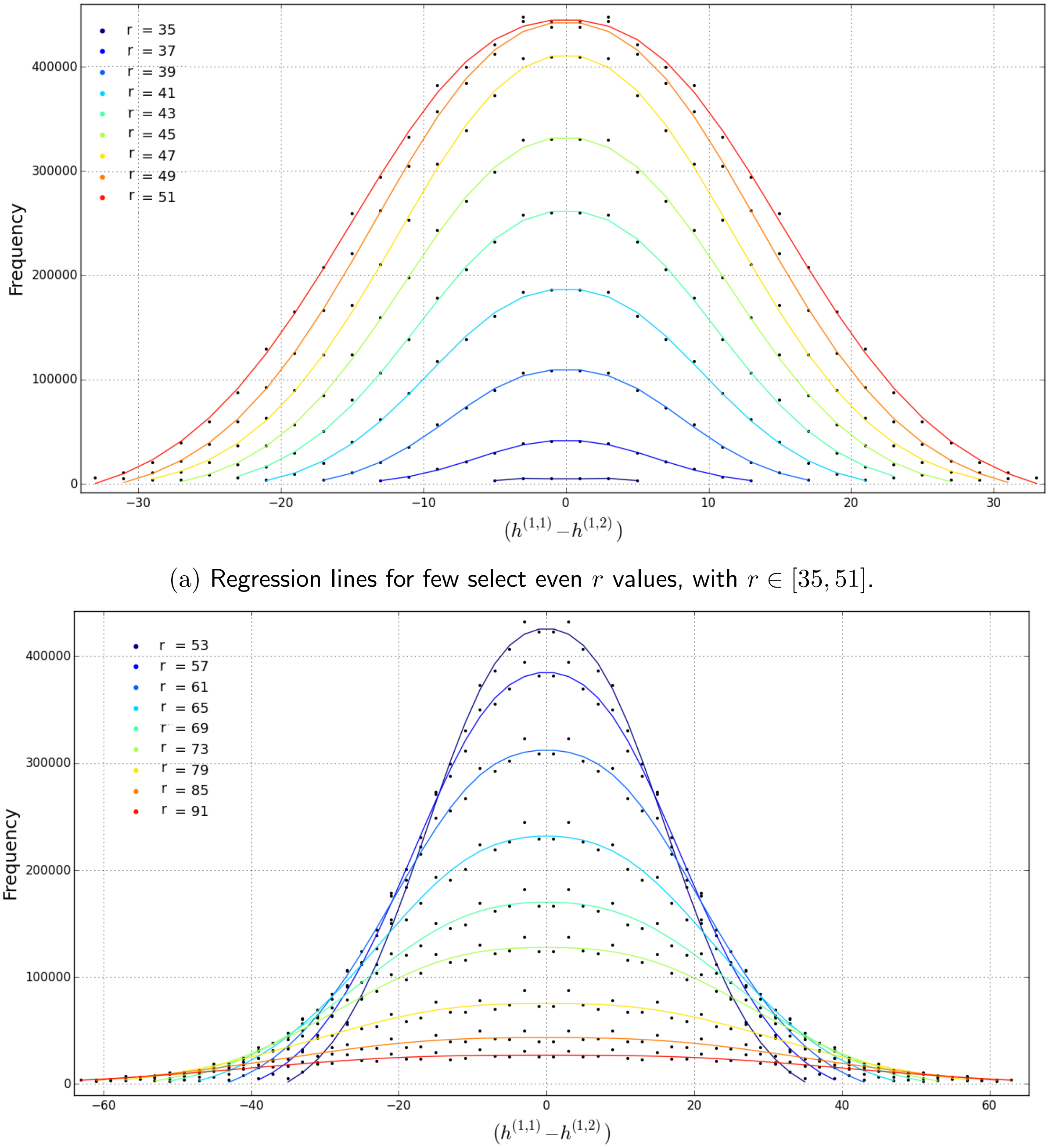


(a) The probability plot for $r = 54$.

Figure A.-4: For all models, the left hand graph is for $r = 54$ and the right is for $r = 51$. The probability plot presents all the models together. All the above mentioned modeled are included to compare their resemblance with the actual data. The larger the p value the better the line $y = x$ fits the data, implying the better the model is at describing the data.

A.1.3 A first approximation to the data

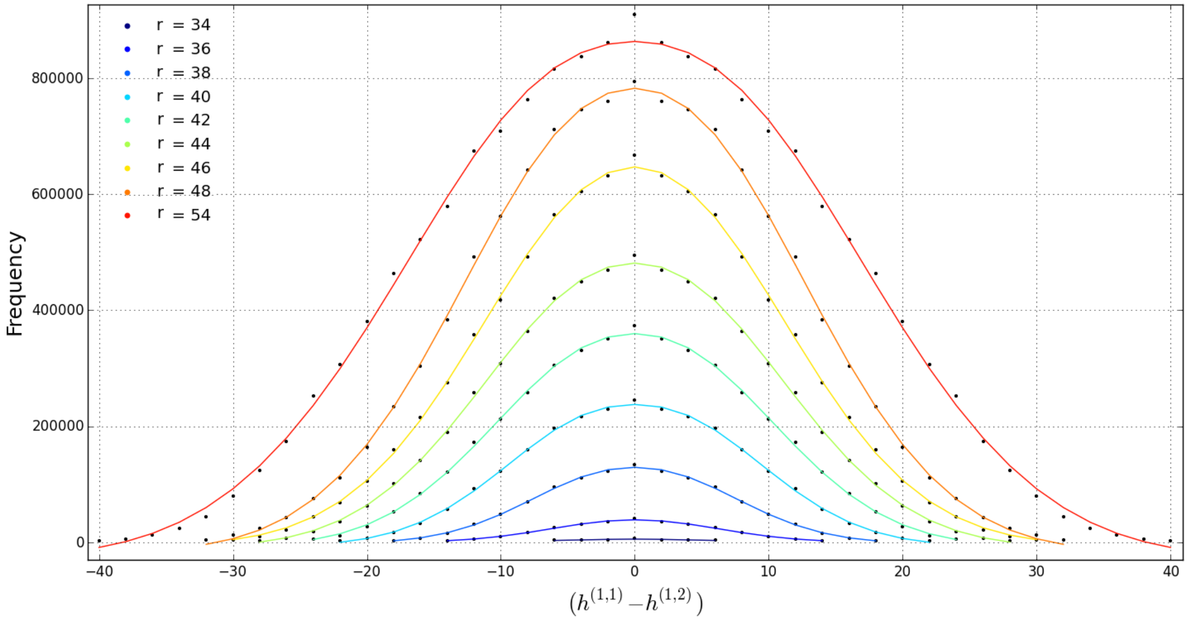
The overall behavior of the data across each curve is modeled extremely well using the Pseudo-Voigt model. Here we present a few plots illustrating a first approximation to the data. A second approximation can be made by introducing an oscillating amplitude as described in Section 2.1



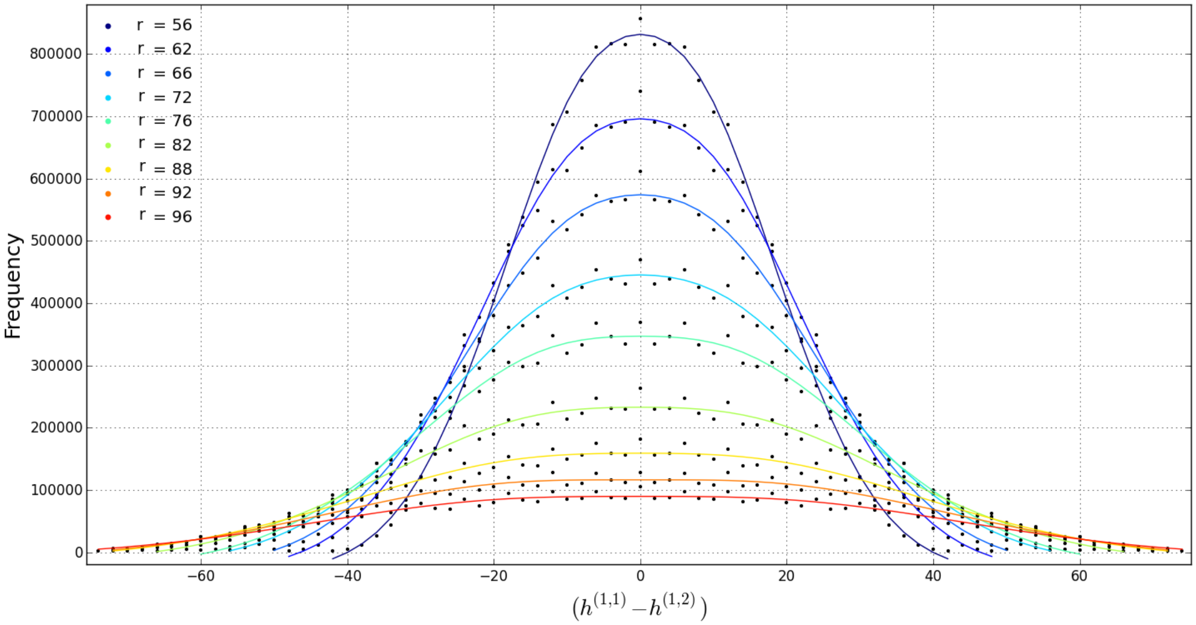
(a) Regression lines for few select even r values, with $r \in [35, 51]$.

(b) Regression lines for few select even r values, with $r > 51$.

Figure A.-3: Best fit curve based on the Pseudo-Voigt model for the same sets of curves as seen in Figure 5.



(a) Regression lines for few select even r values, with $r \leq 54$.

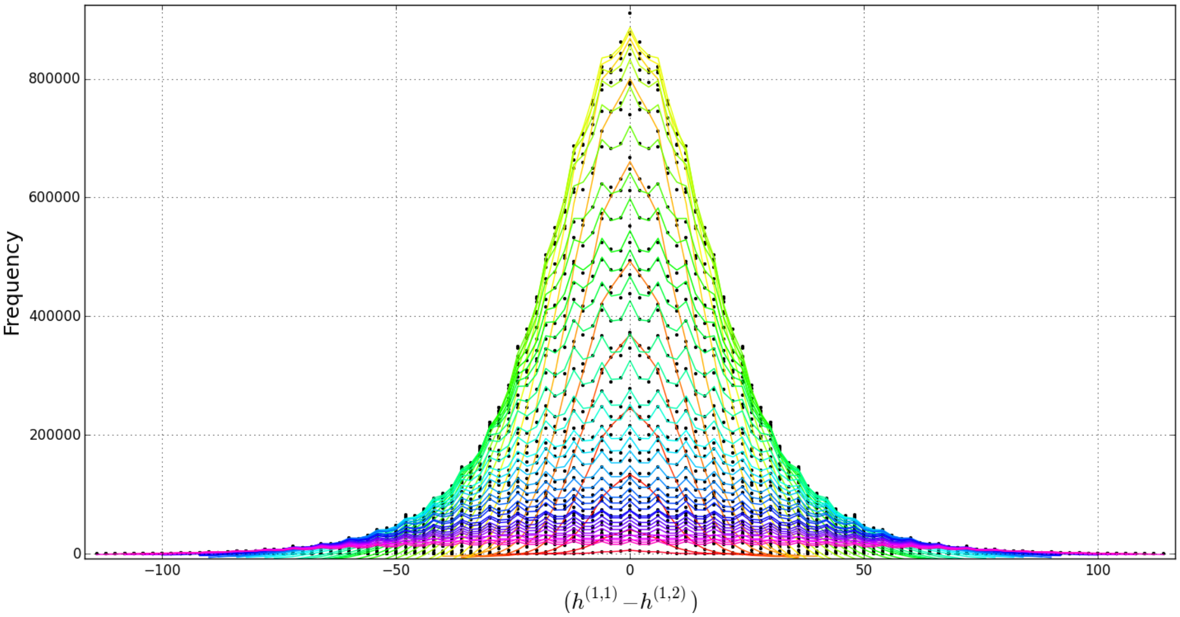


(b) Regression lines for few select even r values, with $r > 54$.

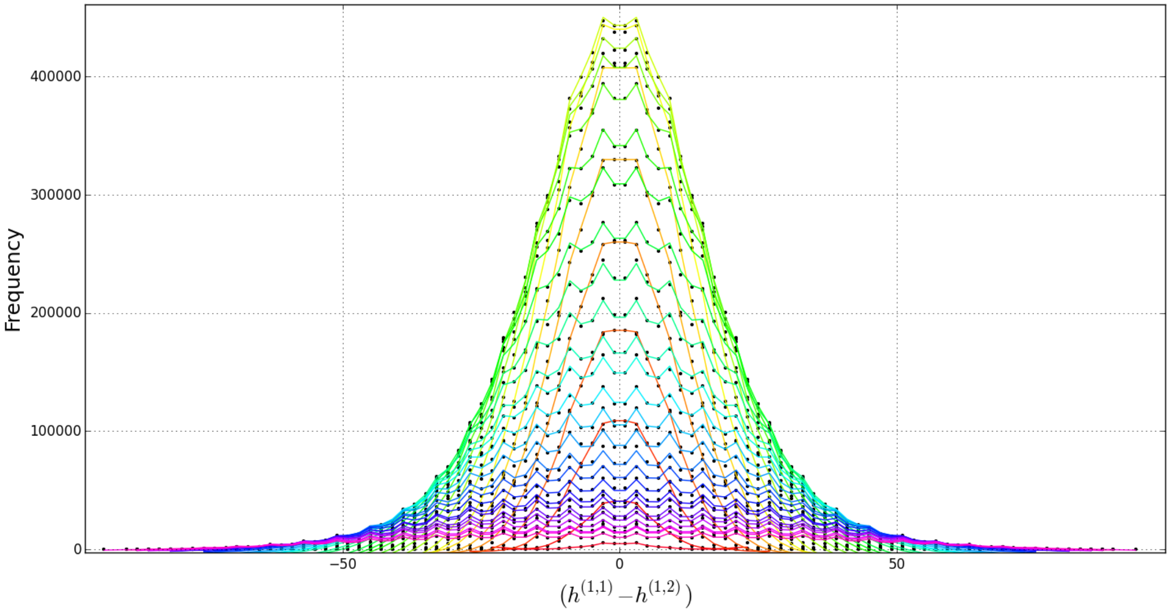
Figure A.-2: Best fit curve based on the Pseudo-Voigt model for the same sets of curves as seen in Figure 6.

A.1.4 Table of parameter values and statistics

Here we present the parameter values as well as the reduced χ value, χ_R, in a tabular format for all even r curves — $r \in [34, 120]$ — and for all odd r curves — $r \in [35, 99]$.

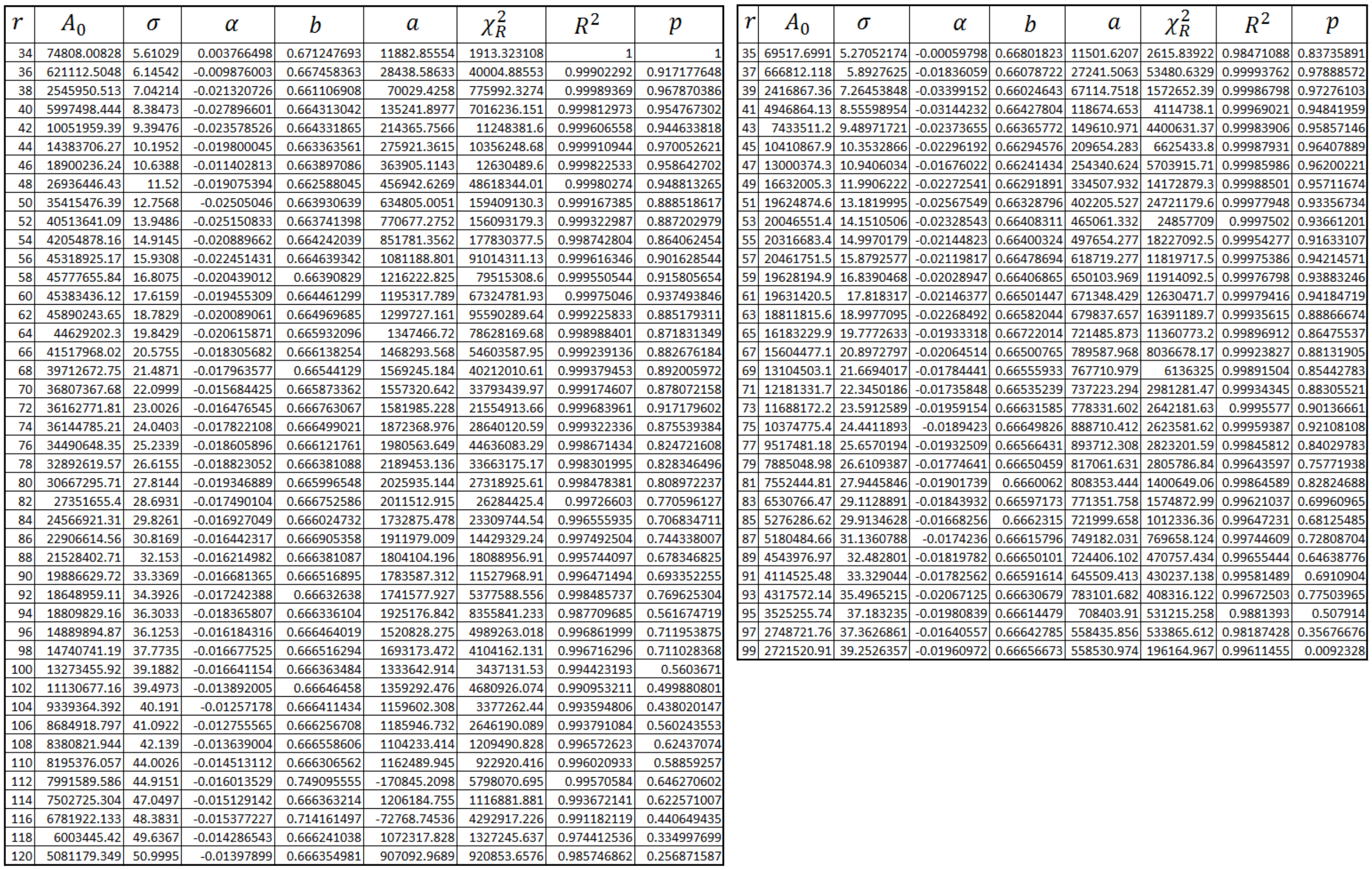


(a) Every fitted even curve from $r = 34$ until $r = 120$.



(b) Every fitted even odd from $r = 35$ until $r = 99$.

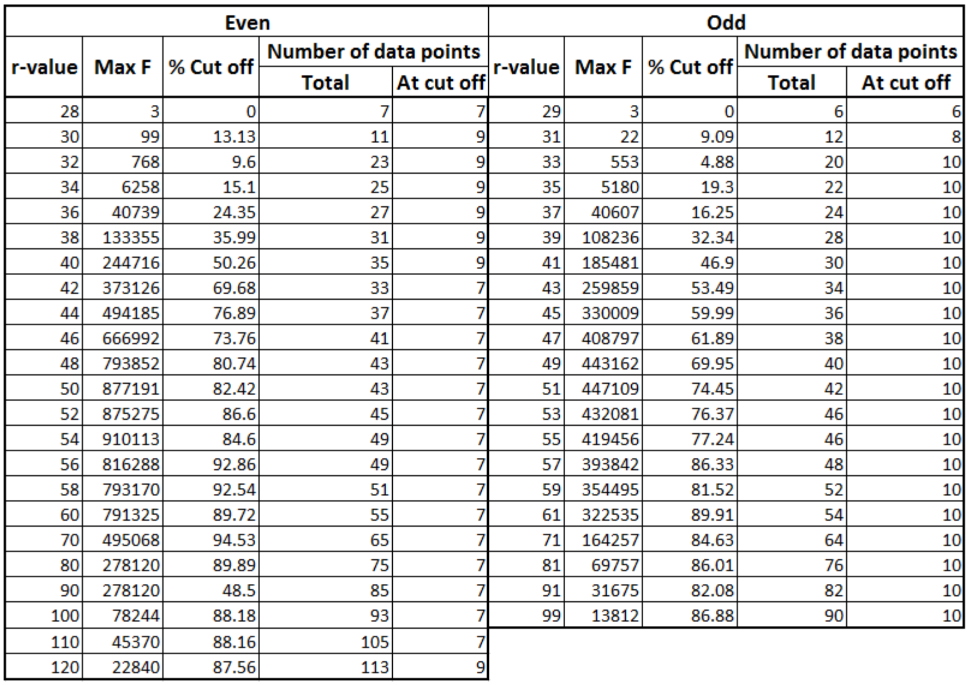
Figure A.-1: This is what the entire distribution looks like using our modified pseudo-Voigt model. See Figure A.0 for the fitted coefficients as well as the fits for every curve given by the probability plots.



r	A_0	σ	α	b	a	χ_R^2	R^2	p
34	74808.00828	5.61029	0.003766498	0.671247693	11882.85554	1913.323108	1	1
36	621112.5048	6.14542	-0.009876003	0.667458363	28438.58633	40004.88553	0.99902292	0.917177648
38	2545950.513	7.04214	-0.021320726	0.661106908	70029.4258	775992.3274	0.99989369	0.967870386
40	5997498.444	8.38473	-0.027896601	0.664313042	135241.8977	7016236.151	0.999812973	0.954767302
42	10051959.39	9.39476	-0.023578526	0.664331865	214365.7566	11248381.6	0.999606558	0.944633818
44	14383706.27	10.1952	-0.019800045	0.663363561	275921.3615	10356248.68	0.999910944	0.970052621
46	18900236.24	10.6388	-0.011402813	0.663897086	363905.1143	12630489.6	0.999822533	0.958642702
48	26936446.43	11.52	-0.019075394	0.662588045	456942.6269	48618344.01	0.99980274	0.948813265
50	35415476.39	12.7568	-0.02505046	0.663930639	634805.0051	159409130.3	0.999167385	0.888518617
52	40513641.09	13.9486	-0.025150833	0.663741398	770677.2752	156093179.3	0.999322987	0.887202979
54	42054878.16	14.9145	-0.020889662	0.664242039	851781.3562	177830377.5	0.998742804	0.864062454
56	45318925.17	15.9308	-0.022451431	0.664639342	1081188.801	91014311.13	0.999616346	0.901628544
58	45777655.84	16.8075	-0.020439012	0.66390829	1216222.825	79515308.6	0.999550544	0.915805654
60	45383436.12	17.6159	-0.019455309	0.664461299	1195317.789	67324781.93	0.99975046	0.937493846
62	45890243.65	18.7829	-0.020089061	0.664969685	1299727.161	95590289.64	0.999225833	0.885179311
64	44629202.3	19.8429	-0.020615871	0.665932096	1347466.72	78628169.68	0.998988401	0.871831349
66	41517968.02	20.5755	-0.018305682	0.666138254	1468293.568	54603587.95	0.999239136	0.882676184
68	39712672.75	21.4871	-0.017963577	0.66544129	1569245.184	40212010.61	0.999379453	0.892005972
70	36807367.68	22.0999	-0.015684425	0.665873362	1557320.642	33793439.97	0.999174607	0.878072158
72	36162771.81	23.0026	-0.016476545	0.666763067	1581985.228	21554913.66	0.999683961	0.917179602
74	36144785.21	24.0403	-0.017822108	0.666499021	1872368.976	28640120.59	0.999322336	0.875539384
76	34490648.35	25.2339	-0.018605896	0.666121761	1980563.649	44636083.29	0.998671434	0.824721608
78	32892619.57	26.6155	-0.018823052	0.666381088	2189453.136	33663175.17	0.998301995	0.828346496
80	30667295.71	27.8144	-0.019346889	0.665996548	2025935.144	27318925.61	0.998478381	0.808972237
82	27351655.4	28.6931	-0.017490104	0.666752586	2011512.915	26284425.4	0.99726603	0.770596127
84	24566921.31	29.8261	-0.016927049	0.666024732	1732875.478	23309744.54	0.996555935	0.706834711
86	22906614.56	30.8169	-0.016442317	0.666905358	1911979.009	14429329.24	0.997492504	0.744338007
88	21528402.71	32.153	-0.016214982	0.666381087	1804104.196	18088956.91	0.995744097	0.678346825
90	19886629.72	33.3369	-0.016681365	0.666516895	1783587.312	11527968.91	0.996471494	0.693352255
92	18648959.11	34.3926	-0.017242388	0.66632638	1741577.927	5377588.556	0.998485737	0.769625304
94	18809829.16	36.3033	-0.018365807	0.666336104	1925176.842	8355841.233	0.987709685	0.561674719
96	14889894.87	36.1253	-0.016184316	0.666464019	1520828.275	4989263.018	0.996861999	0.711953875
98	14740741.19	37.7735	-0.016677525	0.666516294	1693173.472	4104162.131	0.996716296	0.711028368
100	13273455.92	39.1882	-0.016641154	0.666363484	1333642.914	3437131.53	0.994423193	0.5603671
102	11130677.16	39.4973	-0.013892005	0.66646458	1359292.476	4680926.074	0.990953211	0.499880801
104	9339364.392	40.191	-0.01257178	0.666411434	1159602.308	3377262.44	0.993594806	0.438020147
106	8684918.797	41.0922	-0.012755565	0.666256708	1185946.732	2646190.089	0.993791084	0.560243553
108	8380821.944	42.139	-0.013639004	0.666558606	1104233.414	1209490.828	0.996572623	0.62437074
110	8195376.057	44.0026	-0.014513112	0.666306562	1162489.945	922920.416	0.996020933	0.58859257
112	7991589.586	44.9151	-0.016013529	0.749095555	-170845.2098	5798070.695	0.99570584	0.646270602
114	7502725.304	47.0497	-0.015129142	0.666363214	1206184.755	1116881.881	0.993672141	0.622571007
116	6781922.133	48.3831	-0.015377227	0.714161497	-72768.74536	4292917.226	0.991182119	0.440649435
118	6003445.42	49.6367	-0.014286543	0.666241038	1072317.828	1327245.637	0.974412536	0.334997699
120	5081179.349	50.9995	-0.01397899	0.666354981	907092.9689	920853.6576	0.985746862	0.256871587

r	A_0	σ	α	b	a	χ_R^2	R^2	p
35	69517.6991	5.27052174	-0.00059798	0.66801823	11501.6207	2615.83922	0.98471088	0.83735891
37	666812.118	5.8927625	-0.01836059	0.66078722	27241.5063	53480.6329	0.99993762	0.97888572
39	2416867.36	7.26453848	-0.03399152	0.66024643	67114.7518	1572652.39	0.99986798	0.97276103
41	4946864.13	8.55598954	-0.03144232	0.66427804	118674.653	4114738.1	0.99969021	0.94841959
43	7433511.2	9.48971721	-0.02373655	0.66365772	149610.971	4400631.37	0.99983906	0.95857146
45	10410867.9	10.3532866	-0.02296192	0.66294576	209654.283	6625433.8	0.99987931	0.96407889
47	13000374.3	10.9406034	-0.01676022	0.66241434	254340.624	5703915.71	0.99985986	0.96200221
49	16632005.3	11.9906222	-0.02272541	0.66291891	334507.932	14172879.3	0.99988501	0.95711674
51	19624874.6	13.1819995	-0.02567549	0.66328796	402205.527	24721179.6	0.99977948	0.93356734
53	20046551.4	14.1510506	-0.02328543	0.66408311	465061.332	24857709	0.9997502	0.93661201
55	20316683.4	14.9970179	-0.02144823	0.66400324	497654.277	18227092.5	0.99954277	0.91633107
57	20461751.5	15.8792577	-0.02119817	0.66478694	618719.277	11819717.5	0.99975386	0.94214571
59	19628194.9	16.8390468	-0.02028947	0.66406865	650103.969	11914092.5	0.99976798	0.93883246
61	19631420.5	17.818317	-0.02146377	0.66501447	671348.429	12630471.7	0.99979416	0.94184719
63	18811815.6	18.9977095	-0.02268492	0.66582044	679837.657	16391189.7	0.99935615	0.88866674
65	16183229.9	19.7772633	-0.01933318	0.66722014	721485.873	11360773.2	0.99896912	0.86475537
67	15604477.1	20.8972797	-0.02064514	0.66500765	789587.968	8036678.17	0.99923827	0.88131905
69	13104503.1	21.6694017	-0.01784441	0.66555933	767710.979	6136325	0.99891504	0.85442783
71	12181331.7	22.3450186	-0.01735848	0.66535239	737223.294	2981281.47	0.99934345	0.88305521
73	11688172.2	23.5912589	-0.01959154	0.66631585	778331.602	2642181.63	0.9995577	0.90136661
75	10374775.4	24.4411893	-0.0189423	0.66649826	888710.412	2623581.62	0.99959387	0.92108108
77	9517481.18	25.6570194	-0.01932509	0.66566431	893712.308	2823201.59	0.99845812	0.84029783
79	7885048.98	26.6109387	-0.01774641	0.66650459	817061.631	2805786.84	0.99643597	0.75771938
81	7552444.81	27.9445846	-0.01901739	0.6660062	808353.444	1400649.06	0.99864589	0.82824688
83	6530766.47	29.1128891	-0.01843932	0.66597173	771351.758	1574872.99	0.99621037	0.69960965
85	5276286.62	29.9134628	-0.01668256	0.6662315	721999.658	1012336.36	0.99647231	0.68125485
87	5180484.66	31.1360788	-0.0174236	0.66615796	749182.031	769658.124	0.99744609	0.72808704
89	4543976.97	32.482801	-0.01819782	0.66650101	724406.102	470757.434	0.99655444	0.64638776
91	4114525.48	33.329044	-0.01782562	0.66591614	645509.413	430237.138	0.99581489	0.6910904
93	4317572.14	35.4965215	-0.02067125	0.66630679	783101.682	408316.122	0.99672503	0.77503965
95	3525255.74	37.183235	-0.01980839	0.66614479	708403.91	531215.258	0.9881393	0.507914
97	2748721.76	37.3626861	-0.01640557	0.66642785	558435.856	533865.612	0.98187428	0.35676676
99	2721520.91	39.2526357	-0.01960972	0.66656673	558530.974	196164.967	0.99611455	0.0092328

Figure A.0: Left : list of best fit coefficients for all even curves $r \in [34, 120]$. Right: List of best fit coefficients for all odd curves $r \in [35, 99]$. In both tables, the last two columns represent the R^2 and p values for the probability plot for each curve. The p-values were obtained by first performing a Z-Standardization on the data.



Even					Odd				
r-value	Max F	% Cut off	Number of data points		r-value	Max F	% Cut off	Number of data points	
			Total	At cut off				Total	At cut off
28	3	0	7	7	29	3	0	6	6
30	99	13.13	11	9	31	22	9.09	12	8
32	768	9.6	23	9	33	553	4.88	20	10
34	6258	15.1	25	9	35	5180	19.3	22	10
36	40739	24.35	27	9	37	40607	16.25	24	10
38	133355	35.99	31	9	39	108236	32.34	28	10
40	244716	50.26	35	9	41	185481	46.9	30	10
42	373126	69.68	33	7	43	259859	53.49	34	10
44	494185	76.89	37	7	45	330009	59.99	36	10
46	666992	73.76	41	7	47	408797	61.89	38	10
48	793852	80.74	43	7	49	443162	69.95	40	10
50	877191	82.42	43	7	51	447109	74.45	42	10
52	875275	86.6	45	7	53	432081	76.37	46	10
54	910113	84.6	49	7	55	419456	77.24	46	10
56	816288	92.86	49	7	57	393842	86.33	48	10
58	793170	92.54	51	7	59	354495	81.52	52	10
60	791325	89.72	55	7	61	322535	89.91	54	10
70	495068	94.53	65	7	71	164257	84.63	64	10
80	278120	89.89	75	7	81	69757	86.01	76	10
90	278120	48.5	85	7	91	31675	82.08	82	10
100	78244	88.18	93	7	99	13812	86.88	90	10
110	45370	88.16	105	7					
120	22840	87.56	113	9					

Figure A.1: A list showing the number of data points left after increasing the cut off frequency to achieve a perfect fit. Conversely, one may state is as, the number of data points for each curve required such that the model will result in a perfect fit.

A.2 Supplementary plots for the $h^{1,1} + h^{1,2}$ distribution

A.2.1 Plots for the odd distribution as counterparts to the even ones

All even plot counterparts will be referenced in the figures. The plots appear in the same order as in the main body, with descriptions only if necessary.

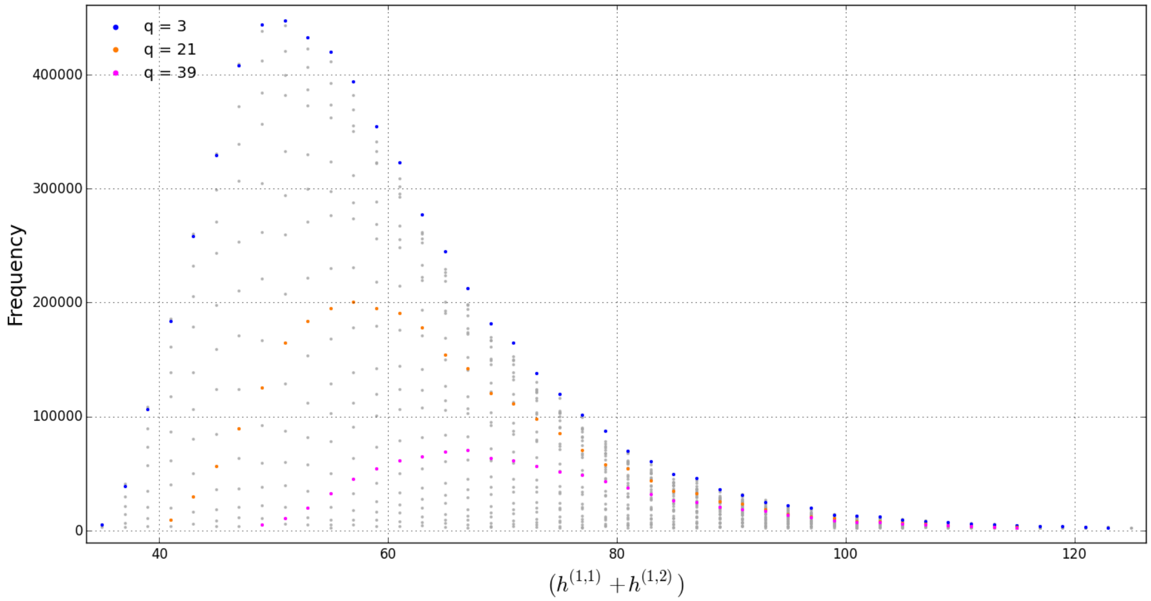
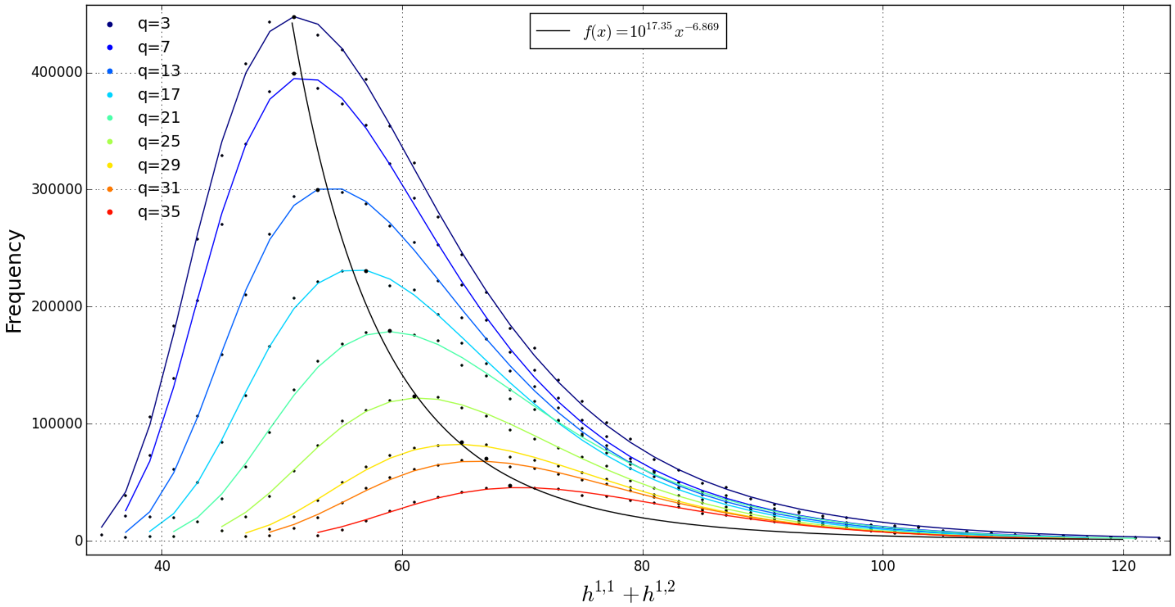
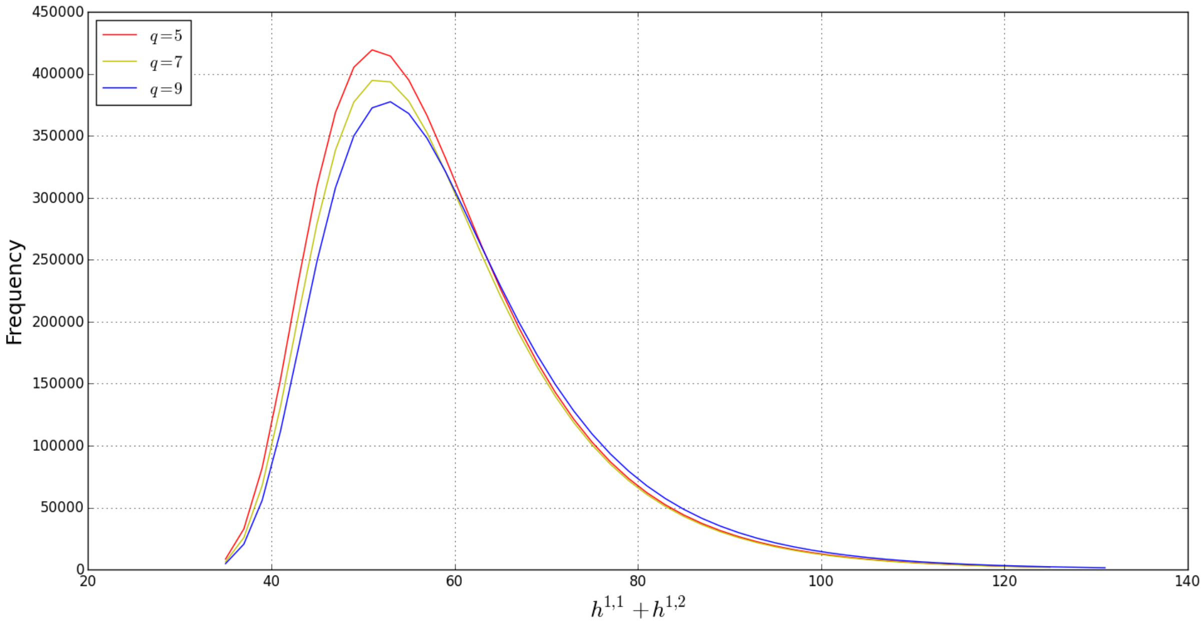


Figure A.2: Three highlighted curves ($q = 3, 19, 31$) within the odd $h^{1,1} + h^{1,2}$ distribution. The transparent grey data dots are all the data plots for the distribution. Refer to Figure 11 for the even plot.

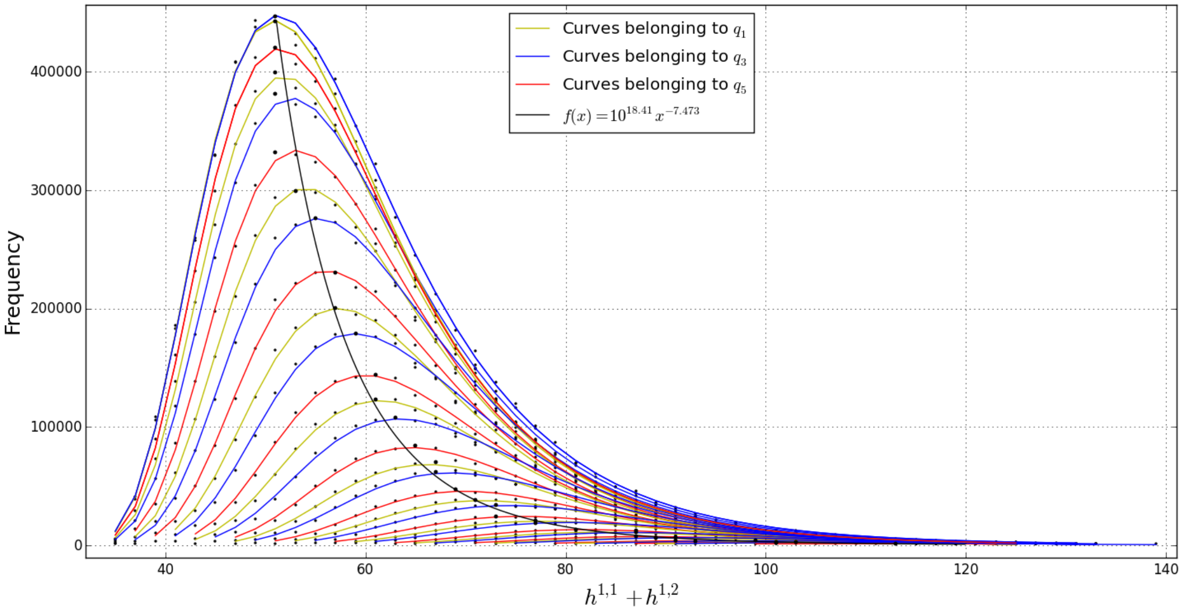


(a) Lines of best fit from a regression analysis for a few select curves. The black data points represent the maximum frequency for that particular $q - curve$. the black line is a line of best fit to describe the points of maximum frequency — this is analogous to a blackbody spectrum. See Figure 12a for the curves within the even distribution.

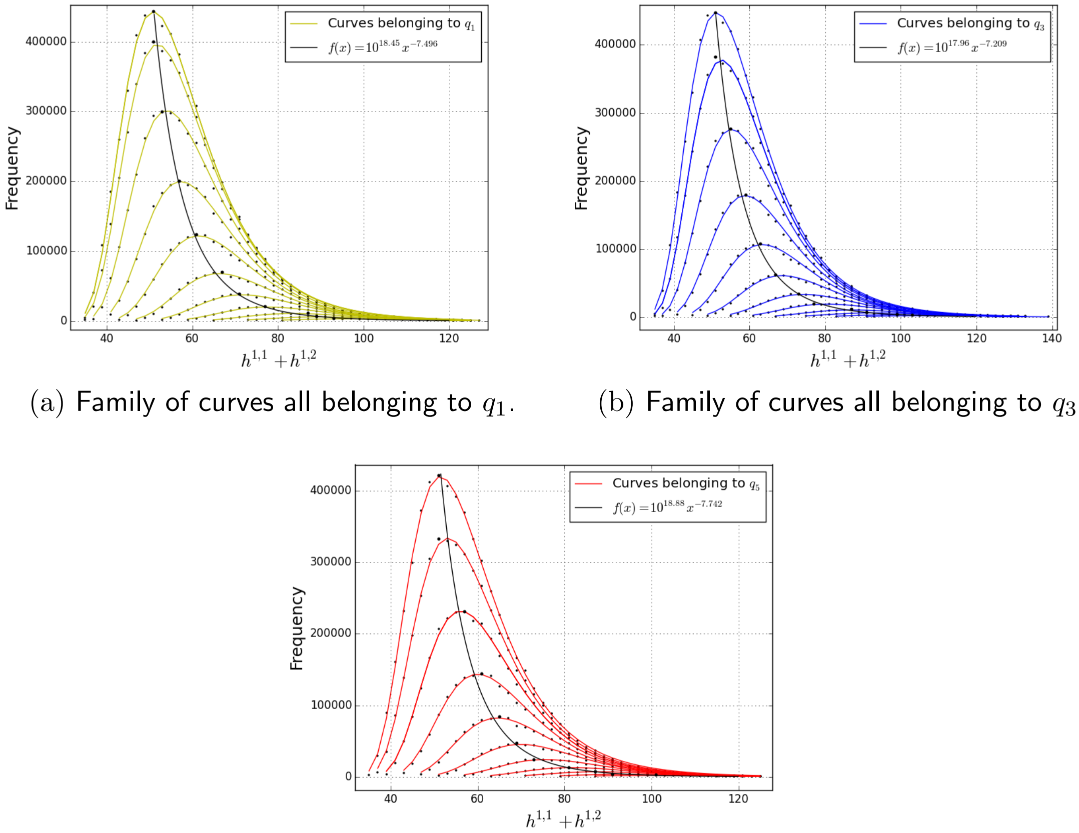


(a) The curves segregate into three classes determined by the value of the even integer modulo 6. A similar pattern occurs in the even distribution; see Figure 11a.

Figure A.2: In the attempt to describe the data analogously to a blackbody distribution (a), we discover some subtle structure, (b). These are the odd counterparts to Figure 11.



(a) All the curves color coded according to what residue class their curves q_n belongs to.

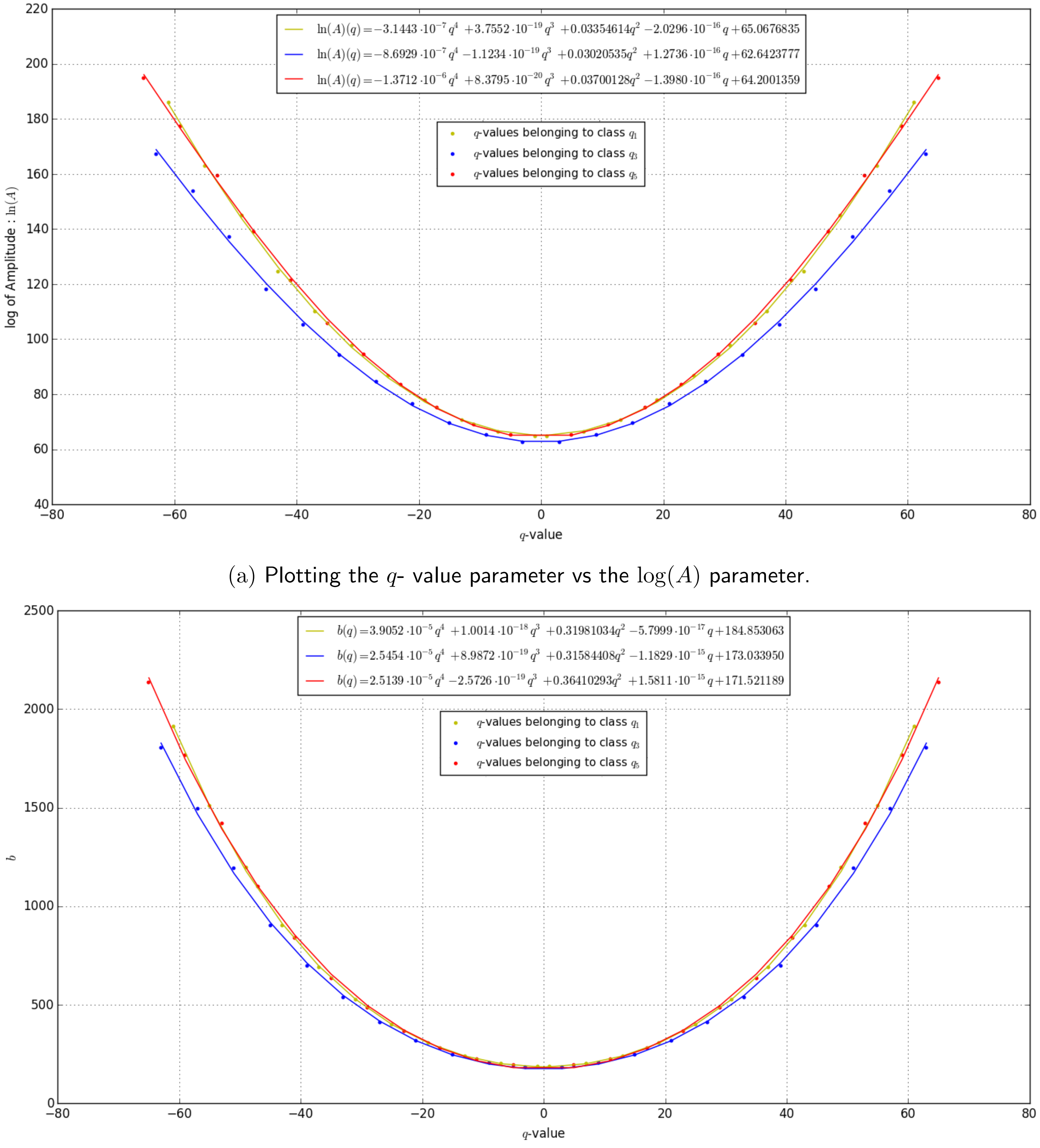


(a) Family of curves all belonging to q_1.

(b) Family of curves all belonging to q_3.

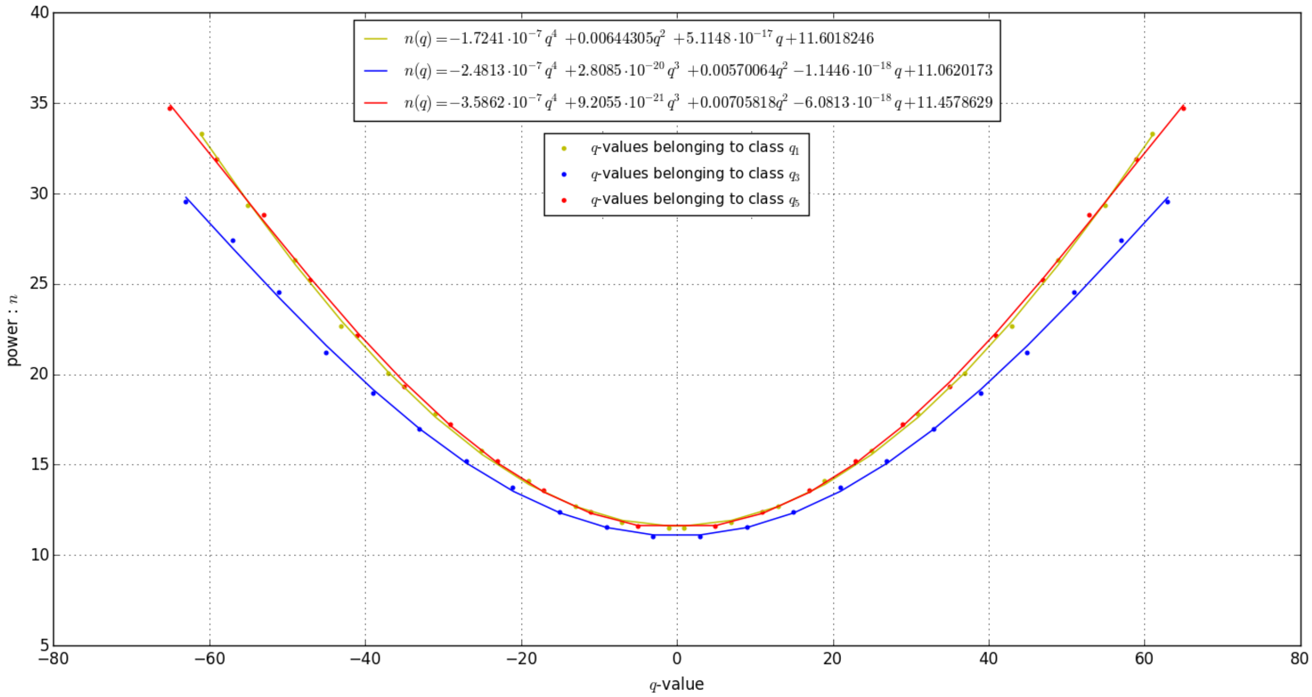
(c) Family of curves all belonging to q_5.

Figure A.2: We illustrate the added structure for odd $h^{1,1} + h^{1,2}$ data, by displaying how the regression curves can be divided into residue classes. For the list of even curves, refer to Figure 12.



(a) Plotting the q- value parameter vs the $\log(A)$ parameter.

(b) Plotting the q- value parameter vs the b parameter.



(a) Plotting the q- value parameter vs the power n parameter.

Figure A.2: The parameter plots are color coded according to what residue class their q value belong to. For the relationships in the even distribution, see Figure 12.

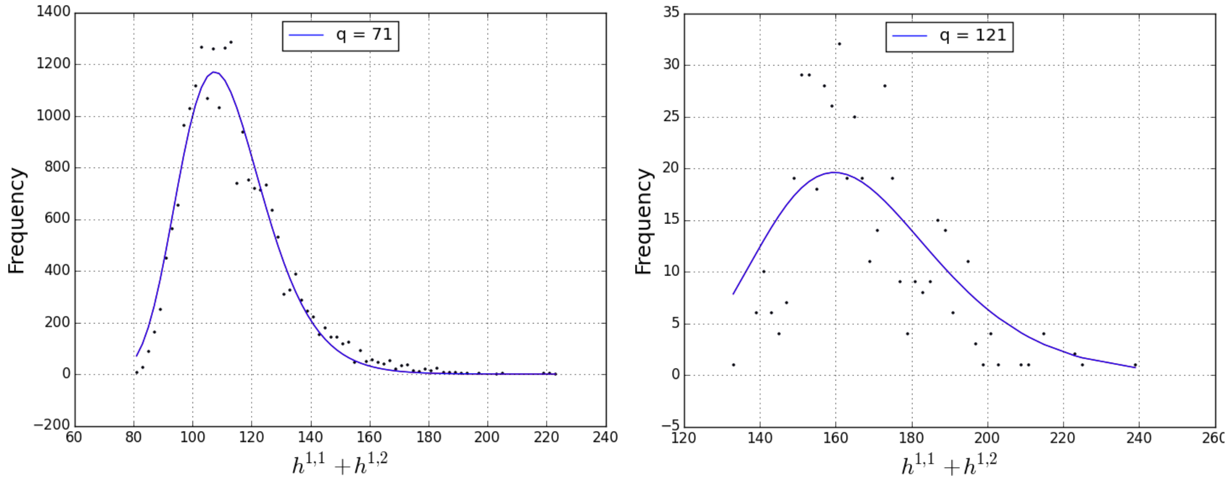
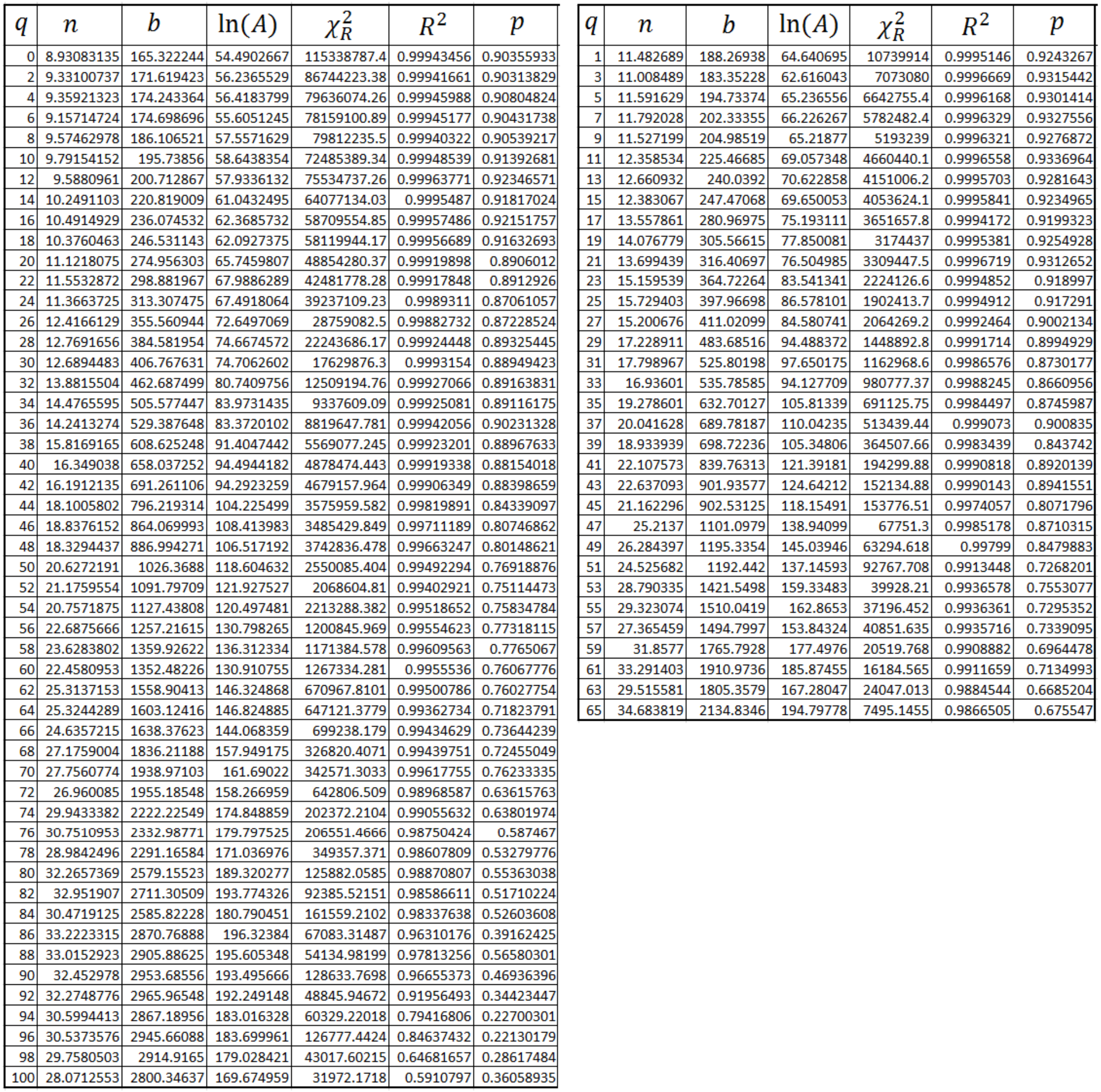


Figure A.3: Left figure is the fitted model(blue line) for a q value of 71 and right has a q value of 121. As the q-value increases, the scattering of the data points within $h^{1,1} + h^{1,2}$ increases to the point where the model works no longer. For an example of how the model begins to break down at large q, see Figure 13.

A.2.2 Table of parameter values, coefficient values and statistics



q	n	b	$\ln(A)$	χ_R^2	R^2	p
0	8.93083135	165.322244	54.4902667	115338787.4	0.99943456	0.90355933
2	9.33100737	171.619423	56.2365529	86744223.38	0.99941661	0.90313829
4	9.35921323	174.243364	56.4183799	79636074.26	0.99945988	0.90804824
6	9.15714724	174.698696	55.6051245	78159100.89	0.99945177	0.90431738
8	9.57462978	186.106521	57.5571629	79812235.5	0.99940322	0.90539217
10	9.79154152	195.73856	58.6438354	72485389.34	0.99948539	0.91392681
12	9.5880961	200.712867	57.9336132	75534737.26	0.99963771	0.92346571
14	10.2491103	220.819009	61.0432495	64077134.03	0.9995487	0.91817024
16	10.4914929	236.074532	62.3685732	58709554.85	0.99957486	0.92151757
18	10.3760463	246.531143	62.0927375	58119944.17	0.99956689	0.91632693
20	11.1218075	274.956303	65.7459807	48854280.37	0.99919898	0.8906012
22	11.5532872	298.881967	67.9886289	42481778.28	0.99917848	0.8912926
24	11.3663725	313.307475	67.4918064	39237109.23	0.9989311	0.87061057
26	12.4166129	355.560944	72.6497069	28759082.5	0.99882732	0.87228524
28	12.7691656	384.581954	74.6674572	22243686.17	0.99924448	0.89325445
30	12.6894483	406.767631	74.7062602	17629876.3	0.9993154	0.88949423
32	13.8815504	462.687499	80.7409756	12509194.76	0.99927066	0.89163831
34	14.4765595	505.577447	83.9731435	9337609.09	0.99925081	0.89116175
36	14.2413274	529.387648	83.3720102	8819647.781	0.99942056	0.90231328
38	15.8169165	608.625248	91.4047442	5569077.245	0.99923201	0.88967633
40	16.349038	658.037252	94.4944182	4878474.443	0.99919338	0.88154018
42	16.1912135	691.261106	94.2923259	4679157.964	0.99906349	0.88398659
44	18.1005802	796.219314	104.225499	3575959.582	0.99819891	0.84339097
46	18.8376152	864.069993	108.413983	3485429.849	0.99711189	0.80746862
48	18.3294437	886.994271	106.517192	3742836.478	0.99663247	0.80148621
50	20.6272191	1026.3688	118.604632	2550085.404	0.99492294	0.76918876
52	21.1759554	1091.79709	121.927527	2068604.81	0.99402921	0.75114473
54	20.7571875	1127.43808	120.497481	2213288.382	0.99518652	0.75834784
56	22.6875666	1257.21615	130.798265	1200845.969	0.99554623	0.77318115
58	23.6283802	1359.92622	136.312334	1171384.578	0.99609563	0.7765067
60	22.4580953	1352.48226	130.910755	1267334.281	0.9955536	0.76067776
62	25.3137153	1558.90413	146.324868	670967.8101	0.99500786	0.76027754
64	25.3244289	1603.12416	146.824885	647121.3779	0.99362734	0.71823791
66	24.6357215	1638.37623	144.068359	699238.179	0.99434629	0.73644239
68	27.1759004	1836.21188	157.949175	326820.4071	0.99439751	0.72455049
70	27.7560774	1938.97103	161.69022	342571.3033	0.99617755	0.76233335
72	26.960085	1955.18548	158.266959	642806.509	0.98968587	0.63615763
74	29.9433382	2222.22549	174.848859	202372.2104	0.99055632	0.63801974
76	30.7510953	2332.98771	179.797525	206551.4666	0.98750424	0.587467
78	28.9842496	2291.16584	171.036976	349357.371	0.98607809	0.53279776
80	32.2657369	2579.15523	189.320277	125882.0585	0.98870807	0.55363038
82	32.951907	2711.30509	193.774326	92385.52151	0.98586611	0.51710224
84	30.4719125	2585.82228	180.790451	161559.2102	0.98337638	0.52603608
86	33.2223315	2870.76888	196.32384	67083.31487	0.96310176	0.39162425
88	33.0152923	2905.88625	195.605348	54134.98199	0.97813256	0.56580301
90	32.452978	2953.68556	193.495666	128633.7698	0.96655373	0.46936396
92	32.2748776	2965.96548	192.249148	48845.94672	0.91956493	0.34423447
94	30.5994413	2867.18956	183.016328	60329.22018	0.79416806	0.22700301
96	30.5373576	2945.66088	183.699961	126777.4424	0.84637432	0.22130179
98	29.7580503	2914.9165	179.028421	43017.60215	0.64681657	0.28617484
100	28.0712553	2800.34637	169.674959	31972.1718	0.5910797	0.36058935

q	n	b	$\ln(A)$	χ_R^2	R^2	p
1	11.482689	188.26938	64.640695	10739914	0.9995146	0.9243267
3	11.008489	183.35228	62.616043	7073080	0.9996669	0.9315442
5	11.591629	194.73374	65.236556	6642755.4	0.9996168	0.9301414
7	11.792028	202.33355	66.226267	5782482.4	0.9996329	0.9327556
9	11.527199	204.98519	65.21877	5193239	0.9996321	0.9276872
11	12.358534	225.46685	69.057348	4660440.1	0.9996558	0.9336964
13	12.660932	240.0392	70.622858	4151006.2	0.9995703	0.9281643
15	12.383067	247.47068	69.650053	4053624.1	0.9995841	0.9234965
17	13.557861	280.96975	75.193111	3651657.8	0.9994172	0.9199323
19	14.076779	305.56615	77.850081	3174437	0.9995381	0.9254928
21	13.699439	316.40697	76.504985	3309447.5	0.9996719	0.9312652
23	15.159539	364.72264	83.541341	2224126.6	0.9994852	0.918997
25	15.729403	397.96698	86.578101	1902413.7	0.9994912	0.917291
27	15.200676	411.02099	84.580741	2064269.2	0.9992464	0.9002134
29	17.228911	483.68516	94.488372	1448892.8	0.9991714	0.8994929
31	17.798967	525.80198	97.650175	1162968.6	0.9986576	0.8730177
33	16.93601	535.78585	94.127709	980777.37	0.9988245	0.8660956
35	19.278601	632.70127	105.81339	691125.75	0.9984497	0.8745987
37	20.041628	689.78187	110.04235	513439.44	0.999073	0.900835
39	18.933939	698.72236	105.34806	364507.66	0.9983439	0.843742
41	22.107573	839.76313	121.39181	194299.88	0.9990818	0.8920139
43	22.637093	901.93577	124.64212	152134.88	0.9990143	0.8941551
45	21.162296	902.53125	118.15491	153776.51	0.9974057	0.8071796
47	25.2137	1101.0979	138.94099	67751.3	0.9985178	0.8710315
49	26.284397	1195.3354	145.03946	63294.618	0.99799	0.8479883
51	24.525682	1192.442	137.14593	92767.708	0.9913448	0.7268201
53	28.790335	1421.5498	159.33483	39928.21	0.9936578	0.7553077
55	29.323074	1510.0419	162.8653	37196.452	0.9936361	0.7295352
57	27.365459	1494.7997	153.84324	40851.635	0.9935716	0.7339095
59	31.8577	1765.7928	177.4976	20519.768	0.9908882	0.6964478
61	33.291403	1910.9736	185.87455	16184.565	0.9911659	0.7134993
63	29.515581	1805.3579	167.28047	24047.013	0.9884544	0.6685204
65	34.683819	2134.8346	194.79778	7495.1455	0.9866505	0.675547

Figure A.4: Left : list of best fit coefficients for all even curves $q \in [0, 100]$. Right: List of best fit coefficients for all odd curves $q \in [1, 65]$.

Coefficient values for the description of the entire $h^{1,1} + h^{1,2}$ distribution

$$A_{k,i} = \begin{pmatrix} 54.2664195 & 2.9066 \times 10^{-16} & 0.02414823 & -5.4137 \times 10^{-20} & -7.2635 \times 10^{-7} \\ 65.0676835 & -2.0296 \times 10^{-16} & 0.03354614 & 3.7552 \times 10^{-19} & -3.1443 \times 10^{-7} \\ 54.8909275 & -2.0323 \times 10^{-16} & 0.02753302 & -2.7091 \times 10^{-20} & -9.1972 \times 10^{-7} \\ 62.6423777 & 1.2736 \times 10^{-16} & 0.03020535 & -1.1234 \times 10^{-19} & -8.6929 \times 10^{-7} \\ 54.5840853 & 2.9011 \times 10^{-16} & 0.02748121 & -9.4235 \times 10^{-20} & -9.3840 \times 10^{-7} \\ 64.2001359 & -1.3980 \times 10^{-16} & 0.03700128 & 8.3795 \times 10^{-20} & -1.3712 \times 10^{-7} \end{pmatrix} \tag{A.8}$$

$$b_{k,i} = \begin{pmatrix} 132.357878 & 3.3411 \times 10^{-15} & 0.32753297 & -8.6619 \times 10^{-19} & 4.5825 \times 10^{-6} \\ 184.853063 & -5.7999 \times 10^{-17} & 0.31981034 & 1.0014 \times 10^{-18} & 3.9052 \times 10^{-5} \\ 117.228782 & -1.2791 \times 10^{-15} & 0.36989364 & -8.5325 \times 10^{-20} & 2.9743 \times 10^{-6} \\ 173.033950 & -1.1829 \times 10^{-15} & 0.31584408 & 8.9872 \times 10^{-19} & 2.5454 \times 10^{-5} \\ 105.298297 & 5.7916 \times 10^{-15} & 0.37843953 & -1.5078 \times 10^{-18} & 1.3974 \times 10^{-6} \\ 171.521189 & 1.5811 \times 10^{-15} & 0.36410293 & -2.5726 \times 10^{-19} & 2.5139 \times 10^{-5} \end{pmatrix} \tag{A.9}$$

$$n_{k,i} = \begin{pmatrix} 8.98205242 & 2.9066 \times 10^{-17} & 0.00434183 & -6.7671 \times 10^{-21} & -1.5512 \times 10^{-7} \\ 11.6018246 & 5.1148 \times 10^{-17} & 0.00644305 & 0 & -1.7241 \times 10^{-7} \\ 9.19515076 & 4.3161 \times 10^{-17} & 0.00496066 & -1.3763 \times 10^{-20} & -1.9163 \times 10^{-7} \\ 11.0620173 & -1.1446 \times 10^{-18} & 0.00570064 & 2.8085 \times 10^{-20} & -2.4813 \times 10^{-7} \\ 9.15798913 & 5.0109 \times 10^{-17} & 0.00493009 & -2.3559 \times 10^{-20} & -1.9210 \times 10^{-7} \\ 11.4578629 & -6.0813 \times 10^{-18} & 0.00705818 & 9.2055 \times 10^{-21} & -3.5862 \times 10^{-7} \end{pmatrix} \tag{A.10}$$

A.3 Supplementary plots for the fourfold data.

When looking for mirror symmetry in the fourfold data, we only observed partial mirror symmetry. Below is the full break down of the data set.

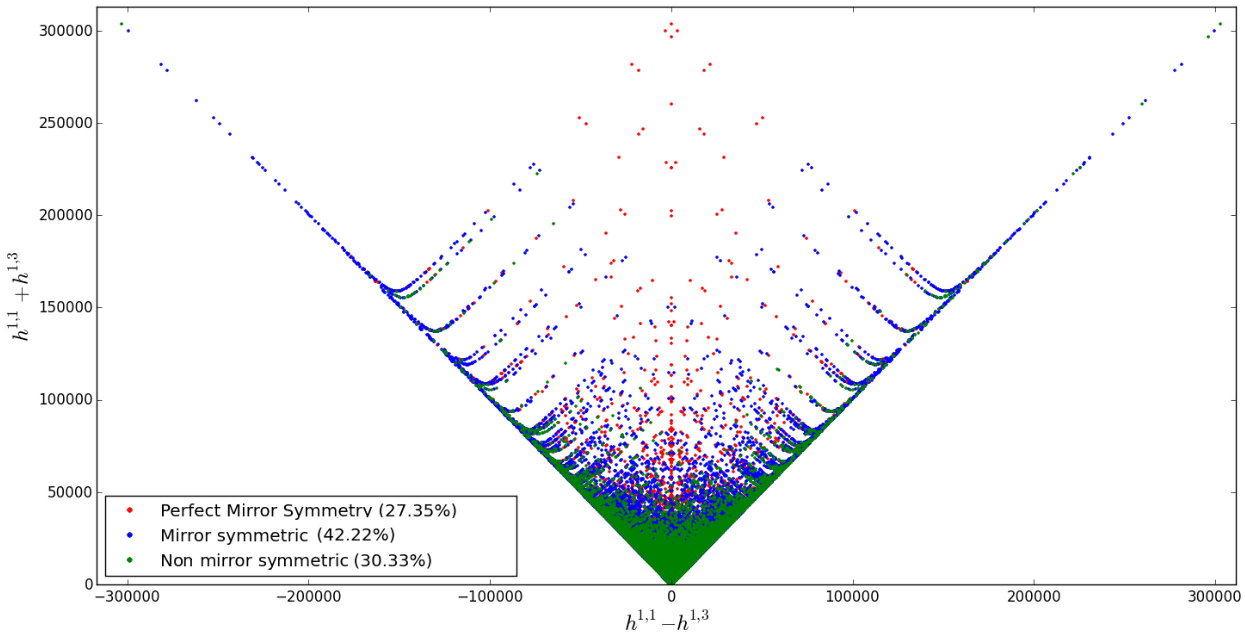


Figure A.5: Mirror symmetry is incomplete in the fourfold data set.

References

[1] P. Candelas, G. T. Horowitz, A. Strominger, E. Witten, "Vacuum Configurations for Superstrings," Nucl. Phys. B **258**, 46 (1985).

[2] P. Candelas, A. M. Dale, C. A. Lutken, R. Schimmrigk, "Complete Intersection Calabi–Yau Manifolds," Nucl. Phys. B **298**, 493 (1988).
P. Candelas, C. A. Lutken, R. Schimmrigk, "Complete Intersection Calabi–Yau Manifolds. 2. Three Generation Manifolds," Nucl. Phys. B **306**, 113 (1988).
M. Gagnon, Q. Ho-Kim, "An Exhaustive list of complete intersection Calabi–Yau manifolds," Mod. Phys. Lett. A **9** (1994) 2235.

[3] N. Hitchin, "Generalized Calabi–Yau manifolds," Quart. J. Math. **54**, 281, arXiv:math.DG/0209099.

[4] M. R. Douglas, "The Statistics of string / M theory vacua," JHEP **0305**, 046 (2003). [hep-th/0303194].

[5] P. Candelas, M. Lynker, R. Schimmrigk, "Calabi–Yau Manifolds in Weighted P(4)," Nucl. Phys. B **341**, 383 (1990).

[6] V. Batyrev, "Dual Polyhedra and Mirror Symmetry for Calabi-Yau Hypersurfaces in Toric Varieties", arXiv:alg-geom/9310003.

[7] Victor V. Batyrev, Lev A. Borisov "On Calabi–Yau Complete Intersections in Toric Varieties", arXiv:alg-geom/9412017

[8] M. Kreuzer and H. Skarke, "On the classification of reflexive polyhedra," Commun. Math. Phys. **185**, 495 (1997) [hep-th/9512204].

[9] A. C. Avram, M. Kreuzer, M. Mandelberg, H. Skarke, "The Web of Calabi–Yau hypersurfaces in toric varieties," Nucl. Phys. B **505**, 625 (1997) [hep-th/9703003].

[10] M. Kreuzer and H. Skarke, "Classification of reflexive polyhedra in three-dimensions," Adv. Theor. Math. Phys. **2**, 847 (1998) [hep-th/9805190].

[11] M. Kreuzer and H. Skarke, "Reflexive polyhedra, weights and toric Calabi–Yau fibrations," Rev. Math. Phys. **14**, 343 (2002) [math/0001106 [math-ag]].

[12] M. Kreuzer and H. Skarke, "Complete classification of reflexive polyhedra in four-dimensions," Adv. Theor. Math. Phys. **4**, 1209 (2002) [hep-th/0002240].

[13] Maximilian Kreuzer,Harald Skarke *Calabi–Yau 4-folds and toric fibrations* `arXiv:hep-th/9701175v1`

[14] J. Gray, A. Haupt and A. Lukas, "Calabi–Yau Fourfolds in Products of Projective Space," Proc. Symp. Pure Math. **88**, 281 (2014).
– "All Complete Intersection Calabi–Yau Four-Folds," JHEP **1307**, 070 (2013) [arXiv:1303.1832 [hep-th]].

[15] L. B. Anderson, F. Apruzzi, X. Gao, J. Gray and S. J. Lee, "A New Construction of Calabi–Yau Manifolds: Generalized CICYs," arXiv:1507.03235 [hep-th].

[16] R. Altman, J. Gray, Y. H. He, V. Jejjala and B. D. Nelson, "A Calabi–Yau Database: Threefolds Constructed from the Kreuzer-Skarke List," JHEP **1502**, 158 (2015) [arXiv:1411.1418 [hep-th]].

[17] R. Davies, "The Expanding Zoo of Calabi–Yau Threefolds," Adv. High Energy Phys. **2011**, 901898 (2011) [arXiv:1103.3156 [hep-th]].

[18] P. Candelas and R. Davies, "New Calabi–Yau Manifolds with Small Hodge Numbers," Fortsch. Phys. **58**, 383 (2010) [arXiv:0809.4681 [hep-th]].

[19] Y. H. He, "Calabi–Yau Geometries: Algorithms, Databases, and Physics," Int. J. Mod. Phys. A **28**, 1330032 (2013) [arXiv:1308.0186 [hep-th]].

[20] L. B. Anderson, Y. H. He and A. Lukas, "Heterotic Compactification, An Algorithmic Approach," JHEP **0707**, 049 (2007) doi:10.1088/1126-6708/2007/07/049 [hep-th/0702210 [HEP-TH]].

[21] M. Gabella, Y. H. He and A. Lukas, "An Abundance of Heterotic Vacua," JHEP **0812**, 027 (2008) doi:10.1088/1126-6708/2008/12/027 [arXiv:0808.2142 [hep-th]].

[22] P. Gao, Y. H. He and S. T. Yau, "Extremal Bundles on CalabiYau Threefolds," Commun. Math. Phys. **336**, no. 3, 1167 (2015) doi:10.1007/s00220-014-2271-y [arXiv:1403.1268 [hep-th]].

[23] L. B. Anderson, J. Gray, A. Lukas and E. Palti, "Heterotic Line Bundle Standard Models," JHEP **1206**, 113 (2012) doi:10.1007/JHEP06(2012)113 [arXiv:1202.1757 [hep-th]].

[24] W. Taylor, “On the Hodge structure of elliptically fibered Calabi–Yau threefolds,” JHEP **1208**, 032 (2012) [arXiv:1205.0952 [hep-th]].

[25] W. Taylor and Y. N. Wang, “A Monte Carlo exploration of threefold base geometries for 4d F-theory vacua,” arXiv:1510.04978 [hep-th].

[26] X. Gao, P. Shukla, “On Classifying the Divisor Involutions in Calabi–Yau Threefolds,” arXiv:1307.1139 [hep-th].

[27] R. Blumenhagen, B. Jurke, T. Rahn, “Computational Tools for Cohomology of Toric Varieties,” Adv. High Energy Phys. **2011**, 152749 (2011) [arXiv:1104.1187 [hep-th]].

[28] J. Gray, Y. -H. He, V. Jejjala, B. Jurke, B. D. Nelson, J. Simon, “Calabi–Yau Manifolds with Large Volume Vacua,” Phys. Rev. D **86**, 101901 (2012) [arXiv:1207.5801 [hep-th]].

[29] P. Candelas, A. Constantin, H. Skarke, “An Abundance of K3 Fibrations from Polyhedra with Interchangeable Parts,” arXiv:1207.4792 [hep-th].

[30] V. Braun, “On Free Quotients of Complete Intersection Calabi–Yau Manifolds,” JHEP **1104**, 005 (2011) [arXiv:1003.3235 [hep-th]].

[31] P. Candelas, X. de la Ossa, Y. H. He and B. Szendroi, “Triadophilia: A Special Corner in the Landscape,” Adv. Theor. Math. Phys. **12**, 429 (2008) [arXiv:0706.3134 [hep-th]].

[32] M. Kreuzer and H. Skarke, “PALP: A Package for analyzing lattice polytopes with applications to toric geometry,” Comput. Phys. Commun. **157**, 87 (2004) [math/0204356 [math-sc]].

[33] A. P. Braun, J. Knapp, E. Scheidegger, H. Skarke and N. O. Walliser, “PALP - a User Manual,” arXiv:1205.4147 [math.AG].

[34] The On-Line Encyclopedia of Integer Sequences, `http://oeis.org`, Number A090045.

[35] Y. H. He, S. J. Lee and A. Lukas, “Heterotic Models from Vector Bundles on Toric Calabi–Yau Manifolds,” JHEP **1005**, 071 (2010) [arXiv:0911.0865 [hep-th]].

[36] M. Lynker, R. Schimmrigk and A. Wisskirchen, “Landau-Ginzburg vacua of string, M theory and F theory at c = 12,” Nucl. Phys. B **550**, 123 (1999). [hep-th/9812195].

[37] DH Stamatis “Six Sigma and Beyond: Statistics and Probability Vol 3, CRC Press; 1 edition (August 28, 2002)”