9-2014

Fitting multiple bell curves stably and accurately to a time series as applied to Hubbert cycles or other phenomena

James A. Conder
conder@geo.siu.edu

The companion code to this article is available at http://opensiuc.lib.siu.edu/geol_comp/4/.

Recommended Citation

Conder, James A. "Fitting multiple bell curves stably and accurately to a time series as applied to Hubbert cycles or other phenomena." *Mathematical Geosciences* 47, No. 6 (Sep 2014): 663-678. doi:doi:10.1007/s11004-014-9557-7.

Fitting multiple bell curves stably and accurately to a time series as applied to Hubbert cycles or other phenomena

by James A. Conder[2]

[1] Received____Dec 30, 2013____; accepted_____________________.

[2] Department of Geology, Southern Illinois University, Carbondale, IL 62901 U.S.A.; e-mail: conder@geo.siu.edu

Corresponding Author:

J.A. Conder

Department of Geology

Southern Illinois University

Carbondale, IL, 62901 U.S.A.

Phone + 1 618 453 7352

Fax +1 618 453 7393

e-mail: conder@geo.siu.edu

Abstract

Bell curves are applicable to understating many observations and measurements across the sciences. Relating Gaussian curves to data is a common because of its relation to both the Central Limit Theorem and to random error. Similarly, fitting logistic derivatives to oil or other non-renewable resource production is common practice. Fitting bell curves to a time series is an inherently non-linear problem requiring initial estimates of the parameters describing the bell-curves. Poor estimates lead to instability and divergent solutions. Fitting to a cumulative curve improves stability, but at the expense of accuracy of the final solution. Jointly fitting multiple bell curves is superior to extraction of curves one at a time, but further exacerbates the non-linearity. Including both the cumulative data and the bell-curve data within the inversion, can exploit the greater stability of the cumulative fit and the greater accuracy of a direct fit. The algorithm presented here inverts for multiple bells by combining cumulative and direct fits to exploit the best features of both. The versatility and accuracy of the algorithm are demonstrated using two different Earth Science examples: a seismo-volcanic sequence recorded by a hydrophone array moored to the seafloor and U.S. coal production. The MatLab function used here for joint curve determination is included in the online manuscript complementary material.

KEY WORDS: Gaussian, Hubbert multicycles, Curve-fitting

1. Introduction

One of the most basic procedures for extracting information from a time series of discrete data points is to fit the data to a curve of known form, thereby reducing the data to a few describable parameters. By reducing a large number of ordered data to a few parameters, the system is not only easier to describe, but simpler to understand, and may provide some predictive capability, for example dealing with resource production (e.g., Rutledge 2011). Ideally, a fit is over-determined (i.e., more data than parameters describing the curve). In simple form this may be a fit to a line, but any function with a set of independent parameters can be used.

A typical measure for the best-fitting curve of a given form is to find the set of parameters that minimizes the sum of the squares of the misfit of the data (SSE) to the adopted curve, stated as

$$SSE = \sum_i (d_{ipre} - d_i)^2 \ , \tag{1}$$

where d_i denotes the ith datum and d_{ipre} denotes the predicted value of the ith data point using the adopted curve. A convenient way to normalize SSE for easy comparison for different solutions is using the root mean square error (RMSE).

$$RMSE = \sqrt{\frac{SSE}{N}} \ , \tag{2}$$

where N is the number of data.

Depending on the data being analyzed, one may not want to use a line or higher order polynomial to fit data, but rather some other basis function, like a sine or bell curve. Any function with the form y ~ exp(-t^2) will form a bell shaped curve. In particular, the Gaussian or Normal - sometimes termed the bell curve - has wide applicability in the physical, natural, and social sciences because of both the Central Limit Theorem (Kirkup and Frenkel 2006, p143-150) and its relevance to random error. However, fitting functions such as bell curves to data is an inherently non-linear problem requiring initial estimates of the desired parameters. If the estimates are not precise enough, calculations may not converge to a realistic solution. One could fit a cumulative bell-curve, which is monotonic, and therefore less sensitive to the precision of initial parameter estimates, but is less sensitive to the exact location of the peak and therefore loses some accuracy in the final solution. In some instances, more than one bell-curve may be desired to fit to a time series. Including additional curves further compounds issues of stability and accuracy. Note, the term time series in this paper refers to any set of ordered data whether or not the independent variable is explicitly time.

The focus of this paper is to exploit the advantages of fitting cumulative and standard data simultaneously, thereby improving stability of fitting bell curves without cost to accuracy. The method provides the most gains when fitting multiple curves to a data set. To show the value of the constructed algorithm, the method is applied to a couple of examples relative to the Earth Sciences: seismo-volcanic detections on a hydrophone array and U.S. coal production. A MatLab function for fitting Gaussians or logistic derivatives to a time series incorporating the methods described here can be found in the complementary materials.

2. Bell-curve fitting

The Gaussian has the form

$$f = A\exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) , \qquad (3)$$

where *t* is the independent variable, μ is the location of the curve peak, *A* is the amplitude of the peak, and σ is a width parameter, often noted as standard deviation.

The major issues in curve fitting algorithms for over-determined problems are 1) stability – the algorithm converges to a solution with a reasonable set of parameters, 2) accuracy – the solution is actually minimizing misfit and that the solution found is for a global minimum and not a local minimum, and 3) speed of the calculation.

An always stable approach to finding a minimum SSE for any problem is a grid search over all the parameters, looking for which set of values has the smallest SSE (e.g., Conder and Forsyth 2000). However, the time necessary for this method can become rapidly prohibitive as more parameters are added. In the case of fitting bell curves, every additional curve adds 3 parameters to be found. So, even limiting to three curves requires nine dimensions to search. While speed is not taken as a high priority in this paper as stability and accuracy, it should be explicit that for any methodology to be of value, it must have much more practicable time to completion than a grid search.

To get away from brute force grid search methods, finding minimum solutions typically relies on inverse theory (Menke 2012) and the simple equation

$$Gm = d. \qquad (4)$$

m is the set of parameters describing the desired curve, *d* denotes the vector of discrete data values, and *G* is a matrix containing the partial derivatives of the predicted data relative to the model parameters. To find *m* from a known vector *d*, one needs to find a suitable inverse, G^{-g}, leading to

$$m = G^{-g}\,d. \tag{5}$$

For linear over-determined problems, the well known least squares solution is obtained by pre-multiplying both sides of Eq. 4 by the transpose of *G*. Solutions for best-fitting functions with a linear set of parameters, where the derivative of with respect to any given parameter does not depend on other parameters, are naturally stable and relatively easy to determine accurately and quickly.

To fit one bell curve to a time series, the problem may be fully linearized by relating the log of the data to a quadratic and fitting the resultant quadratic. So,

$$\log(d) = at^2 + bt + c. \tag{6}$$

For a Gaussian,

$$a = -1/(2\sigma^2), \tag{7}$$

$$b = \mu/\sigma^2, \tag{8}$$

and

$$c = \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2. \tag{9}$$

While stable, rapid, and requires no *a priori* information about the parameters, this does not strictly minimize the misfits as shown in Eq. 1, but rather the log misfits. Even more crucially, this linearization method cannot fit more than one bell curve to a time series. For data with more than one bell present in the signal, the result will subsequently give only a gross representation of all the data together as one bell curve.

Without the linear transform above, fitting Gaussian curves is inherently non-linear, requiring initial estimates of the desired parameters to be solved to build the G-matrix. The solution may then be iterated with progressively improved estimates of the parameters. The better the estimates to the true values, the more accurate the derivatives within *G*, and the faster convergence will be reached. The more interdependent the derivatives are on the various parameters, the more accurate the initial estimates need to be to reach convergence. In particular, the non-monotonic shape of a bell curve tends the problem towards instability without reasonably careful seeding of initial estimates.

2.1 Fitting cumulative curves

One way to reduce the sensitivity of initial seed values on stability is to instead fit the cumulative data to a cumulative Gaussian, *F*.

$$F = \sqrt{\frac{\pi}{2}} A\sigma \left\{ 1 + erf\left(\frac{t-\mu}{\sqrt{2}\sigma} \right) \right\} \quad . \qquad (10)$$

The parameters A, σ, and μ are the same as in Eq. 3 and *erf* is the error function. The advantage of fitting the cumulative rather than the standard curve is that the cumulative is monotonic, and therefore can tolerate a wider range of parameter seed values and still reach convergence. Patzek and Croft (2010) and Anderson and Conder (2011) took this approach to fit multiple bell curves to oil production from various parts of the world. Although more stable, the fit is less sensitive to the precise location of the peak. So, using a cumulative Gaussian comes at some cost to accuracy.

For instance, as fitted curve peaks are often offset slightly relative to data peaks using a cumulative approach because of moderate excesses or deficits in one of the tails, Anderson and Conder (2011) grid searched in final positions of the curves.

2.2 Information density

A useful way to look at the difference in strengths and weaknesses between fitting standard and cumulative curves is to look at which data provide the most information to the different curve fits. The density of information provided by each of the data to the solution can be found by decomposing the partial derivative matrix *G* in Eq. 4 into two eigenvector matrices, *U* and *V*, and a matrix *Λ* with the eigenvalues of the system along the diagonal (Jackson 1972). So,

$$G = U\Lambda V^{T}. \tag{11}$$

The eigenvector matrices, *U* and *V*, describe the data space and model space respectively. The information density provided by each of the data is found along the diagonal of the matrix *D*, where

$$D = UU^{T}. \tag{12}$$

A look at the amount of information provided by each of the data helps for illustration. Figure 1 shows starting seed curves for fitting UK oil production from 1965 to 2008 to two Gaussians. Figure 1(b) shows the information density associated with each datum for a standard fit while 1(d) shows the information density for a cumulative fit. The density of information provided from each of the data is markedly different for the cumulative and standard fits. The data with the greatest importance for the standard fit are those that lie near the peak and near the inflection points of the estimated curves. In essence, nearly all the information is contained between the inflection points. The data that lie within the tails of the seed curves provide little information and thereby add little to the inverse calculation. This gives great sensitivity to the exact location of the peaks – assuming the actual peaks are reasonably described by the estimated peaks - but the inverse problem can easily become unstable if the seeded peaks do not adequately represent the actual peaks in the data.

In contrast, the cumulative data information densities are more evenly distributed because of the monotonic shape of the curve. The greatest density of information in the cumulative data is contained by those data near inflections in the cumulative slope. Importantly, the tails also contribute information about the desired parameters (Fig. 1(d)). The more evenly distributed density makes for a more stable inversion by including information from the extremities, but at a cost of only moderate sensitivity to the exact locations of the peaks in the data, and more sensitivity to excess or deficit accumulations in the tails. Recognizing these differences suggests that using elements of curve fits to both cumulative and standard data could create a more robust algorithm than using either separately.

The derivative of a logistic curve is another bell curve with frequent use, especially as applied to the production of natural resources with finite reserves, often termed Hubbert cycles (Deffeyes 2008), sometimes requiring multiple cycles to fit the data (Nashawi et al. 2010; Patzek and Croft 2010). The discussion above applies equally well to the logistic derivative considering only a slightly different form of the curve to fit. The logistic derivative and its cumulative (the logistic) have the following forms

$$f = A\exp\left(\frac{-\tau(t-\mu)}{\left(1+\exp(-\tau(t-\mu))\right)^2}\right) , \tag{13}$$

and

$$F = \frac{A}{\tau\left(1+\exp(-\tau(t-\mu)\right)^2} , \tag{14}$$

with A and μ having analogous definitions to those above, and τ analogous to $1/\sigma$.

3. Method

The method presented here aims to fit multiple curves simultaneously, stably, and accurately with minimal parameter seeding required by the user. The most common way to address stability is to seed the parameters with starting guesses sufficiently close to actual values. This can be crucial for fitting non-monotonic curves since poor starting estimates will require pushing some portions of the data through regions of poorer fits to get to the best-fitting curve, which may result in instability. Monotonic curves are less susceptible to this issue.

The algorithm presented here takes advantage of the greater stability of a cumulative fit and the greater accuracy of a direct fit by constraining the fits for Eqs. 3 and 10 (or 13 and 14) simultaneously to each desired curve. In essence, each datum is given two values: a cumulative value and a non-cumulative value. These values are then fit to the sum of a set of component curves, while requiring the same parameters for both the cumulative and non-cumulative representations. This approach not only exploits the differences in information density across the data for the two curves, but by using the information jointly, the total amount of information available for determining the best-fitting parameters is effectively doubled.

As this is an iterative inversion, initial estimates (seeds) of the parameters are necessary for calculating the derivatives with subsequently improved values for the parameters used to iterate to a solution. The less stable the problem, the closer the seed needs to be to the actual value to ensure convergence. To free the user of initially estimating the parameters, a few different approaches may be taken to determine starting seeds to use.

One simple auto-seeding approach is to assume a set of equally spaced identical curves. Each curve then provides 1/M portion of the cumulative curve (*M* being the number of Gaussians to fit). As the area occupied by a particular curve scales with the sum of the products of the amplitudes, *A*, and widths, σ, the seeding of these are best accomplished jointly.

Practice shows that assuming widths about a tenth of 1/M of the width of the time series works well. The idea being that the time series is adequately capturing the curves of interest and curve width is not on the order of the time series width or larger. The factor of one tenth helps minimize the overlap between seed curves, and gives the curves room to expand or contract without immediately reaching widths on the order of the time series.

With curve widths established, the amplitudes of the curves can easily be assigned by

$$A = \frac{F_{last}}{M\sigma\sqrt{2\pi}} \ , \tag{15}$$

where F_{last} is the last cumulative datum used as a proxy for the cumulative value at time infinity.

Alternatively, the data may be used directly to help with auto-seeding. For instance, peaks in the data may be tagged as locations for initial guesses of final peaks as well as using peak heights as starting amplitudes. Finding widths is less straightforward to extract from the data. If the data consists of a few well-defined curves, finding the zeros of the second derivative of the data will show the inflection points and therefore the widths may be determined (Goshtasby and O'Neill 1994). However, this method fails for overlapping curves, which is a principal aim of this study. Fortuitously, using width seeds similar to that described above tends to be sufficient.

Another potentially useful way to autoseed is simply through random seeding. A key advantage for using random seeds is the ability for doing the problem a number of times to identify the presence of local minimum solutions. If the same solution is converged to with various sets of seeds, it is in all likelihood a global minimum. If the solution found depends on the seeds given, there are local minima present. With enough sets of random seeds, it is possible to find all the minima and identify the global minimum among them.

No matter which seeding approach is used, it still may be the case (especially when fitting several bell curves) that one or more parameters become unrealistic or a time peak leaves the data space. Fortunately, it is easy enough to flag these instances and randomly reseed that curve back into the realistic model space.

It is often beneficial to put more relative weight on the cumulative data for early iterations and more on direct derivatives for later iterations. This aids stability early when estimated parameters likely need significant shifting and aids accuracy when approaching convergence and sensitivity to peaks and inflection points is most important.

Once seeds are established and estimates of the partial derivatives with respect to each parameter are calculated, a solution may be iterated using any of a number of inverse methods, such as standard least squares, damped least squares (Goshtasby and O'Neill 1994), LSQR (Paige and Saunders 1982), LSMR (Fong and Saunders 2010) or Singular Value Decomposition (SVD) (Jackson 1972). SVD is the most robust, but tends to be slower than other methods as it requires a complete decomposition of the partial derivative matrix to create the inverse. Unless the G matrix is poorly conditioned, when the robustness of SVD is most beneficial, LSQR is used here. LSMR is as fast and reliable (and sometimes more so) than LSQR, but as yet does not come standard with MatLab.

4. Results

All results presented in this section, other than the explicitly linearized case, use the method presented of jointly minimizing the cumulative and non-cumulative curve misfits.

4.1 Hydrophone volcano seismic detections

Bohnenstiehl et al. (2014) recorded seismo-volcanic signals on a hydrophone array moored to the ocean floor in the Lau Basin between Fiji and Tonga. Signals were found to come from several dominant azimuths about the array pointing to specific volcanoes. The numbers of detections binned by azimuthal direction to the signal source tend to behave in a Gaussian manner with energetic bands having widths of a few tenths of degrees. Most bands are clear and easy to fit with a single Gaussian to determine a precise azimuth and event frequency (combining peak height and width) to relate to individual volcanoes. However, a few bands are more complex. For instance, the azimuthal band pointing towards Niuafo'ou Island region in the northern part of the basin appears to be a composite signal (Fig. 2), and provides a useful example for exploring different approaches to fitting bell curves.

Results of various curve fits are shown in Table 1. In this case, a linearized fit using Eqs. 6 to 9 does a poor job of representing the data (Fig. 2). Fitting directly to a single Gaussian does a better job, but leaves a significant portion of the signal unfit, suggesting that a second curve might be warranted. A standard approach is to sequentially fit component curves by removing the first curve from the data and fit the residual to a new curve (Goshtasby and O'Neill 1994). As can be seen in Table 1, this improves the RMSE by more than 40%. Using an F-test (e.g., Anderson and Conder, 2011), the improvement warrants the addition of the second curve at a 99.2% confidence level. Yet, fitting two component curves simultaneously (minimizing the combined misfit) reduces the RMSE by nearly 80%, warranting a second curve at better than 99.999% confidence.

Clearly, the jointly determined component curves give a better description of the data than those found sequentially. In this case, the difference in the approaches affects the interpretation of the activity of the natural system. Not only would a fraction of the energy emanating from one volcano be erroneously attributed to another, but the form of the secondary Gaussian would suggest a different emanation pattern. The secondary Gaussian derives from signals generated over the length of a boomerang shaped volcanic edifice. By not fitting the component curves jointly, the secondary Gaussian would suggest the signals were only generated near the summit of the edifice (Fig. 3).

Using a sequential approach for something like data compression may be less problematic as one would be looking for data characterization as a whole rather than at details between curves. Still, a joint determination of the component curves captures more of the overall character of the data meaning less misfit and better compression.

4.2 U.S. coal production

It is not uncommon for multiple local minimization solutions to exist in non-linear problems, tending to increase with both complexity of the time series and the number of parameters to be fit. Figure 4 shows U.S. coal production. There appear to be three different cycles in the data, with two noticeable peaks prior to 1960 and rapidly increasing production after 1970, possibly peaking around 2005. The middle peak is the most challenging to fit as only data from 1935 to 1950 contribute significantly to the peak. Using logistic derivatives, three different solutions can be found with different initial seeds. The three solutions have RMSE values of ~33, 44, and 50. Clearly, the solution with RMSE of 33 is the global minimum, while the other two are local minima. The primary difference between the solutions is whether the middle peak is found during convergence. Only the lowest RMSE solution closely fits the middle peak (Fig. 4). The second solution treats the first two peaks as one broad peak with an early hump, and the third uses two of the available curves to fit the final (incomplete) peak.

Using 100 sets of random seeds, the global minimum is found 32 times and the others 19 and 49 times, respectively. The RMSE 50 solution that is found most often uses two of the solution curves to build the third cycle. As the third cycle is only one-sided in the data as well as the largest, there is more leeway on how to construct it with a low SSE, making that solution easier to find. Of course, some user input may help point the solution towards finding the middle peak. Using one user input time seed of 1945 results in the lowest RMSE solution being found 72 times to 9 and 19, respectively. Using three time seeds of 1917, 1945, and 2003, improves the frequency of finding the desired solution even further, with the global minimum solution found 89 times, the second solution (RMSE 44) 11 times and the third solution (RMSE 50) 0 times.

5. Discussion

The number of possible component curves to be extracted from a time series is limited by the number of data. As there are three parameters per component curve, the number of possible component curves extracted may not exceed (N-1)/3 and still maintain an over-determined problem. Of course, in practical terms there will be far fewer curves desired. The more component curves used, the more local minima likely to exist. Similarly, the more component curves used, the more likely that noise in the data will be fit rather than true signal.

A few research-grade software packages have good utility for fitting bell curves to a time series, such as SAS PROC NLIN software (Copyright, SAS Institute Inc), the SPSS Curve Estimation routine (IBM Corp., 2013), SciPy's optimize.curvefit (Millman and Aivazis 2011), and the SOLVER module in Microsoft Excel (Walsh and Diamond 1995). Each has limitations and advantages. The SPSS package has an easy to use GUI and can fit a variety of different curves, but will only fit a single curve at a time. The others are more flexible in the number of parameters that may be fit at one time, but require correspondingly more effort on the part of the user. The SOLVER module has been used to good success by several researchers in the Hubbert curve modeling community (e.g., Rutledge 2011). It can minimize a value like SSE by adjustment of any number of designated parameters. The power of adaptable software like SOLVER or the optimize.curvefit module of SciPy is that they calculate numerical derivatives for the various parameters and therefore can be used for virtually any function, though with all the caveats of stability and finding local minima. Tests with SOLVER show comparable, but slightly higher, RMSEs relative to the examples presented in this study - when convergence is reached (typically requiring fairly precise starting estimates of parameters). The algorithm presented here is designed to be simple to use (accessed as a MatLab function) and to expand stability for traditionally low-stability problems without sacrificing accuracy. The greater stability allows for convergence of jointly fit multiple curves with relaxed precision of starting estimates of the parameters to be calculated, often to the point of allowing a simple auto-seeding to do the job. This requires less work on the part of the analyst beforehand, as less *a priori* information about the desired parameters must be deduced, as well as reducing the sensitivity of the final results to user bias through initial seeds.

Both Gaussians (Brandt 2007; Liu et al. 2012) and derivatives of logistics (Nashawi et al. 2010;

Gallagher 2011) have been applied to production data of fossil fuels and other resources. The choice of one basis function over the other is somewhat debatable. That a logistic derivative curve would reasonably describe the production data makes some intuitive sense as a logistic was designed to track carrying capacity for populations of species within an ecosystem supported with limited resources. In this case, the carrying capacity would be analogous to the total extraction or consumption of the finite resource. So the derivative would track the production or consumption in time – convenient since fossil fuel extraction is typically tabulated in annual increments.

Likewise, there is some intuitive value to using a Gaussian if one considers the Central Limit Theorem at work with many concurrent extraction operations combining to a Gaussian-like extraction curve. For the Central Limit Theorem to be applicable, any one production system should behave relatively independently of the others in the same region. This may be a reasonable assumption for production with relative political stability and limited economic barriers, such as in the U.S. or Norway (Laherrere 2000). Although, even those may fall short of actual independence with individual producers likely responding similarly to the same external economic stimuli.

In practice, the difference in goodness of fit between the two different curves is often marginal (Patzek and Croft 2010). Using Gaussians for the three curve fit for U.S. coal production results in a best RMSE of 35.9 compared to 32.7 for logistic derivatives. In this case, the difference of fit is significant. While an F-test is useful for looking at the importance of adding additional parameters to a particular model, the corrected Akaike Information Criterion (AICc), stemming from information theory, can straightforwardly compare models even if non-nested (Burnham and Anderson 2004). For a least squares case,

$$AICc = N\log\left(\frac{SSE}{N}\right) + 2K + \frac{2K(K+1)}{N-K-1}\ , \tag{16}$$

where *K* is the number of model parameters plus one for a given model. The difference in *AICc* between two or more models shows the extent of information loss from one to another, which is a combination of goodness of fit in the first term and a model complexity penalty in the following terms. The best model has the lowest *AICc* (least information loss). Even more beneficial than picking a best model is that each model can be weighted in a probability sense as the probable true model among the given possibilities.

$$w_i = \frac{\exp(-0.5\Delta_i)}{\sum_j \exp(-0.5\Delta_j)}\ , \tag{17}$$

where w_i are the probabilities which sum to one, and Δ_i is the difference in *AICc* between the *i*th model and the lowest *AICc* in the set. For example, if three models have w_i = 0.5, 0.4, and 0.1. The first and second model would be 5 and 4 times more probable to be the true model than the third, but the first would only be 20% more probable than the second. For the Gaussian and logistic derivative 3 component curve cases for U.S. coal, the logistic derivative has a $w_i >$ 0.99999. Much of the information loss of the Gaussian relative to the logistic comes in a systematically poorer fit in the initial tail (Fig. 5). This w_i may not be particularly precise in that Eq. 16 assumes normally distributed errors across the fit, whereas the errors scale with production. But, even omitting data prior to 1907 to remove the most non-normally distributed portion, w_i is still greater than 0.98 for the logistic, because of systematic improvement to fits of the peaks and valleys throughout the 20th century (Fig. 5), demonstrating a significantly superior fit for logistic derivatives over Gaussians for U.S. coal production. Importantly, it should be kept in mind that using one basis function over the other carries a different set of implicit assumptions as to why that curve is appropriate, which should be used as primary criterion as to which curve to use regardless of statistical criteria (Burnham and Anderson 2004).

Of course, there are regions that may not fit either set of curves reasonably well. For example, Libya may not be expected to have much independence among producers to be Gaussian or act economically and technologically unfettered enough to act like a logistic. Indeed, despite efforts to interpret production of Libya and other complex production curves in terms of Hubbert cycles (Nashawi et al. 2010), an unreasonably large number of cycles must be included to fit the production data (Anderson and Conder 2011). Beyond using statistical tests, a red flag that too many cycles are likely imposed or that some other set of basis functions should be called for is finding smaller cycles contained within larger cycles to fit the data. For a Gaussian approach, allowing cycles within cycles violates the Central Limit Theorem assumption because internal cycles are what make up the larger Gaussians. That is, the data should not distinguish between individual components, as it is the individual components that make up the Gaussian. For a logistic approach, one could expect a suite of logistics following a power-law or log-normal distribution (Sorrell et al. 2012) with few large and many smaller ones. In a time series where size of the scatter scales with the signal (e.g., annual production), small curves will be largely lost in the scatter of the larger curves, with any imposed smaller curves largely fitting the noise of the system.

6. Conclusion

Gaussian, Hubbert, and other bell curves serve as useful basis functions for describing and understanding many time series. However, because of the inherent non-linearity and instability of the problem there are few tools available for simultaneously fitting multiple component curves reliably and robustly. Much of the difficulty lies in estimating the desired parameters accurately enough to then solve for them. Because the data provide a more even distribution of information about the desired parameters for cumulative curves than peaked curves, fitting data to cumulative curves eases the necessary degree of precision for estimating parameters beforehand. However, fitting cumulative curves comes at some cost to accuracy of fitting the peaks. Jointly solving for the parameters that best fit both the standard curves and the cumulative curves concurrently can improve the stability without sacrificing accuracy. This is especially true if early iterations more strongly weight cumulative curves and later iterations more strongly weight standard curves. If one or more parameters become unreasonable through instability, they can be easily flagged and reseeded within the relevant portion of the time series, strongly increasing the likelihood for convergence even for randomly seeded parameter estimates.

A common issue, particularly for complex data and/or the inclusion of many component curves, is the likelihood of multiple sets of curves that locally minimize SSE. In a case like U.S. coal production, it is straightforward to recognize whether a solution is a global minimum as there are three clearly distinct peaks to fit, but it may not be as obvious for many cases. The ability to quickly randomize the times of the seed curves provides a useful way for exploring the various local minima present. Particularly tough cases can still be cracked using a mix of well-valued seeds and autoseeds.

Acknowledgements

I thank Del Bohnenstiehl for the seismo-volcanic azimuthal detection analysis. I thank Ken Anderson whose interest in Hubbert cycles first prompted my thoughts of the problem of simultaneously fitting multiple bell curves to complex systems. This work was improved by the thoughtful comments and suggestions of two anonymous reviewers as well as the journal editor, Roussos Dimitrakopoulos. This work was partially supported by National Science Foundation grant OCE-0825424.

References

Anderson KB, Conder JA (2011) Discussion of Multicyclic Hubbert Modeling as a Method for Forecasting Future Petroleum Production. Energy and Fuels 25:1578–1584. doi: 10.1021/ef1012648

Bohnenstiehl DW, Dziak RP, Matsumoto H, Conder JA (2014) Acoustic Response of Submarine Volcanoes in the Tofua Arc and Northern Lau Basin to Two Great Earthquakes, Geophysical Journal International, 196:1657-1675. doi: 10.1093/gji/ggt472

BP Statistical Review of World Energy 2013, http://www.bp.com/en/global/corporate/about-bp/statistical-review-of-world-energy-2013/review-by-energy-type/oil/oil-production.html (accessed Oct 17, 2013)

Brandt AR (2007) Testing Hubbert. Energy Policy 35:3074–3088. doi: 10.1016/j.enpol.2006.11.004

Burnham KP, Anderson DR (2004) Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociol Methods Res 33:261–304. doi: 10.1177/0049124104268644

Conder JA, Forsyth DW (2000) Do the 1998 Antarctic Plate earthquake and its aftershocks delineate a plate boundary? Geophys Res Lett 27:2309–2312. doi: 10.1029/1999GL011126

Deffeyes KS (2008) Hubbert's Peak: The Impending World Oil Shortage, 2nd ed. Science (80-) 295:208.

Fong D, Saunders M (2010) LSMR: An iterative algorithm for sparse least-squares problems. Perform Comput 4026:21.

Gallagher B (2011) Peak oil analyzed with a logistic function and idealized Hubbert curve. Energy Policy 39:790–802. doi: 10.1016/j.enpol.2010.10.053

Goshtasby A, O'Neill WD (1994) Curve Fitting by a Sum of Gaussians. CVGIP Graph Model Image Process 56:281–288. doi: 10.1006/cgip.1994.1025

IBM Corp. (2013). IBM SPSS Statistics for Macintosh, Version 22.0. Armonk, NY: IBM Corp

Jackson DD (1972) Interpretation of Inaccurate, Insufficient and Inconsistent Data. Geophys J Int 28:97–109. doi: 10.1111/j.1365-246X.1972.tb06115.x

Kirkup L, Frenkel RB (2006) An Introduction to Uncertainty in Measurement: Using the GUM (Guide to the Expression of Uncertainty in Measurement). 248.

Laherrere JH (2000) Learn strengths, weaknesses to understand Hubbert curve. Oil Gas J 98:Special Report.

Liu C, Zhu J, Wang S, Liu W (2012) Oil production forecasts and their uncertainty analyses. Bull Can Pet Geol 60:158–165. doi: 10.2113/gscpgbull.60.3.158

Menke W (2012) Geophysical Data Analysis: Discrete Inverse Theory, Third Edition: MATLAB Edition (International Geophysics Series). 330.

Millman KJ, Aivazis M (2011) Python for Scientists and Engineers, Comput. Sci. Eng., 13, 9. doi: 10.1109/MCSE.2011.36

Nashawi IS, Malallah A, Al-Bisharah M (2010) Forecasting World Crude Oil Production Using Multicyclic Hubbert Model. Energy & Fuels 24:1788–1800. doi: 10.1021/ef901240p

Paige CC, Saunders MA (1982) LSQR - An Algorithm for Sparse Linear-Equations and Sparse Least-Squares. ACM Trans Math Softw 8:43–71.

Patzek TW, Croft GD (2010) A global coal production forecast with multi-Hubbert cycle analysis. Energy 35:3109–3122. doi: 10.1016/j.energy.2010.02.009

Rutledge D (2011) Estimating long-term world coal production with logit and probit transforms. Int J Coal Geol 85:23–33. doi: 10.1016/j.coal.2010.10.012

SAS Institute Inc. (2000) SAS OnlineDoc, Version 8, Cary, NC: SAS Institute Inc.

Sorrell S, Speirs J, Bentley R, et al. (2012) Shaping the global oil peak: A review of the evidence on field sizes, reserve growth, decline rates and depletion rates. Energy 37:709–724.

Walsh S, Diamond D (1995) Non-linear curve fitting using microsoft excel solver. Talanta 42:561–572.

Figure captions

Fig. 1 Data information density comparisons for standard and cumulative Gaussian fits; (a) UK oil production from 1965 to 2008 as representative time series (open circles); Black lines show two seed curves; (b) density of information (crosses) carried by data; Vertical dashed lines show the inflection points of summed curve; (c) and (d) are the same, but for cumulative Gaussians; data from BP Statistical Review of World Energy 2013

Fig. 2 Seismo-volcanic detections binned by azimuth (open circles) from a hydrophone array moored in the central Lau Basin; (a) fits to a linearized Gaussian (dash-dot) and a standard Gaussian fit (dashed); (b) fitting more than one Gaussian by sequential fitting; Component primary and secondary Gaussians (dashed) with sum of two (thin black line); (c) two Gaussians jointly fit to the data; data from Bohnenstiehl et al., 2014

Fig. 3 Seafloor near Niuafo'ou Island; Colorbar denotes seafloor depth in meters; Azimuthal directions from the hydrophone array point to where the signals in Figure 2 emanate; White lines bound the 2-sigma width of the primary Gaussian in Table 1 with the main peak likely emerging from the seamount marked *A*; Dashed lines show the 2-sigma width for the secondary Gaussian using a sequential approach and solid black lines show 2-sigma width of the secondary Gaussian when fit jointly, likely deriving from the boomerang shaped edifice marked *B*

Fig. 4 U.S. coal production from 1800 to 2010 fit by three Hubbert curves; Three discrete peaks are visible in the data (open circles); Middle peak is defined only by production from 1935 to 1950. Three separate local minima are found using different seeding; Composite curves are dashed, full solutions are solid lines

Fig. 5 Comparison of data predictions from three cycle Hubbert and Gaussian models for U.S.

coal production; Upward bars denote better fit for logistic derivatives, downward bars denote better fit for Gaussians; Initial tail is systematically better fit by logistic derivatives; Data from 1907 onward only (vertical dashed line) are still fit statistically better with logistic derivatives