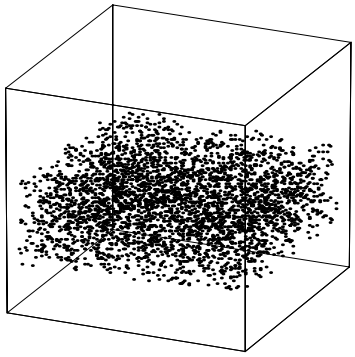
10.18 Surfaces – Approximation of Irregular Noisy Data

The trickiest problems with surface fitting are high dimensional irregular noisy data. Interpolation methods are generally not applicable, but there are few cases where they can provide some assistance.



Approximation Methods are useful in some cases, but require extreme care. Figure 10.18 – 1 illustrates a fictitious dataset with both a linear and a higher order polynomial LS fit, using the formulas and code from Section 10.16. Even by inspection it is already obvious that the fitted surfaces are suspect at best.

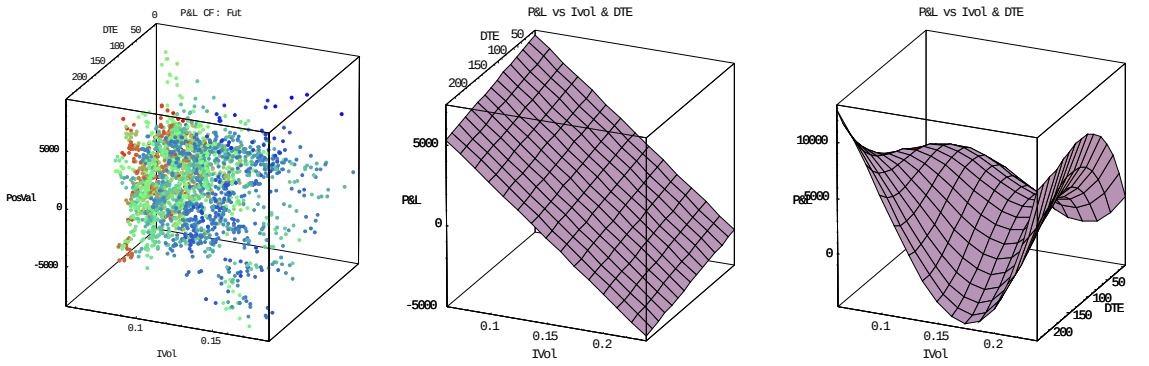


Figure 10.18 – 1. Irregular and noisy trading data with linear and higher order LS approximating surfaces.

This is a very grave matter with high dimension data fitting, since past a few dimensions, visualisation methods are unavailable[46]. That puts reliance on purely algebraic investigations, and that can be a very dangerous situation if applied blindly and in isolation.

In many such cases, a slightly less elegant, but much more reliable approach may be best. In particular, quasi-statistical methods that rely on "bins analysis" may not have the theoretical framework of MLE, but they are very robust and without "hidden" or "internal decisions" that are beyond your access.

The basic idea is to create a sequence of "bins" and apply statistical measure to the data within each bin. Each bin then provides an "average" data point, plus the related statistics as required for "acceptability" and Goodness of Fit (GoF) analysis.

---

[46] It is possible to produce the effect of visualisation of up to the equivalent of 7 dimensions with specialised methods, as for example in the **Pr/rO** ® software's analysis subsections (see www.arbitrage-trading.com).

Figure 10.18 – 2 a) shows a schematic illustration a "binned" volume. Figure 10.18 – 2 b) shows a surface (interpolated) on the average points, and this can be seen to more faithfully represent the "trend" in the data from above, as compared to the traditional LS based results seen above.

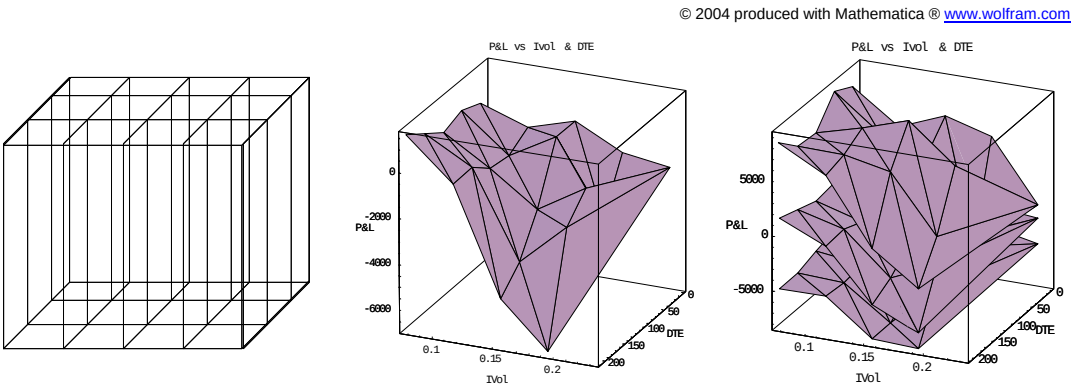


Figure 10.18 – 2. a) "bins" for creating "local" statistical fitting, b) Interpolation of bin averages, and c) Interpolation of bin averages and 5-95%nitel measures.

The calculations take at least a two-step procedure:

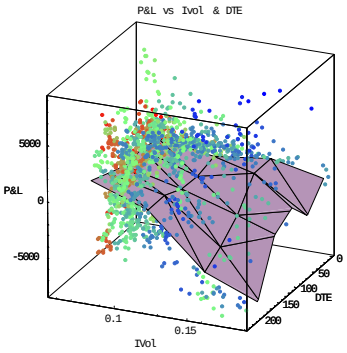
Calculate "bin" values, for example the average in "bin *i*":

$$u(x_i, y_i) = \sum_{k=1}^{N_i} z_{i,k} \qquad (10.95)$$

Then, interpolate the bin values $u(x_i, y_i)$ by any suitable means, such as tiling.

Additionally, Figure 10.18 – 2 c) shows that it is relatively easy to apply statistical measure directly. In this case, the image is showing the 5%ntile and 95%ntile bounding surfaces, and provides easy assessment on the "credibility of the fit". For example, the range or spread of the results is much lower (better) in the lower right corner, than it is in the lower left corner, where the data appear to be in a very wide and "unpredictable" regime.

Other GoF measures are also possible, some of which are illustrated in the following Sections; particularly relevant is the Case Study 10.19.4 on Options Arbitrage.



Notice that the implementation of this type of methodology is very simple and straight forward, since the calculations are simple and straightforward, and (some of the) analysis relies on "greyware" rather than "software".