ON THE NATURE OF THE STOCK MARKET: SIMULATIONS AND EXPERIMENTS

by

Hendrik J. Blok

B.Sc., University of British Columbia, 1993
M.Sc., University of British Columbia, 1995

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES

(Department of Physics and Astronomy)

We accept this dissertation as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

November 2000

Appendix A

Discounted least-squares curve fitting

In this appendix the standard method of least-squares curve fitting is modified in order to make it more amenable to time series. In particular the goal is to use time series data for forecasting by extrapolating from historical data. As will be shown this method can require fewer computations and less storage. Also, by discounting historical data extrapolated forecasts become more robust to outliers.

The reader should keep in mind that, despite the similarity of notation with standard least-squares curve fitting, the following is specifically meant to be applied to time series, where the relevance of past data are *discounted* as newer data arrive.

This appendix borrows heavily from Press et al.'s excellent discussion of generalized least-squares curve fitting [20, Sect. 15.4] which is highly recommended.

A.1 Least-squares curve fitting

We use the index i to label our data points where $i = 0$ indicates the most recently acquired datum and $i = 1, 2, 3, \dots$ indicate successively older data. Each point consists of a triplet (x, y, σ) where x is the independent variable (eg. time), y is the dependent variable, and σ is the associated measurement error in y.

We wish to fit data to a model which is a linear combination of *any* M specified functions of x. The general form of this kind of model is

$$y(x) = \sum_{j=1}^{M} a_j X_j(x) \tag{A.1}$$

where $X_1(x), \dots, X_M(x)$ are arbitrary fixed functions of x, called the *basis functions*. For example, a polynomial of degree $M - 1$ could be represented by $X_j(x) = x^{j-1}$.

(Note that the functions $X_j(x)$ can be wildly nonlinear functions of x. In this discussion "linear" refers only to the model's dependence on its *parameters* a_j.)

A merit function is defined

$$\chi^2 = \sum_{i=0}^{N} \left[\frac{y_i - \sum_j a_j X_j(x_i)}{\sigma_i} \right]^2 . \tag{A.2}$$

which sums the (scaled) squared deviations from the curve of all N points. The goal is to minimize χ^2.

The derivative of χ^2 with respect to all M parameters a_j will be zero at the minimum

$$0 = \sum_i \frac{1}{\sigma_i^2} \left[y_i - \sum_j a_j X_j(x_i) \right] X_k(x_i), \; k = 1, \ldots, M \tag{A.3}$$

giving the best parameters a_j.

If we define the components of an $M \times M$ matrix $[\alpha]$ by

$$\alpha_{kj} = \sum_i \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2} \tag{A.4}$$

and a vector $[\beta]$ of length M by

$$\beta_k = \sum_i \frac{y_i X_k(x_i)}{\sigma_i^2} \tag{A.5}$$

then Eq. A.3 can be written as the single matrix equation

$$[\alpha] \cdot \mathbf{a} = [\beta] \tag{A.6}$$

where $\mathbf{a}$ is the vector form of the parameters a_j.

Eqs. A.3 and A.6 are known as the *normal equations* of the least-squares problem and can be solved for the vector parameters $\mathbf{a}$ by *singular value decomposition* (SVD) which, although slower than other methods, is more robust and is not susceptible to round-off errors [20, Ch. 2].

A.2 Discounting

The discussion above applies to all linear least-squares curve fitting. The variation proposed here is to discount the relevance of historical data as new data arrive. This was motivated by time series where the fitting parameters may vary slowly.

Fitting time series is typically handled with a moving window over the last N data points. Each of the last N points is weighted equally and all prior data is

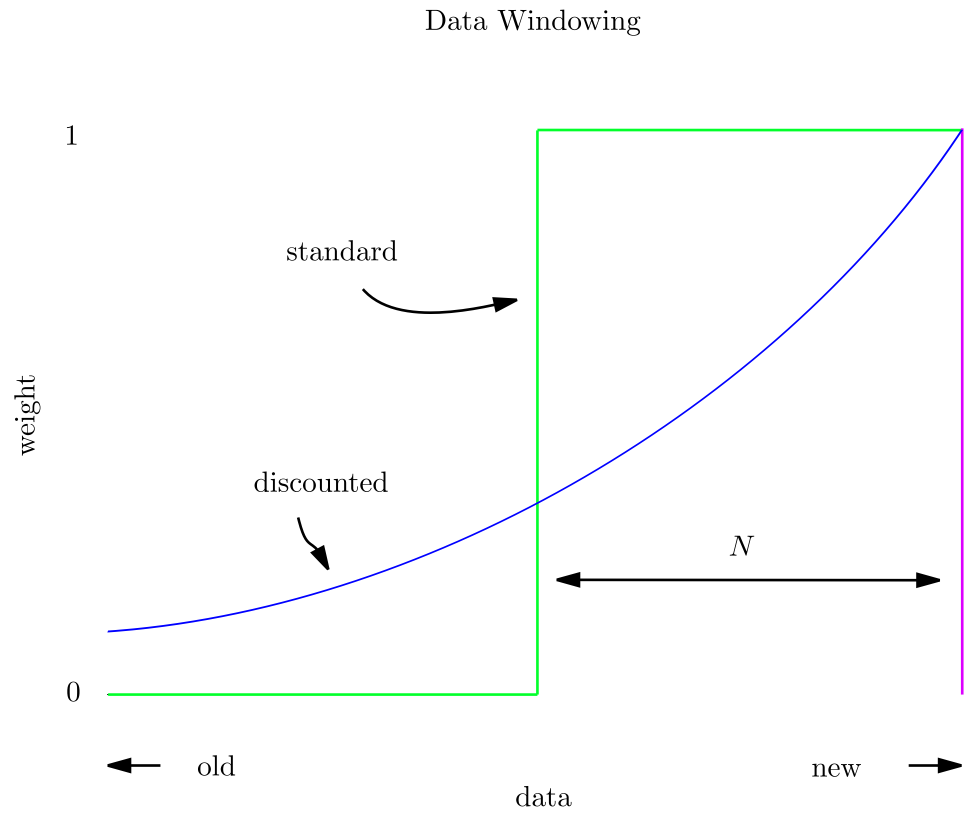

Figure A.1: Comparison of weightings using standard and discounted windows.

discarded as shown in Fig. A.1. The discontinuous weighting function can introduce discontinuities in the fitting parameters a_j as the data is updated, particularly when an *outlier* (a strongly atypical y-value) is suddenly discarded.

These discontinuities can be avoided by steadily discounting old data as new data arrive. As will be shown, this method also has computational and resource advantages.

As before, we use the index i to label our data points with larger i indicating older data. As a new datum arrives (x_0, y_0, σ_0) we shift the indices of prior data and scale up the errors by some factor $0 < \gamma < 1$

$$(x_{i+1}, y_{i+1}, \sigma_{i+1}) \leftarrow (x_i, y_i, \sigma_i/\gamma). \tag{A.7}$$

If we define σ_i^* as the original value of σ_i then after applying i of the above operations

$$\sigma_i = \sigma_i^*/\gamma^i \tag{A.8}$$

so, since $\gamma < 1$, the historical deviations grow exponentially as new information is acquired. Increasing the error effectively decreases the weight of a datum in the fitting procedure.

Calculation of the covariance matrix and the uncertainties of the parameters proceeds as with standard least-squares fitting (see [20, Ch. 15], for instance) so I will just mention the main result, namely that the inverse of $[\alpha]$

$$\mathbf{C} = [\alpha]^{-1} \tag{A.9}$$

gives the covariances of the fitting parameters

$$\mathrm{Cov}\,[a_j, a_k] = C_{jk} \tag{A.10}$$

and the variance of a single parameter is, of course,

$$\mathrm{Var}\,[a_j] = C_{jj}. \tag{A.11}$$

A.3 Storage and updating

So far we have made no mention of N, the number of data points to be fit. From Fig. A.1 it appears we need to store the entire history to apply this technique. But notice that as we acquire a new datum (x_0, y_0, σ_0), from Eqs. A.4 and A.5, the matrix $[\alpha]$ and vector $[\beta]$ update as

$$\alpha_{kj} \leftarrow \frac{X_j(x_0)X_k(x_0)}{\sigma_0^2} + \gamma^2 \alpha_{kj} \tag{A.12}$$

and

$$\beta_j \leftarrow \frac{X_j(x_0)y_0}{\sigma_0^2} + \gamma^2 \beta_j \tag{A.13}$$

so it appears we need not store any data points, but should just store $[\alpha]$ and $[\beta]$ and update them as new data are accumulated.

A useful measure we have neglected to calculate so far is χ^2, the chi-square statistic itself. In (partial) matrix notation Eq. A.2 can be written

$$\begin{aligned}
\chi^2 &= \sum_i \frac{y_i^2}{\sigma_i^2} + \mathbf{a}^T \cdot [\alpha] \cdot \mathbf{a} - \mathbf{a}^T \cdot [\beta] - [\beta]^T \cdot \mathbf{a} && \text{(A.14)} \\
&= \sum_i \frac{y_i^2}{\sigma_i^2} + \mathbf{a}^T \cdot ([\alpha] \cdot \mathbf{a} - [\beta]) - [\beta]^T \cdot \mathbf{a} && \text{(A.15)} \\
&= \sum_i \frac{y_i^2}{\sigma_i^2} - [\beta]^T \cdot \mathbf{a} && \text{(A.16)} \\
& && \text{(A.17)}
\end{aligned}$$

which appears to still depend on the data history in the first term. Let us define this term as a new variable δ,

$$\delta \equiv \sum_i \frac{y_i^2}{\sigma_i^2}. \tag{A.18}$$

Then, similarly to Eqs. A.12 and A.13, δ can be updated as more information is accumulated

$$\delta \leftarrow \frac{y_0^2}{\sigma_0^2} + \gamma^2 \delta \tag{A.19}$$

without requiring the entire data history.

Finally, it may be useful to record the number of points accumulated. But because each point loses relevance as it gets "older" we should likewise discount this measure, giving an effective memory

$$N^* \leftarrow 1 + \gamma^2 N^* \tag{A.20}$$

(not to be confused with the number of parameters M.)

So, to store all relevant historical information we need only remember $[\alpha]$, $[\beta]$, δ, and N^* for a total of M^2+M+2 numbers, regardless of how many data points have been acquired. Fig. A.2 shows that for many practical problems discounted least-squares fitting requires less storage than the standard moving window. Although it has not been tested, I expect a similar condition to hold for processing time.

As the reader can justify, all of these values should be initialized (prior to any data) with null values: $[\alpha] = \mathbf{0}$, $[\beta] = \mathbf{0}$, $\delta = 0$, and $N^* = 0$.

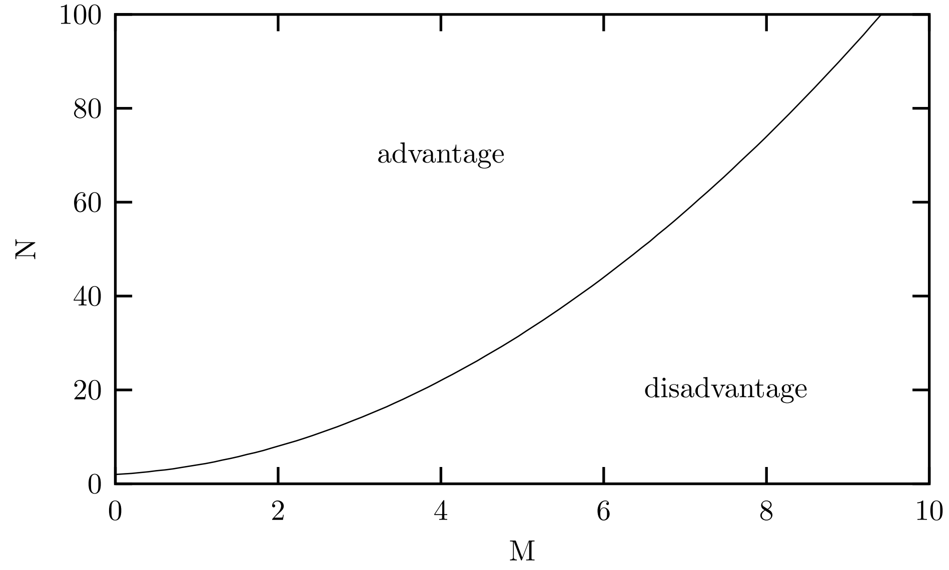

Figure A.2: Discounted least-squares fitting has a computational storage advantage over moving windows of N data points when $N > M^2 + M + 2$ where M is the number of parameters to be fitted.

A.4 Memory

For traditional least-squares fitting it is well known that if the measurement errors of y_i are distributed normally then the method is a *maximum likelihood estimation* and the expectation value of Eq. A.2 evaluates to

$$\left\langle \chi^2 \right\rangle = N - M \tag{A.21}$$

because each term $(y_i - y(x_i))/\sigma_i$ should be distributed normally with mean zero and variance one and there are $N - M$ degrees of freedom to sum the variances over.

Similarly with discounting, assuming $(y_i - y(x_i))/\sigma_i^*$ has variance one (notice this is the unscaled error),

$$\begin{aligned} \left\langle \chi^2 \right\rangle &= \sum_{i=0}^{N} \gamma^{2i} \left\langle \left[\frac{y_i - y(x_i)}{\sigma_i^*} \right]^2 \right\rangle - M && \text{(A.22)} \\ &= \sum_i \gamma^{2i} - M && \text{(A.23)} \\ &= N^* - M && \text{(A.24)} \end{aligned}$$

from Eq. A.20.

Notice that as the amount of data collected grows

$$N^*_{max} \equiv \lim_{N\to\infty} N^* = \frac{1}{1-\gamma^2} \tag{A.25}$$

which relates the discounting factor γ to the effective memory N^*. Conversely, it is more natural to set γ such that it produces the desired memory via

$$\gamma(N^*_{max}) = \sqrt{1 - \frac{1}{N^*_{max}}}. \tag{A.26}$$

A.5 Unknown measurement errors

On occasion measurement uncertainties are unknown and least-squares fitting can be used to recover an estimate of these uncertainties. Be forewarned that this technique assumes normally distributed (around the curve) y data with identical variances. If this is not the case, the results become meaningless. It also precludes the use of a "goodness-of-fit" estimator (such as the incomplete gamma function, see [20, Sect. 6.2] because it *assumes* a good fit.

We begin by assuming $\sigma^*_i = 1$ for all data points and proceeding with our calculations of $\mathbf{a}$ and χ^2. If all (unknown) variances are equal $\sigma^* \equiv \sigma^*_i$ then Eq. A.24 actually becomes

$$\left\langle \chi^2 \right\rangle = (N^* - M)\sigma^{*\,2} \tag{A.27}$$

so the actual data variance is best estimated by

$$\sigma^{*\,2} = \frac{\chi^2}{N^* - M}. \tag{A.28}$$

We can update our parameter error estimates by recognizing that, from Eqs. A.4 and A.9, the covariance matrix is proportional to the variance in the data, so

$$C_{jk} \leftarrow \sigma^{*\,2} C_{jk}. \tag{A.29}$$

A.6 Forecasting

Forecasting via curve fitting is a dangerous proposition because it requires extrapolating into a region beyond the scope of the data, where different rules may apply and, hence, different parameter values. Nevertheless, it is often used simply for its convenience. We assume the latest parameter estimations apply at the forecasted point x and simply use Eq. A.1 to predict

$$y_f = y(x) = \sum_j a_j X_j(x). \tag{A.30}$$

The uncertainty in the prediction can be estimated from the covariance matrix. Recall, the definition of variance is

$$\mathrm{Var}\left[z\right] \equiv \left\langle \left(z - \langle z \rangle\right)^2 \right\rangle \tag{A.31}$$

and the covariance between two variables is defined as

$$\mathrm{Cov}\left[z_1, z_2\right] \equiv \left\langle \left(z_1 - \langle z_1 \rangle\right)\left(z_2 - \langle z_2 \rangle\right) \right\rangle \tag{A.32}$$

so Eq. A.1 has variance

$$\begin{aligned}
\mathrm{Var}\left[y(x)\right] &= \mathrm{Var}\left[\sum_j a_j X_j(x)\right] && \text{(A.33)}\\
&= \left\langle \left(\sum_j \left(a_j - \langle a_j \rangle\right) X_j(x)\right)^2 \right\rangle && \text{(A.34)}\\
&= \sum_{jk} X_j(x) \left\langle \left(a_j - \langle a_j \rangle\right)\left(a_k - \langle a_k \rangle\right) \right\rangle X_k(x) && \text{(A.35)}\\
&= \sum_{jk} X_j(x) C_{jk} X_k(x) && \text{(A.36)}
\end{aligned}$$

where **C** is the covariance matrix with possible updating, in the absence of measurement errors, according to Eq. A.29.

The above gives the uncertainty in $y(x)$ but in the derivation it was assumed that the observed y-values were distributed normally around the curve where $y(x)$ represents the mean of the distribution. Similarly for the prediction, $y(x)$ is the prediction of the mean with its own uncertainty—on top of which there is the measurement uncertainty of data around the mean σ_{meas}. These two uncertainties are mutually independent so the variances of the two simply add to give the cumulative variance of the prediction

$$\begin{aligned}
\mathrm{Var}\left[y_f\right] &= \mathrm{Var}\left[y(x)\right] + \sigma^2_{\mathrm{meas}} && \text{(A.37)}\\
&= \sum_{jk} X_j(x) C_{jk} X_k(x) + \sigma^2_{\mathrm{meas}}. && \text{(A.38)}
\end{aligned}$$

A.6.1 Unknown measurement errors

If the measurement errors are not known in advance, but are calculated from Eq. A.28 then the above formula should be rewritten

$$\mathrm{Var}\left[y_f\right] = \sigma^{*\,2} \left(\sum_{jk} X_j(x) C_{jk} X_k(x) + 1\right) \tag{A.39}$$

where C_{jk} in this equation, are the covariances *without* rescaling.

A.7 Summary

Discounted least-squares curve fitting differs from the traditional linear least-squares method in that the uncertainties of older data are artificially amplified as new data are acquired, effectively discounting the relevance of older data. Discounting provides a very efficient method of storing the entire data series in only $M^2 + M + 2$ values, where M is the number of parameters to be fit, regardless of the length of the series. Discounting also smooths the fit, reducing the effects of outliers.

It has been demonstrated how discounted least-squares can be used for forecasting. Whether it is valid depends very much on the time series in question, and its consistency. If the fitting parameters vary on time scales of the same order or smaller than the memory N^* of the fit then the forecasts will not be reliable. (Of course, a suitable model of the time series is necessary as well.)

I have found no evidence of discounting being applied to curve fitting before; the only similar procedure I have found is “exponential smoothing”, a technique which uses damping coefficients to smooth forecasts. However, being such a simple premise I am confident this technique has already been discovered, I just don’t know where to look.