Title
Extracting Information from Mega-Panels and High-Frequency Data

Author
Granger, Clive W.J.

Publication Date
1998

98-01

UNIVERSITY OF CALIFORNIA, SAN DIEGO

DEPARTMENT OF ECONOMICS

EXTRACTING INFORMATION FROM MEGA-PANELS AND HIGH-FREQUENCY DATA

BY

CLIVE W.J. GRANGER

DISCUSSION PAPER 98-01
JANUARY 1998

Extracting Information From Mega-Panels and High-Frequency Data[1]

Clive W.J. Granger
Department of Economics
University of California, San Diego

January 1998

Abstract

Very large data sets in economics are already available and will soon become commonplace. The econometric techniques currently in use may not be relevant and new techniques will have to be devised. It can be argued that most tests of significance, linear models, assumptions of normality, and procedures to reduce bias, for example, will be replaced. The usefulness of asymptotic theory is discussed. It is suggested that methods for extracting conditional distributions will be becomes especially useful and a few particular possible techniques are suggested.

---

[1] I would like to thank Richard Carlson and also participants at a workshop at the Federal Reserve Bank, Washington, DC, and at the CIPES conference on large data sets in Rotterdam for their useful comments. Work conducted with the help of NSF Grant SBR-9708615

1. Introduction.

Economic statisticians are in the midst of a regime shift from a period of scarce data in most situations to one in which some data sets are comparatively enormous. It is helpful to establish some basic notation. *K* will indicate "thousand" as usual, with 4*K* meaning four thousand, and similarly *M* means "million" and *B* means "billion", so that 2*M* times 3*K* equals 6*B*.

If we have a panel in which *Q* variables are measured for each of *N* regions (or agents) at each of *T* time periods, this gives a data set $\underline{X}_{jt}$ $j = 1, ..., N;\ t = 1, ..., T$ where $\underline{X}$ is a vector with *Q* components. A mega-panel may well have *Q* about 100 or so, *N* in the thousands and *T* in the millions, producing a total set of size $A10^{11}$, for some *A*. Although it will be important to return to different types of mega-panels, with one or more of *Q*, *N*, and *T* being large, some of the early discussion will consider just a generic "large sample" of size *n*. To get an impression of how large *n* may be in practice, think of all of the super-market items recorded using bar codes - prices, quantities, brands, commodities, locations, what else bought - in a major developed country in a week or the details of transactions using a major credit card company in a day throughout Europe, or the prices of every transaction for every share and bond on a major exchange over a ten year period. Substantial subsets of such data sets are already available.

When discussing the analysis of large data sets I will assume that the form of the analysis used is not constrained by any computer limitations, such as speed of calculation, memory size, or cost. Although the absence of such limitations will never be strictly true they are

progressively becoming less important for most sciences. Computer speed has been growing about 20% annually and data storage abilities about 40% annually and further similar growth is expected. What matters for our purposes is the number of calculations per second. When I started my career 40 years ago a young woman working on an electric desk calculator could, at the very most, achieve one calculation per ten seconds, whereas recently an "ultra computer" consisting of 7*K* PC microprocessors working in parallel achieved a speed of one trillion ($10^{12} = M^2$) calculations per second (*Science News*, January 4, 1997, page 7). (I hope I will be forgiven for noting how little the contents of elementary textbooks of statistics have changed over these same forty years.) Wegman (1995) has discussed the interaction of huge data sets and the frontiers of computational feasibility, using and extending a taxonomy of data set sizes due to Huber (1994). It is pointed out that many of the algorithms used in statistical analysis are $O(n)$ in complexity, such as those calculating moments and kernal density estimates, some are $O(n \log n)$, such as those calculating fast Fourier transforms, and some may be $O(n^2)$, such as some clustering algorithms. Thus, my assumption that there will be no computing constraints may not always be a sensible one, but it is more interesting to continue discussion with it operating and then to face practical issues later.

If one moves from the classical situation of having small or data sets of limited size to having substantial amounts of data to analyze, there are three basic questions:

(i) which of our standard procedures and concepts should be discarded?

(ii) will some of our standard procedures evolve and then perform better with n large?

(iii) what new techniques need to be developed?

The answers to these questions will partly depend on the objective of the analysis.

For the following discussion it is often helpful to have a specific example in mind. For this purpose, consider the records of purchases charged to a popular credit card in a large country, providing about one million (so $n = 1M$) pieces of data per day. I should state that I have no connection with any such company and thus no inside information about the actual properties of their data.

My general attitude will be that large data sets are not a problem, they are an opportunity to do superior analysis, if only we could think of how to do it! Naturally, I do not expect to solve all possible questions in this paper.

2. Concepts And Procedures to Discard or Improve

This section considers some familiar concepts and techniques that are candidates to be discarded when n is large.

2(i) Small-sample Adjustments.

There will be no need to worry about whether to use $1/n$ or $1/(n-1)$ in the sample estimate of variance, or use R_C^2 rather than R^2.

2(ii) Jackknife.

Or any other $O(1/n)$ bias adjustment technique.

2(iii) t-tests, F-tests, chi-squared tests.

Any test which has a degree or degrees of freedom that depend on n will now take simple asymptotic forms, with t- and chi-squared distributions going to normal and F to unity. It will be much easier to teach linear regression - should one wish to continue to do so - without worrying about degrees of freedom!

2(iv) The Higher Moments.

Skewness and Kurtosis have never been much use and it is doubtful if we will be interested in fitting Pearson curves. The mean might survive for old times sake but the variance may not, as it is too strongly related to the normal distribution (which is discussed below).

2(v) 95%, 99% Confidence Intervals.

In simple cases, where confidence intervals are $O(n^{-1})$ they will be effectively zero so that virtually any parsimonious parametric model will produce a very low *p*-value and will be strongly rejected by any standard hypothesis test using the usual confidence intervals. Virtually all specific null hypotheses will be rejected using present standards. It will probably be necessary to replace the concept of statistical significance with some measure of economic significance.

Consider the mythical credit car company which observes $n = 1M$ transactions per day and suppose that there is an interest in the proportion of purchases above a certain monetary amount (say \$200 or 400 guilders). If the observed proportion is $\hat{p} = 0.4152$, say, then the 99.9% confidence interval is $\hat{p} \pm 3.30\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, using the binomial being approximated by normality and the *Biometrika Tables for Statisticians*, vol. 1, 1966, which gives the band 0.4152 ± .001. It is hard to believe that a move of one digit in the third decimal place has any economic significance for a single day although a different interpretation will be given later. However a particular null hypothesis $H_o : p = 0.4$, will be clearly rejected, for example, or even $H_o : p = 0.418$.

2(vi) Model Selection Criteria, BIC, AIC, etc.

Parsimony may or may not be a desirable property for a model but it is not required when data is plentiful. The terms in the criteria that depend on *n* will dominate its size and so

searching over the number of parameters to use will be ineffective. A subsequent consideration is that the whole concern with data mining - that is considering models until one is found that fits the data particularly well - disappears with very large data sets unless extremely complicated models are also considered. Thus if m, the number of parameters in a model is $o(n)$, data mining should not be a problem and unrealistic models will not be found.

2(vii) Bootstrap.

The bootstrap extends the sample size in a somewhat artificial way. If n is large enough there is no need for such an extension. Many of the questions being tackled by the use of the bootstrap do not arise with n large. This will be true of many other simulation techniques. There will be little use for artificial data when real data is plentiful.

2(viii) Bayesian Estimation.

The use of the Bayesian prior in a likelihood estimation procedure is essentially an extension of the data set if the prior is correct, in some sense. As n becomes large the data generated component of the likelihood will eventually dominate the prior and the estimates achieved will not be affected by the choice of prior. Of course, other aspects of the Bayesian approach to the analysis and particularly the interpretation of results will remain as relevant as now.

2(ix) The Assumption of Linear Regression.

The basic model of interest, according to virtually all but the most advanced statistics and econometric texts, is a regression that is linear in the explanatory variables. Of course, the statistical theory of linear regression is very well understood by now but this is not a good enough reason to use linearity as a null hypothesis. In many fields there seems to be a strong

belief that relationships between variables could well be non-linear. If these beliefs are correct it may not be appropriate to use linearity as a null, but rather start with some non-linear specification. It is easy to assume linearity and it will be difficult to break away from thinking in terms of such models as the base case from which one tests and generalizes.

2(x) The Assumption of Normality.

Virtually everything that was just said about linearity can also be said about normality. It is a familiar, frequently and all-to-easily made assumption that will almost certainly not be accepted in reality as a null for the population distribution. This remark should not be interpreted as saying that normality will not arise from asymptotic considerations, that possibility is considered below. Of course, not having linear relationships removes the possibility of the multivariate normal distribution occurring.

2(xi) Pooling in Panel Analysis.

In many panels analyzed in the past it has often been the case that the time series dimension *T* has been small whilst the cross-section dimension *N* is large. It is thus impossible to specify and estimate an individual temporal or dynamic relationship for each *N*. This impasse is alleviated by assuming that all regions or agents have the same temporal properties, so that the lagged dependent variable enters with the same coefficient in each region or each region has a unit root or causality between an independent variable and the dependent variables occurs either in all regions or nowhere, for example. Simplifying assumptions of this form fall within the "pooling" idea where a panel data set is considered to be a cross-section sample but will not be required for mega-panels as *T* will no longer be small. Naturally, if *T* remains fairly small this remark does not apply.

2(xii) Measures of Relative Efficiency of Estimators.

If $\hat{m}_1, \hat{m}_2$ are both consistent estimates of a parameter *m*, then attention should be aimed at the one that is easiest to calculate and/or interpret as relative efficiency is of little importance. For example, if the distribution is $N(m, \sigma^2)$ Gaussian then the variance of sample mean is σ^2/n and of the sample median is $\frac{\pi}{2} \cdot \sigma^2/n$. If $n = 1M$ both variances are so small, multiplying by $\pi/2$ is of little practical relevance.

3. Asymptotics.

For large *n* the natural reaction of a statistician will almost certainly be that various parts of asymptotic theory will become immediately relevant. One should expect that the law of large numbers, the law of the iterated logarithm and various limit theorems, particularly central limit theorems, should be applicable, for example. Most of these theorems, in their many forms, start by considering sequences of i.i.d. variables. A good survey of this area is Sen and Singer (1993), although it rarely mentions non-i.i.d. cases. Many of the results discuss rates of convergence of specific random variables to constants, bounds or distributions under particular sets of assumptions. One would certainly expect, for *n* in the 100*K* or millions, to be allowed to use the asymptotic critical values of a test statistic, for example, although what probability levels would be appropriate is unclear. Is there any point in testing if an asymptotic theorem is true with one's data? If the assumptions are correct, the theory certainly is correct and so the test is of the assumptions and not of the theory, but that itself may be interesting.

Thinking about the example of the credit card company, at first it seems to be highly likely that a transaction charged by one customer on a day will be independent from a transaction by another customer on the same day. One can argue that my making a transaction does not

affect your decisions. Aggregating to find the average size of transaction or the proportion greater than some given monetary amount will obey the assumptions for which the asymptotic theory can be applied. Thus, if one thinks of a typical transaction having size being taken from some distribution with mean *m* and variance σ^2, the daily sample mean transaction size can be expected to have a normal distribution with mean *m* and variance σ^2/n, when there are *n* transactions in the day. If one day's sample of transactions is good, then surely several days are better, so one could form a sample using two days, a week, or more. For the amalgamated two-day sample with *n* transactions per day, the sample average should be normally distributed with mean *m* and variance $\sigma^2/2n$. This assumes that the parameters of the distributions have not changed. However, it is quite possible in subtle ways that the distribution does change, according to weekly, monthly, seasonally, plus business-cycle, and other economic effects. Thus, the mean on a Wednesday in early April in a business-cycle down turn may be quite different from the mean on a Friday at the end of November at a cycle peak. Similarly for the variance. Putting samples together from different days will be mixing distributions and information is obscured. One can consider transaction sizes on day *t* coming from a distribution $D(x, m_t, \sigma_t, \theta_t)$ where θ_t represents parameters other than the mean and standard deviation. The mixing distribution idea takes the parameters to be drawn at random from their own distribution, but in a time-series context, as here, they may change systematically with time, the seasons, and with causal and explanatory variables. This is discussed further in Section 7. If one is interested in the proportion of transactions whose size is less than some specified amount *k*, then the expected value of this proportion will be a function of the parameters of *D* and will be very well estimated by the observed proportion on a given day, provided that this is not in a tail.

The problem of tail estimation is considered further in Section 6.

In this discussion it has been implicitly assumed that the parameters of the distribution do not change *during* the day but this may not be correct. Further the customers may be categorized in many ways; there may be three distinct types of card (gold, company, regular) and divisions by gender (M/F), age (3 categories), income (4), regions (5), race (3), say, giving just over 1*K* categories. Thus, with a daily sample of 1*M* the average number in each category will be 1*K* and some categories will contain much smaller samples. The large sample case is easily lost be sub-division, unless assumptions of independence across categories are made, in the usual contingency table fashion. The possibility that more data does not produce more information is discussed in Section 5.

4. Summary Statistics. Density Estimation.

In this section discussion will continue of the large *n* case, essentially a large cross-section such as the transaction records for the credit card company on a single day, although soon more complicated and realistic cases will have to be considered. One of the tasks that statisticians face is how to summarize complicated data sets. The idea is to try to approximate a large number of values by just a few and also, possibly, to display the data in such a way that some clear-cut simplification becomes visually obvious. In this area there are obviously producers and consumers, the statisticians are the producers of the statistical devices and the consumers are the potential users. It is usually the case that the producers have little idea what the consumers actually want, and so provide what they hope will be satisfactory, possibly after some education.

It is clear that if the statistician, as producer, provides a good estimate of the distribution

function, or density function of the variable or variables of interests, then the consumer can easily deduce any particular summary statistics of interest, such as mean, standard variance, median, or quantiles, from them. The same has been shown to be true in the area of time series forecasting, if the producer provides a predictive (conditional) density function for x_{n+1} given an information set I_n then a consumer can derive from it the optimum point forecast for particular cost function, plus corresponding confidence intervals, see Granger and Pesaran (1996), Granger (1997a).

Methods of estimating density functions are well developed using kernel smoothers for the univariate case and are starting to be developed in the multivariate case, see for example Jones, Marron, and Streather (1996) and the references given there. Alternatively one can estimate a set of quantiles, for example Sin and Granger (1996). The special problems arising when attempting to estimate the tails of the distribution will be discussed in Section 6.

The question of how to visualize a large data set is an important area of research which is getting a lot of attention from computer specialists. There is a section of the *IEEE* specializing in this area and which produces annual reports. Papers in the *Journal of Computational and Graphical Statistics* consider specific approaches. However it is unclear if it is worthwhile presenting visualizations of multi-dimensional data sets in which the noise level is high, as might typically be expected in economics. One can only think of the "patterns" that some people believe they see in stock market prices that provide the forecasts which comprise "technological analysis."

For the case N large, Q not large, $T = 1$, one possibly useful way to summarize the data is by using multivariate quantiles. A useful account of how these are defined, estimated and a

listing of some useful properties are given by Chaudhuri (1996). A brief summary is given in Appendix A.

For large N, estimators of the quantiles will be consistent and normally distributed. These quantiles could also be made functions of explanatory variables and non-linearly so by using neural network type representations.

A major research question is how to summarize data in the case when Q, the number of variables is fairly large, even 20 say, T is very large and for any size of N. Viewing a Q-dimensional joint distribution in any useful form is difficult for Q greater than 3. Possibly looking at a variety of conditional distributions for univariate and bivariate distributions and then rotating them for better viewing, will be a useful start.

5. More Data But Not More Information.

Asymptotic theory is based on the idea that as the amount of data increases so does the amount of information. This section is just making the simple observation that there are many situations where things do not workout that simply, so that asymptotics are not always relevant. A naive example is of a series that is generated by a single cosine wave without any background noise. If you know a few terms the whole process is known perfectly, observing more terms tells us nothing extra, except perhaps that the generating mechanism has not changed. Other examples involve rare large events, such as exceptionally large changes in the Dow Jones Stock Market Index, we do not find more October 1987's by observing the index more frequently, or particularly large corporations by moving from those listed on the New York Stock Exchange to those on the National Association of Security Dealers Automated Quotations (NASDAQ) system. By observing aggregate consumption more frequently we do not find more business

cycle turning points, by observing earth movements more carefully we do not observe more large earthquakes, but these examples depend on the definition used of a turning point or of a large quake.

A serious question occurs in the time series context, which is whether one can learn anything more about low frequency (long-run, low period) components of a series by observing the series more frequently? Some aspects of this question will be considered in Section 7.

6. Tails, Extremes, and Outliers.

For normal size samples the most difficult part of the distribution to estimate is the details of properties in the tails, simply because there is typically little data there. At first sight, large *n* should solve that problem. As pointed out in Sen and Singer (1993), for example, if

$$F_n(x) = n^{-1} \text{ (number of sample } \leq x \text{ in magnitude)}$$

is the empirical distribution function for an i.i.d. sample, drawn from a population with distribution $F(x)$, then

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{D} N(0, F(x)(1 - F(x))$$

for each fixed x, using the Central Limit Theorem. Thus, if $n \equiv 1M\, F_n(x) - F(x)$ will have a standard deviation of not more than 1/2000. The comulative histogram, even without smoothing, should give an estimate of the distribution function that will be satisfactory for economic decision purposes over a wide range.

Of course, if one wants to go to the real extreme, values which are the minima or maxima of the sample or the order statistics $X_{n,k}$, where this is the k^{th} largest term is a sample of size n, X_t, $i = t, \ldots, n$, in the case where k/n is near one or near zero, then a different asymptotic theory holds. As is well known, these extremes, after a suitable normalization, will come from a group

of just three possible distributions, as proved by Gumbel (1958); see Sen and Singer (1993, chapter 6) or David (1981, chapter 9).

The theory of extreme values loses relevance when the data is contaminated by outliers of certain kinds, particularly mis-recordings, reading and entry errors. One can expect that some percentage, say λ %, of a data set will involve such errors and that this percent will remain, approximately, the same regardless of the size of *n*. Thus, the number of errors will increase proportionately with *n*, at λn. The "discovery"of outliers, using parametric distribution functions will produce a similar outcome. Suppose that one arbitrarily excludes a certain percentage of the sample, outside particular empirical quantiles, and then fits a particular class of parametric probability distribution functions to the remaining data using likelihood techniques to estimate parameters, giving $f(x,\hat{\theta})$. Then outliers X_i can be defined as those data points so that $F(x_i,\hat{\theta})$ are outside some prescribed range, say 0.001 to 0.999. Again, the percentage of outliers "found" will be roughly constant and the number will be proportional to *n*. Of course, this number can be reduced if some outliers are so clearly different from the rest of the sample, their values can be selected and investigated and then possibly replaced with a true, corrected value. However, the costs of doing this may become immense for *n* large.

If the outliers are actually irrelevant figures, there is nothing lost by dropping them as in a regression technique using truncation. For large *n*, one does not need to worry about efficiency, so there is little point in optimizing the number of points excluded. Of course, when considering relationships between several variables, the concept of an outlier becomes more subtle. If there are *Q* variables and an exceptional value affects only one of them, it will disrupt an attempt to find a relationship, but if it affects several variables it will dominate the relationship and be an

influential point. These influential or leverage points are probably not outliers due to measurement errors but may occur because a common factor has a long-tailed distribution. One can think of there being two types of generating processes for the economy, one is the typically occurring process, the other is the exceptional process. For example, there can be the typical Los Angeles economy, which is observed virtually all of the time, but just occasionally there is a substantial earthquake, when that economy becomes quite different for a period. There will be many small earthquakes occurring during the typical period but they have little effect, but when an extreme earthquake occurs most variables in the LA economy will be affected, such as consumption, travel, employment, house construction, investments, and so forth. Removing such data points allows one to better model the typical economy. Increasing the amount of data does not necessarily increase the number of influential points, such as large earthquakes, so the ability to model the exceptional economy may not increase, unless one goes to a panel of economies that occasionally are influenced by quakes.

7. Time Series Data (*Q*, *N* not large, *T* large).

If one has a few long vector series a number of opportunities open up that have not historically been available but there are also a number of questions that arise that have not been considered enough in the past. To get a longer time series one has to either view the series over a longer time span, so that rather than having a monthly series over 1948 to 1997 ($T = 600$) one could observe it monthly from 1868 to 1997 ($T = 1560$) for example, or one could observed the same span but at a higher frequency rate, going from monthly to daily. So for 1948 to 1997 T will increase from 600 months to about 12,500 days (taking 250 working days per year). [Naturally, one could consider changing both span and frequency but I will not do so.] Both

methods of increasing *T* have associated difficulties. Increasing the time span involves mixing data of different quality, based on different definitions, and of possibly dubious relevance for current models and for some problems of the present-day or the immediate future. Observing data more frequently obviously does provide more data but it also does not necessarily provide more information. If the data span contains four swings of the business cycle it will continue to do so however often you observe it. If a series has a "typical spectral shape" one learns little more about the long-run (or long-period or low-frequency) components of the series by observing the data more frequently.

There is an expositional difficulty that one faces at this point as there may be a need to discuss the high frequency components of a series, which I will call the short-period components, and also refer to high-frequency data, which is data observed at short time intervals compared to low-frequency data. The problem is confused by the fact that what is a short-period component at one frequency (say monthly) may be a long-period component if the data is observed more often, such as daily. For linear models, the concept of aliasing is useful for untangling such questions, as discussed in Koopmans (1974, p 135).

Of all the topics that could be discussed here I will mention only three and discuss just two. The topic I will mention but not explore in detail is the fact that the spacing between observations of many very high frequency series in economics is not constant, appearing to be random, but that the spacing may themselves contain relevant information. Examples are the times of trades on a speculative market or at a store or supermarket. I do not discuss it because although it is certainly an interesting area it is already being considered, for example, by Engle and Russell (1996), and it is a property of just some high frequency data sets and is not

specifically a property of large data sets.

The two topics that will be discussed at least briefly are what type of models should be attempted with our long series and how to link together series observed at different frequencies.

At present time series analysis is in a rather confused situation. There are an enormous number of models available to specify the conditional mean, they can be linear or non-linear, involve a few or many lags, be univariate or multivariate, I(d) or not, where d can be an integer or a fraction, can contain seasonals or trends, can have seasonally, randomly or stochastically changing coefficients, be subjected to structural breaks, and so forth. Then the whole exercise can be repeated for the conditional variance. A better way to proceed will be to try to produce an estimate of the predictive or conditional distribution of $\underline{x}_{t+1}$ given $\underline{w}_t$ where $\underline{w}_t$ is a wide enough state-space variable, i.e. $p_t(x|\underline{w}_t)$. This is, of course, not at all a new idea and has been central to work by Aoki (1990), Kitagawa (1989), and an early development was an algorithm from Control Engineering by Sorensen and Alspach (1971). Details are shown in Appendix B.

Of course, once the conditional distribution is known one can obtain the conditional mean, conditional variance, conditional quantiles, and other parameters. It is also possible to iterate the procedure to give conditional distributions several steps ahead. The method essentially approximates a distribution by a mixture of Gaussians, uses a Kalman filter on each and then sums. It has the strengths of the Kalman filter, the ease of use and interpretation and the possibility of certain types of time-changing parameters such as season or breaks, but it also suffers from the same weaknesses. A process cannot be exploding or trending or have a unit root in variance so the distribution cannot be non-mixing, for example. Thus many forms of data found in finance and macro-economics, as well as demographics, development economics,

trades, etc. may not fit into the present framework. It seems that further development is required to cope with series that are cointegrated, in its present or generalized forms, as discussed in Granger (1997b), for example.

When observing a time series at a very short interval, that is with high frequency, it is difficult to relate the slow-moving components of the market or economy with rapidly changing parts of the variables. What one sees in data depends on the tools used to look at it - the models and techniques that are utilized. For example if stock prices are measured every minute (when the market is open) and they are analyzed using a simple Box-Jenkins approach one almost certainly gets an ARIMA (p, 1, q) model, with p, q possibly small but certainly with a unit root. In fact, the series may not actually contain such a root but it probably has a daily cycle - partly due to the lunch-time slump in volume - but within the class of models being considered such a "long" cycle, in daily data, can only be captured by having a unit root in the model. The slow movements of high-frequency data may not be due to changes in slow moving variables, but it is difficult to obtain this relationship in a regression, say. Rather than tackle the complete problem, consider a similar situation but of a lesser magnitude, that of including a variable measured monthly into a model of variables which are otherwise all observed weekly. For ease of consideration, suppose that a VAR type of model is being considered relating the vector of weekly series $\underline{W}_t$ and a vector of monthly series $\underline{M}_t$. Although $\underline{M}_t$ is observed monthly one could envisage a model that exists weekly, estimating a value for the process each week and then, with each week as a base, forecasting the value of M_{jt} for later weeks. As some of these forecasts can be evaluated, corresponding to forecasts of the monthly values, the model can be evaluated and thus estimated. Although the weekly values of the variables, which can depend on

lagged monthly, "weekly" values of M_{jt} and also of other variables including lagged $\underline{W}_t$, do not actually exist they can be thought as being of a virtual reality variable (VRV). The idea can obviously be extended to higher frequencies, as shown in the following table:

Input Series	GNP Form	Used to Model (say)
	Quarterly GNP ➜	GNP in another country
	↓	
Index of Indus. Prod. ➜	Monthly Virtual GNP ➜	Consumption
	↓	
Employment Data ➜	Weekly Virtual GNP ➜	Auto Sales
Money Supply/ Interest Rates ➜	Daily Virtual GNP ➜	Stock Prices

Few of the virtual GNP figures are actually observed, they are estimated from some model using previous virtual values and other data of similar frequency, but can be compared to the actual value when it is observed each quarter. The model will need to contain an "error-correction" term in case the predicted GNP figures deviates from the observed value, to bring the sequence back onto course.

As shown in the table, one could use virtual values as inputs in models for other high frequency variables. Some work has already been conducted in this area, for example, by Rahjens and Robins (1993) but further development is needed.

8. Megapanels (*M, T* both large, *Q* not very small).

The ultimate challenge for large data set analysis is when one measures a vector of *Q* variables, say 20 or so, for each of *N* agents (companies, stocks, municipalities, cities), say 2*K*, over *T* time periods, where *T* may be 10*K* or more. Even if each agent or district is analyzed

individually, producing *N Q*-dimensional predictive distributions will not only be computationally expensive, which is of no concern here, but will be extremely difficult to absorb and interpret by a potential user.

It will be important to be clear about the **purpose** of the analysis before attempting any analysis as this should allow one to concentrate on a subset $\tilde{Q}$ of the Q variables. For each such subset and region (or individual agent), previous reasoning suggests that we should attempt to find approximations to conditional distributions $P(\tilde{Q}_{t+k} \mid I_t)$. The dimensions of these subsets should not be large, for interpretational reasons and the techniques outlines about could be used **if** I_t includes all lags of all variables, the specification search even for linear models will be too extensive.

What is needed is some technique to allow an assumption that some of the data set can be treated as weakly exogenous and thus irrelevant for the specification and estimation of the final model. The criterion for inclusion of further data should not be a statistical one but rather be of economic relevance - are better decisions to be made if a more complicated model is used than a simpler one?

It would be a good idea if some of the data in the full information set could be usefully summarized, such as mean growth rates of all variables (like CAPM), mean growth rates of "near-by" variables (regional analysis) or interquartile ranges of these growth rates. Clearly a great deal of experimentation and learning still need to be done in what will be a very exciting area.

A pair of useful reference for background material are Advances in Knowledge Discovery and Data Mining, ed. by U.M. Fayyed, et al., AAAII and MIT Press, 1996 and Massive Data Sets

(Proceedings of a Workshop) National Research Council, National Academy Press, 1996.

References

Aoki, M. (1990): State Space Modeling of Time Series. Springer-Verlag, Berlin.

David, H.A. (1970): Order Statistics. J. Wiley & Sons, New York.

Chaudhari, P. (1996): "On A Geometric Notion of Quantiles for Multivariate Data," *Journal of the American Statistical Association* 91, 862-872.

Granger, C.W.J. (1997a): "Outline of Forecast Theory Using Generalized Cost Functions," unpublished paper.

Granger, C.W.J. (1997b): "Introducing Nonlinearity Into Cointegration," *Revista de Econometria* (Brazil) 16, 25-35.

Granger, C.W.J. and H. Pesaran (1996): "A Decision-Theoretic Approach to Forecast Evaluation," Department of Applied Economics, Cambridge University working paper 9618.

Gumbel, F.J. (1958): Statistics of Extremes. Columbia University Press, New York.

Huber, (1994): "Huge Data Sets" in Compstat 1994: Proceedings, eds. R. Dutter and W. Grocsmann, Heidelberg, Physica Verlag.

Kitagawa, G. (1989): "Non-Gaussian Seasonal Adjustments," *Computers Math. Applica.* 18, 503-14.

Koopmans, L.H. (1974): The Spectral Analysis of Time Series. Academic Press, New York.

Sen, P.K. and J.M. Singer (1993): Large Sample Methods in Statistics. Chapman & Hall, New York.

Sin. C.-Y. and C.W.J. Granger (1995): "Estimating and Forecasting Quantiles with Asymmetric Least Squares," unpublished.

Sorenson, H. W. And D.L Alspach (1971): "Recursive Bayesian Estimation Using Gaussian Sums," *Automatica* 7, 465-479.

Rathjens, P. and R.P. Robins (1993): "Forecasting Quarterly Data Using Monthly Information," *Journal of Forecasting* 12, 321-330.

Russell, J.R. and R.F. Engle (1996): "Econometric Analysis of Discrete Value, Irregularly Spaces Financial Transaction Data Using a New Autoregressive Conditional Multinomial Model." Working paper University of Chicago.

Wegman, E.J. (1995): "Huge Data Sets and The Frontiers of Computational Feasibility," *Journal of Computer and Graphical Statistics* 4, 281-295.

Appendix A
Chauduri's (1996) Multivariate Quantile. A Summary

For a pair of Q- dimensional vectors ($\underline{u}$, $\underline{s}$), define:

$$\varphi(\underline{u}, \underline{s}) = |\underline{s}| + (\underline{u}, \underline{s})$$

$$\text{where} \quad |\underline{s}| \equiv (\sum_j s_i^2)^{1/2} \qquad ,i = 1, \ldots, Q$$

$$\text{and} \quad (\underline{u}, \underline{s}) \equiv \sum_i u_i s_i .$$

Now, for data $\underline{X}_1, \ldots, \underline{X}_n$ all Q-dimensional and for a given vector $\underline{u}$ in the open unit ball, so that $|\underline{u}| < 1$. Define $\hat{\underline{q}}(\underline{u})$ to be the quantity that minimized

$$\sum_{j=1}^{n} \varphi(\underline{u}, \underline{X}_j - \underline{q}).$$

A value of with $\hat{\underline{q}}(\underline{u})$ $|\underline{u}|$ close to 1 corresponds to an extreme quantile, but if $|\underline{u}|$ is close to zero, one geta a central quantile. $\underline{u} = \underline{0}$ gives the "spatial median" $\hat{\underline{M}}_n$ which minimizes

$$\sum_{j=1}^{n} |\underline{X}_t - \underline{M}_n| .$$

Chaudhuri (1996) proves the following theorems (written in an informal fashion):

1. $\hat{\underline{q}}(\underline{u})$ Exists for any given $\underline{u}$ and is unique if $Q > 2$ and if the $\underline{X}$'s do not lie on a straight line in R^Q space.

2. The quantiles are found fairly easily using an algorithm of the "Newton-Raphson" type.

3. Concerning linear transformation

   (i) if $\underline{X}_n \rightarrow \underline{Y}_n = \underline{X}_n + \underline{a}$, a constant, then the quantiles of $\hat{\underline{q}}_n(\underline{u}) + \underline{a}$.

   (ii) if $\underline{Y}_n = c\,\underline{X}_n$, c a constant, then the new quantiles are $c\,\hat{\underline{q}}_n(\underline{u})$; and

   (iii) If $\underline{A}$ is an orthogonal matrix (so that $A = (A')^{-1}$ and $\underline{Y}_i = \underline{A}\,\underline{X}_i + \underline{a}$, a constant, then the quantiles of $\underline{Y}$ are

$$\hat{\underline{q}}_n(v) = \underline{A}\,\hat{\underline{q}}_n(u) + \underline{a}$$

where $\underline{v} = \underline{A}\,\underline{u}$.

Appendix B
Estimating Conditional Distributions Using Gaussian Sums
due to H. Sorenson and D. Alspach (1971)

Suppose that the state variable $\underline{X}_t$ evolves by

$$\underline{X}_{t+1} = f_t(\underline{X}_t, W_{t+1}) \tag{B1}$$

but that the state is observed imperfectly through out the measurement equation

$$\underline{Z}_t = \underline{h}_t(\underline{X}_t, \underline{V}_t) \tag{B2}$$

where $\underline{W}_t, \underline{V}_t$ are iid series, independent of each other, $\underline{X}_t$ is not observed but $\underline{Z}_t$ is observed. The basic equations for the system, in terms of probability density functions are:

$$P(\underline{X}_t \mid \underline{Z}_{t-j},\ j \geq 0) = \frac{P(\underline{X}_t \mid \underline{Z}_t - j,\ j \geq 1)\, P(\underline{Z}_t \mid \underline{X}_t)}{P(\underline{Z}_t \mid \underline{Z}_{t-j},\ j \geq 1)} \tag{B3}$$

the state up-dating equation, and

$$P(\underline{Z}_t \mid \underline{Z}_{t-j},\ j \geq 1) = \int P(\underline{X}_t \mid \underline{Z}_{t-j},\ j \geq 1)\, P(\underline{Z}_t \mid X_t)\, d\underline{X}_t \tag{B4}$$

the transition equation for $\underline{Z}_t$.

It is generally very difficult to integrate (B4) except in the linear|Gaussian case and so an approximation is used. The Guassian sum approximation takes the form

$$P_n(x) = \sum_{i=1}^{m} \alpha_i N \sigma_i (x - \mu_i) \tag{B5}$$

where $N_\lambda(x) = (2\pi\lambda^2)^{1/2} \exp(-x^2/2\lambda^2)$ with all $\alpha_i \geq 0, \sigma_i \geq 0$ and $\sum_{i=1}^{m} \alpha_i = 1$. By construction $P_n(\lambda) \geq 0$ and $\int_{-\infty}^{\infty} P_n(x)\theta_x = 1$.

In equations (B3) and (B4) all conditional distributions plus the start-up distributions are assumed to take the form (B5), but a multivariate form!

It seems that one can get good approximations to multivariate conditional distributions $P(\underline{X}_{t+h} \mid I_t)$ using these approximations.