Computing power and the power of econometrics

April 12, 2006

James D. Hamilton

Department of Economics, 0508

University of California, San Diego

La Jolla, CA 92093-0508

jhamilton@ucsd.edu

Prepared in celebration of the 50th anniversary of the Econometric Institute at Erasmus University

1 Introduction.

I bought my first computer in 1984, an IBM PC-AT. It was a marvelous machine, with a clock speed of 6 MHz, a hard disk that could hold 20 megabytes of data, and capable of something like 8,000 floating point operations in a single second.

And every single year since then, the computers available for anyone to buy have become even more marvelous. We've gone from measuring speed in flops (floating point operations per second) to megaflops (one million flops), gigaflops (a billion) and the now-standard teraflops (one trillion or 10^{12}). I carry around in my pocket today a tiny jump drive that by itself can hold more data than 40 of those PC-AT's, and transfer it almost instantly to any computer.

The phenomenal progress in computing power has generated tremendous new opportunities for the discipline of econometrics, a process of methodological advance that is ongoing with the technological progress in computing. In this essay I call attention to a few of the ways that the practice of econometrics has been transformed by advances in computing and will continue to evolve in the years to come.

2 Very large data bases.

2.1 High-frequency data.

Huge improvements in data storage and access technology have opened up new classes of data sets that can be analyzed. One very interesting area of research concerns high-frequency data, as we have gone from quarterly or monthly data sets to looking now at daily or even tic-by-tic data that capture every single transaction during the day, even on exchanges where such transactions can be huge in number.

One of the things one sees immediately in such data sets is that a much richer concept of "seasonality" is needed than that with which those using quarterly or monthly data have grown accustomed. There can be profound time-of-day effects (e.g., Andersen and Bollerslev, 1998) and day-of-the-week effects (e.g., Hamilton, 1996), which can interact in complicated ways with the dates of events of institutional importance (e.g., Hamilton, 1998). Despite the size of these data sets, it is therefore critical for the researcher to try to get a visual representation of these effects. Parsimonious descriptions of such seasonal patterns almost certainly need to be informed by institutional knowledge and the observed behavior of the actual data set at hand, rather than something that one can pull "off-the-shelf" such as the old Box-Jenkins (1976) approaches to seasonality.

One also finds that time-series behavior can be very different for high-frequency data than for low-frequency data. Consider for sake of discussion the volatility of stock returns. Suppose we thought of the log of the stock price at time t, denoted $p(t)$, as something that

exists at every continuous instant ($t \in [0, T]$). Suppose our data has been discretely sampled at h evenly-spaced points in time, so that we only observe $p(0)$, $p(m_h)$, $p(2m_h)$, ..., $p(T)$ for $m_h = T/h$. The "realized volatility", introduced by Andersen, et. al. (2001), summarizes the variability of returns in the data set in terms of the following magnitude:

$$V_h = \sum_{j=1}^{h} \{p(j \cdot m_h) - p[(j-1)m_h]\}^2 .$$

In other words, we simply take the sum of squares of all observed returns. If the value of $p(t)$ follows a continuous-time diffusion process with instantaneous variance $\sigma^2(t)$, then V_h should converge to $\int_0^T \sigma^2(t)dt$ as $h \to \infty$.

In practice, however, that's not what's found. Figure 1, taken from Andersen, Bollerslev, and Diebold (2000), plots V_h as a function of k_h, the length of time in minutes associated with the time interval m_h for a couple of representative cases. According to the theory just mentioned, this should be converging to some constant as k_h goes to zero. In fact, for the first panel, V_h seems to continue to increase as k_h decreases. The authors describe this as a typical pattern for a highly liquid asset, and attribute it to the negative correlation of $p(t-m_h)$ for small m_h that arises from bid-ask bounce for such securities. The second panel, for an illiquid asset, shows an opposite pattern of plunging V_h as k_h gets small, arising from long episodes in which there are no trades in the security. One can learn about the mechanics of the microstructure of how financial markets operate, such as the ultimate determinants of bid-ask spreads and differences between observed prices and the price that would hold in a frictionless world, by studying how the properties of the time series change as m_h shrinks. See Hansen and Lunde (2006) for an interesting discussion of these possibilities.

Another issue one needs to deal with as one goes all the way to tic-by-tic data is that the particular instants $(t_1, t_2, ...)$ at which securities were traded are themselves random, and these dates are ultimately the result of the same forces that shaped the prices $p(t_1), p(t_2), ...$ themselves. Here again is another rich area for future research that is only just beginning; Engle (2000) has an intriguing overview.

In addition to prices of securities or exchange rates, another exciting new source of high-frequency data comes from retail scanners, which keep track of every single retail purchase. Here again one finds the actual data to be quite different from many of our preconceptions, with prices often returning to previously fixed prices after short sales (Levy, Dutta, and Bergen, 2002) that may be part of a "loss-leader" strategy (Chevalier, Kashyap, and Rossi, 2003). Again the seasonality in such data sets is quite rich (Fok, Franses and Paap, 2006), and again predicting the timing as well as the value of events is of independent methodological and substantive interest (Davis and Hamilton, 2004). Developing econometric methods that are useful for such data sets is another of the open challenges for the coming decade; for some promising ideas see (Bijwaard, Franses and Paap, 2006, and Fok, et. al., 2006).

2.2 Large panels.

Another area that advancing computing power has opened up is combining the information of large numbers of parallel time series. Given observations on n different time series observed at date t, as represented by the $(n \times 1)$ vector $\mathbf{y}_t$, one can in principle summarize a linear

dynamic structure simply enough, for example, with a p-th order vector autoregression:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t. \tag{1}$$

The issue is that the above representation includes $n^2 p$ unknown parameters in the $\mathbf{\Phi}_j$ matrices and another $n(n+1)/2$ in the covariance matrix $\mathbf{\Omega} = E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t')$. For large n, the pressing need is to develop parsimonious representations of the dynamic interactions that capture most of what is going on. Generalizations of dynamic factor models are one promising approach, such as Forni, et. al. (2000) and Stock and Watson (2002). Using informative Bayesian priors is another idea of active ongoing research. These priors can either be general shrinkage of all coefficients toward zero, favoring simple parsimonious structures as in De Mol, Giannone, and Reichlin (2006), shrinkage toward specific typical dynamic patterns as in Sims and Zha (1998), or shrinkage toward fully specified dynamic stochastic general equilibrium models as in del Negro and Schorfheide (2004).

The rich possible contemporaneous correlation structure in $\mathbf{\Omega}$ arising from geographic proximity or economic similarity of the different individual elements of $\mathbf{y}_t$ is another area that econometricians are still in the early stages of exploring. Recent exciting ideas include Chen and Conley (2001) and Conley and Dupor (2003). Integrating the methods and advances in geographic information systems with econometrics is another topic that may well feature prominently in new econometric developments over the next few years.

3 The simulation revolution in Bayesian inference.

To me, the most striking development in econometrics over the last decade has been what van Dijk (1999) referred to as the "simulation revolution in Bayesian econometric inference." Taking the VAR in equation (1) for illustration, let $\boldsymbol{\theta}$ denote a vector containing the unknown elements of $\mathbf{c}, \boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, ..., \boldsymbol{\Phi}_p, \boldsymbol{\Omega}$ and let $\mathbf{y} = (\mathbf{y}_T^{'}, \mathbf{y}_{T-1}^{'}, ..., \mathbf{y}_1^{'})'$ denote the vector containing all the observations for all dates. A classical econometrician forms an estimate of $\boldsymbol{\theta}$ such as the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, and asks what the distribution of $\hat{\boldsymbol{\theta}}$ would be if one repeated the inference on a large number of samples just like the present one. By contrast, the Bayesian views the population parameter $\boldsymbol{\theta}$ as something inherently unknown and about which we are uncertain, and is willing to summarize that subjective uncertainty in terms of a probability density $p(\boldsymbol{\theta})$, e.g., making statements such as, "I believe there is a 20% probability that θ_1 exceeds 2.5." The Bayesian econometrician had beliefs of this form, summarized by the density $p(\boldsymbol{\theta})$ called the prior, before seeing any data, though these might be quite vague, allowing some probability of virtually any possibility. The Bayesian's goal is to use the observed data $\mathbf{y}$ to calculate the posterior density $g(\boldsymbol{\theta}|\mathbf{y})$, or a mathematical summary of what we believe now that we've seen the data. This posterior density is found using Bayes' Law from

$$g(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}) f(\mathbf{y}|\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{2}$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ denotes the likelihood function and the integral in the denominator of (2) represents a definite integral over all possible values of $\boldsymbol{\theta}$.

The Bayesian knows his or her prior $p(\boldsymbol{\theta})$ by definition, and the likelihood function

$f(\mathbf{y}|\boldsymbol{\theta})$ is often something we know how to calculate as well. The problem is that, except for certain nice classes of densities $p(\boldsymbol{\theta})$ and $f(\mathbf{y}|\boldsymbol{\theta})$, the integral in the denominator of (2) is not known analytically. The key insight of the numerical Bayesian revolution is that one in fact never needs to calculate this denominator. The revolution started with the development of algorithms that allow one to simulate draws from the distribution $g(\boldsymbol{\theta}|\mathbf{y})$ without ever needing to calculate the distribution itself.

One of the key tools of Bayesian simulation, importance sampling, dates back to Hammersly and Handscomb (1964) and Kloek and van Dijk (1978), the latter published six years before I purchased the IBM PC-AT described in the introduction. As noted by van Dijk (1999), it was really the engineering breakthroughs in computer technology that allowed such methods to become prominent over the last decade.

Another key development besides importance sampling was the Gibbs sampler developed by Geman and Geman (1984), Tanner and Wong (1987), and Gelfand and Smith (1990). The Gibbs sampler can be used in situations where even the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ itself is too complicated to calculate and $g(\boldsymbol{\theta}|\mathbf{y})$ too difficult to simulate from, but there exists a way of breaking $\boldsymbol{\theta}$ into blocks such that draws of $g_1(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2)$ and then $g_2(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1)$ can be simulated. This frequently arises in economic applications in which there is an unobserved latent variable, such as a regime s_t that characterized the system at date t. From a Bayesian perspective, the randomness of these unobserved latent variables is fundamentally no different from the randomness of the inherently unknown parameters, so that one could collect the unknown values of the regimes into a sublock of $\boldsymbol{\theta}$, viz., $\boldsymbol{\theta}_2 = (s_T, s_{T-1}, ..., s_1)$. In

such settings, typically knowledge of the probability law for the latent variable would allow one to simulate draws from $g_2(\boldsymbol{\theta}_2|\mathbf{y}, \boldsymbol{\theta}_1)$. With these given numerical values for $\boldsymbol{\theta}_2$, simulation from $g_1(\boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_2)$ may likewise be straightforward, and the Gibbs sampler is implemented simply by zigzagging back and forth between these.

Such methods, popularized in econometrics by Chib and Greenberg (1996), have allowed estimation and inference for broad classes of econometric models for which the likelihood function itself $f(\mathbf{y}|\boldsymbol{\theta})$ is impossible to calculate. These include stochastic volatility models (Jacquier, Polson, and Rossi, 1994; Kim, Shephard and Chib, 1998), multinomial probit models (McCulloch and Rossi, 1994), regime-switching models with stochastic, time-varying transition probabilities (Filardo and Gordon, 1998), nonlinear filtering (Pitt and Shephard, 1999), nonlinear diffusions (Elerian, Chib, and Shephard, 1998), and state-space models with changes in regime (Kim and Nelson, 1999).

And this of course is why econometricians who do not subscribe to the Bayesian perspective nevertheless find these methods of considerable interest and use. Classical methods such as maximum likelihood are infeasible if the likelihood function cannot even be calculated. The Bayesian posterior mean calculated from such simulation methods can be viewed as an approximation to the classical maximum likelihood estimate (Chernozhukov and Hong, 2003), which, in the absence of Bayesian simulation algorithms, could not be calculated.

The ability to estimate parameters for such complicated, intractable models is quite liberating, and there are many more applications to be developed in the specialties of interest to economists, including finance, macroeconomics, labor economics, and industrial organi-

zation.

4 Computer-generated theory.

A final point I wish to address on this theme of how advances in computing ability are in the process of reshaping the practice of econometrics is related to the role of economic theory in guiding econometric estimation and inference. Our economic theory itself is increasingly something that is arrived at by computer calculations and simulations. Certainly in macroeconomics, dynamic stochastic general equilibrium (DSGE) models are now typically analyzed by log-linearizing around the steady state and then using some of the powerful algorithms such as King and Watson (1988) or Klein (2000) to solve numerically. These imply a state-space structure for the observed variables which can then be related to the actual data.

I believe we are currently on the verge of an important breakthrough in these efforts. Traditionally, many macroeconomists have been content to look at a few summary correlations in judging the usefulness of DSGE modeling efforts. However, for policy purposes, one really needs models with much more concrete predictions, and, more importantly, models that are not rejected by the data. The state-space representation of a typical DSGE model implies a huge number of restrictions on the likelihood function, restrictions that are trivial to reject using standard hypothesis tests. The challenge facing researchers is to add enough complications to these frameworks so as to make them fully data coherent. Interesting initial efforts along these lines include Smets and Wouters (2003) and Jung (2006). I expect

this to be one of the most important priorities for macroeconomic research over the next five years.

It may be that as these models become more data-coherent, they will lose some of the flavor of the early DSGE's as they incorporate more realism and institutional details. One of the early success stories along these lines that I think we can claim is coming with our descriptions of the market for interbank trades in Federal Reserve deposits, as in Clouse and Dow (2002).

5 Conclusions.

In these brief remarks, I have explored some of the possibilities that advances in computing have opened up for the field of econometrics and implications for future research. The reader may have noted another theme as well– all of the areas I have discussed are very much application-focused. I'm basically arguing for a research attitude in which the econometrician is very closely tuned in to the questions that applied researchers need to ask and the particular features of the data being studied. The more we try to take advantage of the opportunities that greater computing power provides us, the more rewards I believe are to be obtained from such an attitude.

6 References

Andersen, Torben, and Tim Bollerslev (1998). "DM-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies," *Journal of Finance* 53: 219-265.

_____, _____, Francis X. Diebold, Francis X. (2000). "Great Realizations," *Risk:* 105-108.

_____, _____, _____, and Paul Labys (2001). "The Distribution of Realized Exchange Rate Volatility," *Journal of the American Statistical Association*: 96, 42-55.

Bijwaard, Govert E., Philip Hans Franses, and Richard Paap (2006). "Modeling Purchases as Repeated Events," *Journal of Business & Economic Statistics*, forthcoming.

Box, George E. P., and Gwilym M. Jenkins (1976). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Chen, Xiaohong and Timothy G. Conley (2001). "A New Semiparametric Spatial Model for Panel Time Series, *Journal of Econometrics* 105: 59-83.

Chernozhukov, Victor, and Han Hong (2003). "An MCMC Approach to Classical Estimation," *Journal of Econometrics* 115: 293-346.

Chevalier, Judith A., Anil K. Kashyap, and Peter E. Rossi (2003). "Why Don't Prices Rise During Periods of Peak Demand? Evidence from Scanner Data," *American Economic Review* 93: 15-37.

Chib, Siddhartha, and Edward Greenberg (1996). "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory* 12: 409-431.

Clouse, James A., and James D. Dow, Jr. (2002). "A Computational Model of Banks' Optimal Reserve Management Policy," *Journal of Economic Dynamics and Control* 26: 1787-1814.

Conley, Timothy G., and Bill Dupor (2003). "A Spatial Analysis of Sectoral Complementarity," *Journal of Political Economy* 111: 311-352.

Davis, Michael C., and James D. Hamilton (2004). "Why Are Prices Sticky? The Dynamics of Wholesale Gasoline Prices," *Journal of Money, Credit, and Banking* 36: 17-37.

van Dijk, Herman K. (1999). "Some Remarks on the Simulation Revolution in Bayesian Econometric Inference," *Econometric Reviews* 18: 105-112.

Elerian, Ola, Siddhartha Chib and Neil Shephard (2001). "Likelihood Inference for Discretely Observed Nonlinear Diffusions," *Econometrica* 69: 959-994

Engle, Robert F. (2000). "The Econometrics of Ultra-High-Frequency Data," *Econometrica* 68:1-22.

Filardo, Andrew J., and Stephen F. Gordon (1998). "Business Cycle Durations," *Journal of Econometrics,* 85: 99-123.

Fok, Dennis, Philip Hans Franses and Richard Paap (2006). "Seasonality and Non-linear Price Effects in Scanner-data based Market-response Models," *Journal of Econometrics*, forthcoming.

_____, Csilla Horváth, Richard Paap, and Philip Hans Franses (2006). "A Hierarchical Bayes Error Correction Model to Explain Dynamic Effects of Price Changes," *Journal of Marketing Research*, forthcoming.

Forni, Mario, Marc Hallin, Marco Lippi, and Lucrezia Reichlin (2000). "The Generalized Dynamic Factor Model: Identification and Estimation," *Review of Economics and Statistics* 82:540-554.

Gelfand, A.E., and A.F.M. Smith (1990). "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association* 85: 398-409.

Geman, S., and D. Geman (1984). "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12: 609-628.

Hamilton, James D. (1996). "The Daily Market for Federal Funds," *Journal of Political Economy*, 104: 26-56.

_____ (1998). "The Supply and Demand for Federal Reserve Deposits," *Carnegie-Rochester Conference Series on Public Policy,* Volume 49, pp. 1-52, edited by Bennett T. McCallum, et. al.

Hammersly, J.M., and D.C. Handscomb (1964). *Monte Carlo Methods,* London: Methuen & Co.

Hansen, Peter R., and Asger Lunde (2006). "Realized Variance and Market Microstructure Noise with Comments and Rejoinder," *Journal of Business and Economic Statistics,* forthcoming.

Jacquier, Eric, Nicholas G. Polson, and Peter E. Rossi (1994). "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics,* 12: 371-388.

Jung, Yong-Gook (2006). "Investment Lags and Macroeconomic Dynamics," working

paper, University of California, San Diego.

Kim, Chang-Jin and Charles R. Nelson (1999). *State-Space Models with Regime Switching*, MIT Press.

Kim, Sangjoon, Neil Shephard and Siddhartha Chib (1998). "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *Review of Economic Studies* 65: 361–93.

King, Robert G., and Mark W. Watson (1988). "The Solution of Singular Linear Difference Systems under Rational Expectations," *International Economic Review* 39: 1015-1026.

Klein, Paul (2000). "Using the Generalized Schur Form to Solve a Multivariate Linear Rational Expectations Model," *Journal of Economic Dynamics and Control* 24: 1405-1423.

Kloek, Teun, and Herman K. van Dijk (1978). "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica* 46: 1-19.

Levy, Daniel, Santanu Dutta, and Mark Bergen (2002). "Heterogeneity in Price Rigidity: Evidence from a Case Study Using Microlevel Data," *Journal of Money, Credit, and Banking* 34: 197-220.

McCulloch, Robert, and Peter E. Rossi (1994) "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics* 64: 207-240.

De Mol, Christine, Dominico Giannone, and Lucrezia Reichlin (2006). "Forecasting Using a Large Number of Predictors: Is Bayesian Regression a Valid Alternative to Principal Components?" working paper, Université Libre de Bruxelles.

del Negro, Marco, and Frank Schorfheide (2004). "Priors from General Equilibrium

Models for VARS," *International Economic Review* 45: 643-673.

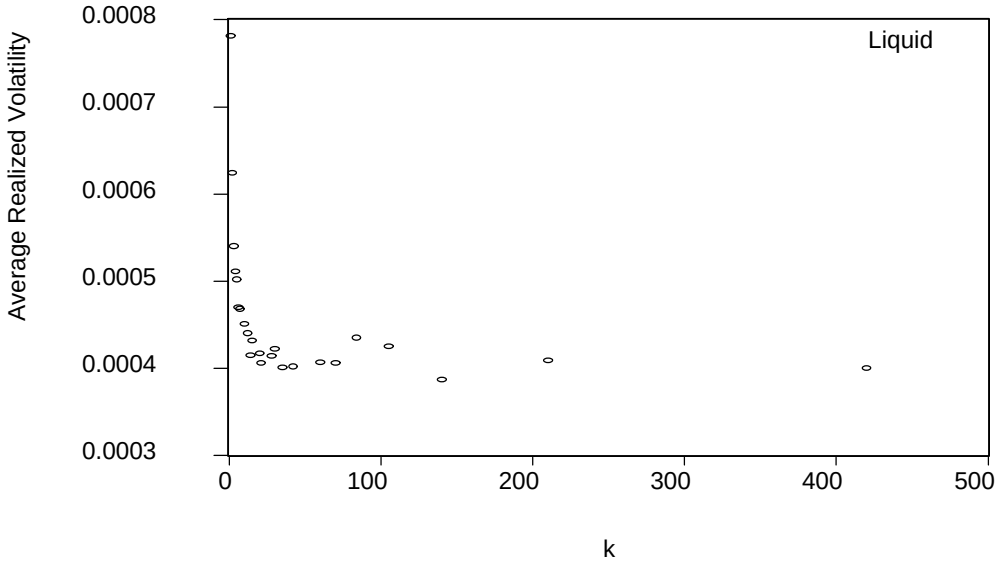
Pitt, Michael K., and Neil Shepard (1999). "Filtering via Simulation: Auxiliary Particle Filter," *Journal of the American Statistical Association* 94: 590-9.

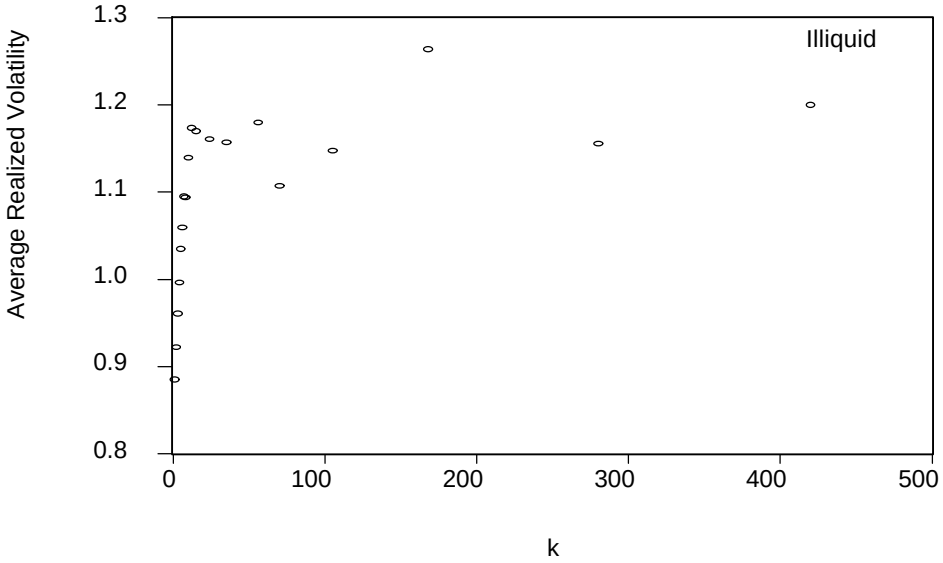
Sims, Christopher A., and Tao Zha (1998). "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review* 39: 949-968.

Smets, Frank and Raf Wouters (2003). "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area," *Journal of the European Economic Association* 1: 1123-1175.

Stock, James H., and Mark W. Watson (2002). "Forecasting Using Principal Components from a Large Number of Predictors ," *Journal of the American Statistical Association* 97: 1167-1179.

Tanner, M.A., and W.H. Wong (1987). "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association* 82: 528-549.


Liquid
Average Realized Volatility
0.0008
0.0007
0.0006
0.0005
0.0004
0.0003
0
100
200
300
400
500
k


Illiquid
Average Realized Volatility
1.3
1.2
1.1
1.0
0.9
0.8
0
100
200
300
400
500
k