Editorial

Big Data in Dynamic Predictive Econometric Modeling

The modern economic environment is awash in Big Data. The business models of many of today's largest and most innovative firms, for example, are intimately connected to data collection and processing. Closely related, all of science is now similarly awash in Big Data, due to massive advances in data capture and storage technology, typically in conjunction with the internet.[1]

Time-series econometrics is no exception, and the time-series community is beginning to confront Big Data. Whereas leading texts of a few decades ago like Hamilton (1994) had no mention of Big Data topics, recent texts like Ghysels and Marcellino (2018) cover regularization methods, factor models for large panels, etc. Nevertheless the time-series econometrics of Big Data is still in its infancy, with many important issues explored incompletely or not at all.

Against this background, in this volume we present new econometric research at the interface of Big Data and predictive time-series modeling. Topics include, but are not limited to: shrinkage, selection, sparsity, dimensionality reduction, structural change, high-frequency data, mixed-frequency data, and network topology description.

A simple taxonomy is useful for understanding the types of Big Data that will concern us, and hence how the volume's papers relate and cohere. To that end, consider a time-series regression involving T time periods and K regressors, with intra-period sampling m times per period. Then the "X matrix" is $mT \times K$, and Big Data corresponds to situations of very large T, K, and/or m.

T, K, and m are usefully considered separately, although more than one can of course be large at once. As T gets large we have "tall data", in reference to the tall X matrix, due to the large number of periods, i.e., the long calendar span of data. As K gets large (potentially even so large that $K>T$) we have "wide data", in reference to the wide X matrix due to the large number of regressors. As m gets large we have "dense data", in reference to the high-frequency intra-period sampling, regardless of whether the data are tall.

Still maintaining our simple regression motivation, there is nevertheless a fourth source of Big Data: the regression might be *multivariate*, say N-dimensional. As N gets very large we have "high-dimensional data". More generally we refer to situations involving large N or K, or both, as high-dimensional. (Large K is effectively high-dimensional because endogenizing the regressors in a large-K univariate regression would produce a large-N vector autoregression.) We confront the problem raised by high dimensionality – a small number of degrees of freedom relative to the estimation task at hand – by regularization, that is, by imposing restrictions of one sort or another that allow us to recover some degrees of freedom.

The regularization spectrum runs from unconstrained through deterministically constrained. At one end is completely unconstrained estimation. Next comes stochastically constrained estimation; that is, with estimates coaxed in a certain direction without being

[1] The origin of the term "Big Data", in addition to the phenomenon itself, is also interesting. See Diebold (2012).

forced. We speak of "shrinkage", and shrinkage strength can of course vary. (In a Bayesian interpretation, shrinkage is toward the prior mean, with strength governed by prior precision.) In the limit we have deterministic restrictions such as exact zeros, promoting sparsity and variable selection, and other restrictions such as reduced rank, associated with factor structure and cointegration.[2]

Many of the papers in this volume involve large K and/or N and proceed via some sort of shrinkage-type regularization. Examples include Billio, Casarin, and Rossini, "Bayesian Nonparametric Sparse Vector Autoregressive Models"; Carriero, Clark and Marcellino, "Large Bayesian Vector Autoregressions with Stochastic Volatility and Non-Conjugate Priors"; and Chen, Li, and Linton, "A New Semiparametric Estimation Approach of Large Dynamic Covariance Matrices with Multiple Conditioning Variables".

Others proceed with reduced-rank restrictions. Examples include Andreasen, Christensen, and Rudebusch, "Term Structure Analysis with Big Data: One-Step Estimation Using Bond Prices"; Bai and Ng, "Principal Components and Regularized Estimation of Factor Models"; Onatskiy and Wang, "Extreme Canonical Correlations and High-Dimensional Cointegration Analysis"; and Fan, Gong, and Zhu, "Generalized High-Dimensional Trace Regression via Nuclear Norm Regularization".

Still others proceed with dynamic restrictions, as in Korobolis and Pettenuzzo, "Adaptive Hierarchical Priors for High-Dimensional Vector Autoregressions".

One might naively assert that tall data are not really a part of the Big Data phenomenon. (Time has not started moving more quickly, so the sample calendar span, T, is still typically not very big.) But a more sophisticated assessment of the size of T involves whether it is big enough to make structural change a potentially serious concern. And structural change *is* a serious concern, routinely, in time-series econometrics. Hence structural change effectively makes T big, and confronting strutuctural change is a part of confronting the large-T aspect of Big Data. Contributions in this volume include Petrova, "A Quasi-Bayesian Nonparametric Approach to Time Varying Parameter VAR Models" (smoothly time-varying parameters), and Smith, Timmerman, and Zhu, "Variable Selection in Panel Models with Breaks" (abruptly time-varying parameters).

Large-m Big Data have featured prominently in recent decades, as automated data collection now proceeds in near-continuous time in a variety of contexts, from measuring climactic conditions to measuring trades in financial markets. Moreover, most no-arbitrage financial economic models are written in continuous time, and uncovering their primitives, particularly stochastic volatility and jumps, is facillitated by high-frequency data. The large realized volatility literature, for example, emphasizes estimating aspects of semimartingales (drifts, stochastic volatilities, jumps) using high-frequency data.[3] Contributions in this volume include Andersen, Fusari, Todorov, and Varneskov, "Unifed Inference for Nonlinear Factor Models from Panels with Fixed and Large Time Span", and Bollerslev, Meddahi, and Nyawa, "High-Dimensional Multivariate Realized Volatility Estimation".

Interestingly, large K and/or N Big Data often involve *mixed* m. That is, Big Data tend to be mixed-frequency data: when many series are examined, it is highly unlikely that all will be measured at the same frequency, unless all frequencies but one are arbitrarily discarded. Hence mixed-frequency data arises naturally in Big-Data contexts. Related contributions in this volume include Babii, Chen and Ghysels, "Commercial and Residential Mortgage

[2]Note that all restrictions, not just explicit zero restrictions, effectively promote sparsity, appropriately interpreted.

[3]See Aït-Sahalia and Jacod (2014) for a unified overview.

Defaults: Spatial Dependence with Frailty", and Hautsch and Voigt, "Large-Scale Portfolio Allocation Under Transaction Costs and Model Uncertainty".

Regularization is largely concerned with estimation in large K and/or N environments, but there remains the issue of interpreting and understanding regularized estimation results. For example, a 1000-dimensional vector autoregression will still have a huge number of hard-to-interpret estimated parameters, even if successfully regularized. Methods for network topology summarization and visualization can greatly facillitate model interpretation in Big Data environments, as suggested by Demirer et al. (2018). Related contributions in this volume include Hale and Lopez, "Monitoring Banking System Connectedness with Big Data", and Zhu, Wang, Wang, and Härdle, "Network Quantile Autoregression".

Thus far we have classified our papers by the variety of methodological areas emphasized: shrinkage, selection, sparsity, dimensionality reduction, structural change, high-frequency data, mixed-frequency data, network topology, etc. But we hasten to add that we could equally have classified them by the variety of substantive applications addressed: asset pricing, portfolio allocation, risk measurement and management, bond markets, macroeconomics, financial networks, mortgage markets, etc. For example, Mykland's paper, "Combining Statistical Intervals and Market Prices: The Worst Case State Price Distribution", combines aspects of risk measurement and asset pricing. In addition, many papers span *multiple* methodological and substantive areas.

Finally, we gratefully acknowledge encouragement from Yacine Aït-Sahalia and the editorial Board of *Journal of Econometrics*, and financial support from the University of Pennsylvania Warren Center, the University of Chicago Stevanovich Center, and the Penn Institute for Economic Research.

Francis X. Diebold*
University of Pennsylvania
E-mail address: fdiebold@sas.upenn.edu.

Eric Ghysels
University of North Carolina

Per Mykland
University of Chicago

Lan Zhang
University of Illinois at Chicago

* Corresponding Editor.

References

Aït-Sahalia, Yacine and Jean Jacod (2014), *High-Frequency Financial Econometrics*, Princeton University Press.

Demirer, Mert, Francis X. Diebold, Laura Liu, and Kamil Yilmaz (2018), "Estimating Global Bank Network Connectedness," *Journal of Applied Econometrics*, 33, 1–15.

Diebold, Francis X. (2012), "On the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and and the Discipline," Manuscript, Department of Economics, University of Pennsylvania, `http://www.ssc.upenn.edu/~fdiebold/papers/paper112/Diebold_Big_Data.pdf`.

Ghysels, Eric and Massimiliano G. Marcellino (2018), *Applied Economic Forecasting Using Time Series Methods*, Oxford University Press.

Hamilton, James D. (1994), *Time Series Analysis*, Princeton University Press.