14 Big Data in Economics: Evolution or Revolution?

Christine De Mol, Eric Gautier, Domenico Giannone, Sendhil Mullainathan, Lucrezia Reichlin, Herman van Dijk and Jeffrey Wooldridge

Abstract

The Big Data Era creates a lot of exciting opportunities for new developments in economics and econometrics. At the same time, however, the analysis of large datasets poses difficult methodological problems that should be addressed appropriately and are the subject of the present chapter.

14.1 Introduction

'*Big Data*' has become a buzzword both in academic and in business and policy circles. It is used to cover a variety of data-driven phenomena that have very different implications for empirical methods. This chapter discusses some of these methodological challenges.[1]

In the simplest case, '*Big Data*' means a large dataset that otherwise has a standard structure. For example, Chapter 13 describes how researchers are gaining increasing access to administrative datasets or business records covering entire populations rather than population samples. The size of these datasets allows for better controls and more precise estimates and is a bonus for researchers. This may raise challenges for data storage and handling, but does not raise any distinct methodological issues.

However, '*Big Data*' often means much more than just large versions of standard datasets. First, large numbers of units of observation often come with large numbers of variables, that is, large numbers of possible covariates. To illustrate with the same example, the possibility to link different administrative datasets increases the number of variables attached to each statistical unit. Likewise, business records typically contain all consumer interactions with the business. This can create a tension in the estimation between the objective of 'letting the data speak' and obtaining accurate (in a way to be specified later) coefficient estimates. Second, Big Data sets often have a very different structure from those we are used to in economics. This includes web search queries, real-time

geolocational data or social media, to name a few. This type of data raises questions about how to structure and possibly re-aggregate them.

The chapter starts with a description of the '*curse of dimensionality*', which arises from the fact that both the number of units of observation and the number of variables associated with each unit are large. This feature is present in many of the Big Data applications of interest to economists. One extreme example of this problem occurs when there are more parameters to estimate than observations. In this case, standard estimators (such as ordinary least squares) do not yield a unique solution. The section, which borrows heavily from De Mol et al. (2008), describes the econometric problems raised by the curse of dimensionality. It describes some of the methodological solutions called regularization methods that have been proposed.

Section 14.3 then discusses recent research on recovering policy effects using Big Data. In many fields of economics, we are interested in measuring a (causal) relationship between some variable of interest (for example, a policy) and its effects. In other words, although there might be many variables, some of them (related to a specific policy) are of special interest to the researcher. The section describes current efforts to develop methods that combine the ability of regularization methods to harness the information contained in these richer datasets with the possibility to identify the impact of specific policy relevant effects.

Section 14.4 turns to prediction problems. Here we are not interested in specific coefficients per se but in our ability to forecast a variable of interest, for example, inflation, growth or the probability of default. Forecasting has a long tradition in macroeconomics and the greater availability of highly granular microdata is creating renewed interest in prediction problems also at the microeconomic level. A priori, regularization methods are well-suited for this type of problem. However, 'off-the-shelf' regularization methods are agnostic regarding the data generation process. On the basis of the experience with macro forecasting models, the section argues for the need to develop regularization methods that account for the specificities of the data generation processes in economics, such as serial correlation or mixed frequencies.

Recent progress in computing power and storage capacities has allowed researchers to handle and analyse increasingly big datasets. For some of the Big Data (e.g., high frequency trading data, browsing data), this may not be enough. Section 14.5 discusses how simulation-based methods can be refined to leverage the potential of parallel computing.

Section 14.6 concludes. The availability of unprecedented amounts of data offers exciting research opportunities in economics. While researchers will be able to exploit some of the methods developed in other fields, such as statistics and computer science, it is essential that some of these methods be tailored to the specificities of economic research questions and economic data. On these fronts, there is still much to be done.

14.2 The Curse of Dimensionality and Regularization

An early occurrence of the term '*Big Data*' in economics is to be found in a discussion by Diebold (2003, 2012). To quote, 'I stumbled on the term Big Data innocently enough, via discussion of two papers that took a new approach to macro-econometric dynamic factor models (DFMs), Reichlin (2003) and Watson (2003), presented back-to-back in an invited session of the 2000 World Congress of the Econometric Society.'

The two authors referenced above were presenting their research on factor models in high-dimensional time series (Forni et al., 2000, Stock and Watson, 2002), which mainly consisted in deriving asymptotic results for the case where both the number of time samples and the cross-sectional dimension, that is, the number of time series, tend to infinity. The approach relied on a factor model dating back to Chamberlain and Rothschild (1983) in finance, but generalized to take serial correlation into account. Stock and Watson (2002) considered so-called '*static*' factor models, whereas Forni et al. (2000) derived asymptotics in the case of '*dynamic*' factor models. The estimators they used are based on a few principal components either in the time domain for the static case or in the Fourier domain for the dynamic case. This factor-model literature was probably the first in economics to address the difficulties arising from the high dimensionality of the data, albeit under rather strong assumptions (namely factor models) on the data generating process.

In statistics, the difficulties pertaining to the analysis of high-dimensional data are well-known issues, often referred to as the '*curse of dimensionality*'. Some of the facets of this curse can be explained using the familiar example of the linear regression model. To introduce some background notation useful for our discussion, let

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}, \tag{14.1}$$

where $\mathbf{X}$ is a $n \times p$ matrix containing the observed predictors (covariates), $\mathbf{Y}$ is the outcome or $n \times 1$ vector of the observed responses and $\mathbf{U}$ is an unobserved zero-mean error or nuisance term. The $p \times 1$ vector β contains the regression coefficients. In the case of time series, n is the number of time samples and p is the number of time series used for prediction. In the case of cross-section data, n is the number of observations and p is the number of covariates. In the discussion in this section, we will consider the matrix $\mathbf{X}$ as deterministic.

Depending on the application under study, two different problems can be highlighted. The first is prediction (also referred to as 'generalization' by the machine-learning community), in which case one is only interested in estimating the outcome for future times or new examples to come. This requires the estimation of the regression parameters, but only as an auxiliary step to the estimation of the outcome. The second problem, the identification of the

model, pertains more to the vector β of regression coefficients itself, in the linear regression example in (14.1). This is essential for interpreting the estimated coefficients in terms of their relevance in predicting the response. For example, some coefficients can be zero, indicating that the corresponding predictors are not relevant for this task. The determination of these zeroes, hence of the relevant/irrelevant predictors, is usually referred to as '*variable selection*'.

As is well known, the most straightforward solution for the linear regression problem is Ordinary Least Squares (OLS). The OLS estimator for β in (14.1) minimizes the least-squares loss

$$\Phi(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 , \tag{14.2}$$

where $\|\mathbf{Y}\|_2 = \sqrt{\sum_{i=1}^{n} |\mathrm{Y}_i|^2}$ is the L_2-norm of the vector $\mathbf{Y}$. It is given by

$$\hat{\beta}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{14.3}$$

($\mathbf{X}'$ denotes the transpose of the matrix $\mathbf{X}$).

For the OLS estimator, expression (14.3), to make sense we need the $p \times p$ matrix $\mathbf{X}'\mathbf{X}$ to be of full rank, hence invertible. This cannot be the case in high-dimensional situations where the number of coefficients, p, is larger than the number of observations, n.[2] In that case, the minimizer of the least-squares loss is nonunique, but uniqueness can be restored by selecting the so-called '*minimum-norm least-squares solution*', orthogonal to the null-space, that is, by ignoring the subspace corresponding to the zero eigenvalues.

Notice that although this remedy may work well for prediction, the identification problem remains hindered by this nonuniqueness issue. An additional difficulty arises when the matrix $\mathbf{X}'\mathbf{X}$ has eigenvalues that are close to zero, or more precisely, when its '*condition number*', that is, the ratio between the largest and the smallest of its nonzero eigenvalues, becomes large. This situation prevents a stable determination of the least-squares (or minimum-norm least-squares) estimator: small fluctuations in the outcome vector $\mathbf{Y}$ will be amplified by the effect of the small eigenvalues and will result in large uncontrolled fluctuations (high variance/volatility) on the estimation of β, again preventing meaningful identification.

The pathologies described above contribute to what is often referred to as the '*curse of dimensionality*' or else the '*large p, small n paradigm*' in high-dimensional statistics. As early as in the 1950s, Stein (1956) introduced a 'high-dimensional' surprise in statistics by showing that the maximum-likelihood estimator of the unknown mean vector of a multivariate Gaussian distribution is not 'admissible' in a dimension higher than three, that is, that it is outperformed by '*shrinkage*' estimators. Heuristically, 'shrinking' means that a naive estimate is improved by combining it with other information or priors.

Many remedies have been proposed to address these pathologies under the common designation of '*regularization methods*', which provide in one form or

Box 14.1: Principal Component Regression (PCR)

The Principal Component Regression consists in estimating β by

$$\hat{\beta}_{pcr} = \sum_{i=1}^{k} \frac{\langle \mathbf{X}'\mathbf{Y}, \mathbf{V}_i \rangle}{\xi_i^2} \mathbf{V}_i \tag{14.4}$$

where the $\mathbf{V}_i$'s are the eigenvectors of $\mathbf{X}'\mathbf{X}$ with eigenvalues ξ_i^2, and $\langle \cdot, \cdot \rangle$ denotes the scalar product.

another the dimensionality reduction necessary to reduce the variance/volatility of unstable estimators, or in other words, to avoid '*overfitting*'. Overfitting refers to the fact that, when using a model with many free parameters (here the p components of β), it is easy to get a good fit of the observed data, that is, a small value for the residual (14.2), but that this does not imply that the corresponding (unstable) value of β will have a good predictive power for responses corresponding to new observations. For time series, good in-sample fit does not imply good out-of-sample forecasts.

One of the simplest regularization methods is principal component regression (PCR), a statistical procedure that transforms the possibly correlated variables into a smaller number of orthogonal new variables (the components, see Box 14.1). The truncation point k for the number of components, usually much smaller than the true rank of $\mathbf{X}'\mathbf{X}$, has to be carefully chosen to overcome instabilities. In this method, also referred to as '*Truncated Singular Value Decomposition*' (TSVD), the truncation introduces a bias in order to reduce variance.

Until recently alternative estimators were less well known in econometrics. Other regularization methods introduce constraints or penalties on the vector β of the regression coefficients. Probably the oldest penalized regression method is '*Ridge regression*'(see Box 14.2), due to Hoerl and Kennard (1970). This method is also known in the applied mathematics literature as Tikhonov's regularization. It consists in adding to the least-squares loss a penalty proportional to the size of β, measured by its squared L_2-norm. As for the truncation point in PCR, the regularization parameter has to be chosen carefully in order to provide a proper balance between the bias introduced by shrinkage, and the variance of the estimator and its value is usually determined by cross-validation.

Ridge regression introduces a form of linear '*shrinkage*', where the components of $\hat{\beta}_{ols}$ are shrunk uniformly towards zero, as can be easily seen in the case of orthonormal regressors (i.e., for $\mathbf{X}'\mathbf{X} = \mathbf{I}$), where $\hat{\beta}_{ridge} = \frac{1}{1+\lambda_2} \mathbf{X}'\mathbf{Y}$. More generally, quadratic penalties provide estimators which depend linearly on the response $\mathbf{Y}$ but do not allow for variable selection, since typically all regression coefficients are different from zero.

Box 14.2: The Ridge Regression Estimator

The ridge regression estimator is given by

$$\begin{aligned}\hat{\beta}_{ridge} &= \text{argmin}_{\beta}\left[\|\mathbf{Y}-\mathbf{X}\beta\|_2^2+\lambda_2\|\beta\|_2^2\right]\\ &= (\mathbf{X}'\mathbf{X}+\lambda_2\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}\end{aligned} \tag{14.5}$$

where $\mathbf{I}$ is the identity matrix and $\lambda_2 > 0$ is the so-called '*regularization parameter*', which, as seen from (14.5), reduces the impact of the smallest eigenvalues of $\mathbf{X}'\mathbf{X}$, at the origin of the instability of the OLS estimator.

An alternative to quadratic penalties that allows for variable selection by enforcing sparsity, that is, the presence of zeroes in the vector β of the regression coefficients, has been popularized in the statistics and machine-learning literature under the name of '*Lasso regression*' by Tibshirani (1996). It consists in replacing the L_2-norm penalty used in ridge regression by a penalty proportional to the L_1-norm of β (see Box 14.3).

In the case of orthonormal regressors, it is easily seen that the Lasso penalty provides a nonlinear shrinkage of the components of $\hat{\beta}_{ols}$, which are shrunk differently according to their magnitude, as well as sparsity, since the jth coefficient $[\hat{\beta}_{lasso}]_j = 0$ if $|[\mathbf{X}'\mathbf{Y}]_j| < \lambda_1/2$. Unfortunately, there is no closed-form expression for $\hat{\beta}_{lasso}$ in the case of general matrices $\mathbf{X}$, and the Lasso estimator has to be computed numerically as the solution of a (nonsmooth) convex optimization problem.

The previous estimators can be given a Bayesian interpretation, since $\hat{\beta}_{ols}$ can be viewed as the maximum (log-)likelihood estimator for a Gaussian error term and the penalized maximum likelihood estimators $\hat{\beta}_{ridge}$ and $\hat{\beta}_{lasso}$ can be interpreted as maximum a posteriori (MAP) estimators, the penalty resulting from a prior distribution for the regression coefficients. In Ridge regression, it corresponds to a Gaussian prior whereas in Lasso regression it is a Laplacian or double-exponential prior.

The regularization techniques described above are paradigmatic since they convey the essential ideas in dealing with high-dimensional settings. There are however numerous extensions and generalizations. For example, more general types of penalties can be used such as $\|\beta\|_\gamma^\gamma = \sum_{j=1}^p |\beta_j|^\gamma$, *i.e.*, the L_γ-norms used in '*bridge regression*' (Frank and Friedman, 1993). Notice that in this family, though, only the choice $\gamma = 1$ yields both convexity and sparsity. Moreover, weights or even a nondiagonal coupling matrix can be introduced in the penalty to cover the case of non i.i.d. (independent and identically distributed) regression coefficients. Composite penalties are also used, for example, in elastic-net or group-lasso regularization. Finally, different loss functions can be considered such as those used in robust statistics, logistic regression, etc. A good pointer to this variety of techniques is the book by Hastie et al. (2009).

Box 14.3: The Lasso Regression Estimator

Lasso consists in replacing the L_2-norm penalty used in ridge regression by a penalty proportional to the L_1-norm of β, that is, to the sum of the absolute values of the regression coefficients, $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$, yielding the estimator

$$\hat{\beta}_{lasso} = \text{argmin}_\beta \left[\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right], \tag{14.6}$$

In the special case of orthonormal regressors ($\mathbf{X}'\mathbf{X} = \mathbf{I}$), the Lasso estimator is easily seen to be given by

$$[\hat{\beta}_{lasso}]_j = S_{\lambda_1}([\mathbf{X}'\mathbf{Y}]_j)$$

where $S_{\lambda_1}(x)$ is the '*soft-thresholder*' defined by

$$S_{\lambda_1}(x) = \begin{cases} x + \lambda_1/2 & \text{if } x \leq -\lambda_1/2 \\ 0 & \text{if } |x| < \lambda_1/2 \\ x - \lambda_1/2 & \text{if } x \geq \lambda_1/2\,. \end{cases}$$

Let us remark that global variable selection methods, preferably convex to facilitate computation, such as the Lasso and its relatives, are essential to deal with high-dimensional situations. Indeed, considering all possible submodels and selecting the best among them, for example according to the Akaike Information Criterion (AIC) proposed by Akaike (1974) or the Bayesian Information Criterion (BIC) proposed by Schwarz (1978), leads to a complexity growing exponentially with the number of variables involved and renders the methods totally unpractical. To paraphrase the title of a paper by Sala-I-Martin (1997): 'You cannot just run two million regressions!' (and, incidentally, two million would not even suffice for $p \geq 22$).

As concerns asymptotic and consistency results, the settings have to go beyond the classical scheme of keeping the number of parameters p constant (and usually small), while letting the number of observations n of the dependent variable tend to infinity. In high-dimensional situations, both n and p may tend to infinity, while assuming or not some relationship between their respective growth rates. The theory is more subtle in this case and is still developing. This question has been studied for principal component regression for time series under a factor model assumption. Results in this line for the case of penalized regression, and in particular of Ridge regression, have been derived by De Mol et al. (2008, 2015). The first paper also contains an empirical part where predictive accuracy of PCR, Ridge and Lasso regression is evaluated based on a dataset of about 100 time series. It is shown that all three methods perform similarly and that results of Lasso are uninformative when used for applications

where, as is typically the case for macroeconomics, data are cross-correlated. Moreover, in that case Lasso is unstable in selection.

14.3 Policy Analysis and Causal Inference

In the actual big data activity sphere, in parallel with the developments of powerful machine-learning techniques, the emphasis is on predictive rather than causal models. As we shall further discuss in the next section, successful predictive algorithms are rapidly developing in response to the increasing demand coming from all kinds of applications. These algorithms convert large amounts of unstructured data into predictive scores in an automatic way and often in real time.

Whether this trend is desirable may be a matter of debate but it is clear that it implies a significant shift from the single-covariate causal-effect framework that has dominated much empirical research, especially in microeconomics. Being nonstructural, predictive models are subject to the Lucas critique (Lucas, 1976) and their success should not obscure the fact that many economic applications are about inference on a causal effect. In microeconomics, for example, a successful literature has developed methods to assess the effectiveness of a given policy or treatment.

In the case where the intervention is binary in nature, we define a binary variable, W, equal to unity for the treatment group and zero for the control group. We typically have in mind a counterfactual setting, where it makes sense to think of potential outcomes for each unit in the control and treated states. These outcomes are often denoted $Y(0)$ and $Y(1)$, and then we observe the treatment status, W, and the outcome under the corresponding treatment status, $Y = (1 - W)Y(0) + WY(1) = Y(0) + W\,[Y(1) - Y(0)]$. For unit i in the population, the treatment effect is $Y_i(1) - Y_i(0)$ – which is not observed. Instead, attention typically centres on the average treatment effect, $\tau = \mathrm{E}\,[Y(1) - Y(0)]$, or the average over an interesting subpopulation (such as those actually receiving the treatment). A special case is when the treatment effect is constant, in which case we can write $Y = \tau W + Y(0)$, and the $Y(0)$ plays the role of the unobserved factors affecting Y.

The potential outcomes setting can be extended to cases where the policy variable, W, is not binary. If the policy effect is constant across units and across levels of the treatment, we can write a simple regression equation

$$Y = \tau W + R\,, \tag{14.7}$$

where Y, W and R are random variables and τ is a scalar coefficient of interest. We (eventually) observe data on Y and W. The variable R includes unobserved factors – $Y(0)$ in the simplest setting – affecting Y.

In medicine and the experimental sciences, truly randomized experiments can be carried out, which means the treatment level W can be made independent of R. For example, when W is binary, we can randomly assign individuals into the treatment and control groups. In such cases, (14.7) can be estimated using simple regression, which delivers an unbiased and consistent estimator of τ. In economics, random assignment is much less common, and in general one has access only to so-called observational – not experimental – data. Hence, several strategies, when randomized assignment is not available have been developed. Here we review a few of those strategies, highlighting how high-dimensional regression methods can be applied to estimating causal effects. Good pointers to part of the relevant work in this field are the review papers by Belloni et al. (2013, 2014a).

A traditional and still commonly used method to handle nonrandom treatment assignment is regression adjustment, where one assumes the availability of covariates that render the policy assignment appropriately 'exogenous'. Let X be a $1 \times p$ vector of covariates. Then, if X is thought to predict both Y and the treatment assignment, W, we can 'control for' X in a multiple regression analysis. This leads to a linear model,

$$Y = \tau W + X\beta + U, \tag{14.8}$$

where now $R = X\beta + U$ and, if the elements of X suitably control for the nonrandom assignment, the treatment and covariates satisfy the exogeneity conditions

$$\mathrm{E}[WU] = 0, \ \mathrm{E}[XU] = 0. \tag{14.9}$$

If p, the number of control variables, is large and the $p \times 1$ vector β is sparse, the model can be estimated by means of a Lasso regression, as described in the previous section. However, one has to know in advance the right vector X such as, under the usual exogeneity conditions (14.9), there are no more confounding variables and one recovers the marginal effect τ, holding fixed everything else.

One can relax the linearity assumption in X and just assume

$$\mathrm{E}[R|W, X] = \mathrm{E}[R|X], \tag{14.10}$$

which yields

$$Y = \tau W + g(X) + U, \tag{14.11}$$

where $g(X) = \mathrm{E}[R|X]$ and $U = R - \mathrm{E}[R|W, X]$ is such that $\mathrm{E}[U|W, X] = 0$. Model (14.11) is a so-called partially linear model and g is generally a nonparametric function. Belloni et al. (2014b) use Lasso-type methods to nonparametrically partial out X from both Y and W. They approximate the mean functions $\mathrm{E}[Y|X]$ and $\mathrm{E}[W|X]$ using functions of the form $\sum_{j=1}^{p} \beta_j \phi_j(X)$, for a large dictionary of functions $(\phi_j)_{j=1}^{p}$, and build a confidence interval around τ. This

method is particularly appealing as it does not require one to chose a bandwidth to estimate the nonparametric conditional mean functions. If the approximations $\sum_{j=1}^{p} \beta_j \phi_j(X)$ are sparse, then the method selects the significant $\phi_j(X)$ for each of $\mathrm{E}[Y|X]$ and $\mathrm{E}[W|X]$. As shown in Belloni et al. (2014a), using the union of the functions selected from the methods in a standard regression analysis with Y as the response variable and W as the other regressor, the usual heteroscedasticity-robust standard error produces valid t statistics and confidence intervals. It should be emphasized that, while the approach works well for selecting functions of X that appear in the conditional mean, it does not select the variables X such as (14.10) holds; the researcher is assumed to have already selected the appropriate controls.

When (14.9) or (14.10) do not hold, we can rely on instrumental variables, namely, assume to have at our disposal a vector of random variables Z, called instrumental variables, such as in (14.7),

$$\mathrm{Cov}[Z, R] = 0. \tag{14.12}$$

This yields the relation

$$\mathrm{Cov}[Z, Y] = \tau \mathrm{Cov}[Z, W]. \tag{14.13}$$

If Z is a scalar, (14.13) identifies τ when $\mathrm{Cov}[Z, W] \neq 0$. When we have more than one instrumental variable for W, two stage least squares (2SLS) is a common estimation approach. However, 2SLS only uses the linear projection of W on Z in forming instruments. If we strengthen the exogeneity requirement to $\mathrm{E}[R|Z] = 0$, 2SLS is asymptotically inefficient if $\mathrm{E}[W|Z]$ is nonlinear or $\mathrm{Var}[R|Z]$ is not constant. If we assume homoscedasticity in (14.7), that is, $\mathrm{Var}[R|Z] = \mathrm{Var}[R]$, then the optimal instrument for W is given by $\mathrm{E}[W|Z]$, the best mean square predictor of W. Belloni et al. (2012) propose to use Lasso-type methods to estimate the regression function $\mathrm{E}[W|Z]$ using a large dictionary of approximating functions, and they show how to conduct inference on τ using a heteroscedastic-robust standard error.

Gautier and Tsybakov (2011) propose an instrumental variables method to make inference for high-dimensional structural equations of the form

$$Y = X\beta + U \tag{14.14}$$

where the dimension of X is large (and may include exogenous and endogenous variables). This occurs, for example, in large demand systems, or when treatment variables are interacted with (exogenous) group dummies. In the latter case, the policy might have an effect on only certain groups, and the policymaker would like to determine for which group the policy has an effect. The instrumental variables literature has identified various problems: (i) the instrumental variable candidates, Z, might not be exogenous; (ii) the instrumental variables can be 'weak' and estimating in a first-stage a reduced form equation

can yield multimodal and non-normal distributions of the parameter estimates, even with very large sample size, so that asymptotic theory is not reliable; (iii) in the presence of many instrumental variables, estimating in a first-stage a reduced form equation can give rise to a large bias. Gautier and Tsybakov (2011) rely on a new method which is robust to (ii) and (iii) in order to treat the more challenging case of a high-dimensional structural equation. Confidence sets can be obtained for arbitrary weak and numerous instrumental variables, whether or not the condition $\mathrm{Cov}[Z, U] = 0$ gives rise to a unique β. Therefore, it is also possible to handle the case where the dimension of Z is smaller than the dimension of X, which can yield the identification of β under sparsity of the structural Equation (14.14) or other shape restrictions. To deal with the possibility of (i), a high-dimensional extension of the Sargan and Hansen method is developed.

There is much interest in the literature on heterogeneous treatment effects, and variable selection methods can also be applied in such cases. For example, variable selection methods can be applied to estimating the propensity score when the treatment variable takes on a small number of levels. Moreover, variable selection methods can be applied to both propensity score estimation and regression function estimation to obtain so-called *doubly robust* estimators. Unlike the linear, additive Equation (14.11), methods that weigh by the inverse of the propensity score allow for heterogeneous treatment effects. See, for example, the papers by Farrell (2015) and Athey and Imbens (2015).

Besides these high-dimensional problems, let us mention another important issue which arises in connection with the availability of big datasets, namely, to determine whether the accumulation of data, say, over an entire population, affects the precision of estimates. Abadie et al. (2014) analyse how to compute uncertainty in empirical situations where the sample is the entire population and where the regression function is intended to capture causal effects. Other contributions on *Causal Inference* in a big data setting use machine-learning methods. A recent example is the work by Athey and Imbens (2015). There are many open challenges in this area, pointers to recent progress are available from the site of the Sackler Colloquium on 'Drawing Causal Inference from Big Data', organized in March 2015 at the US National Academy of Science in Washington.[3]

The previous discussion has focused on cross-sectional data, but empirical researchers attempting to estimate causal effects often rely on panel data that exploit changes in policies over time. An important component of panel data models is allowing for time-constant, unobserved heterogeneity. Belloni et al. (2014a) propose first differencing a linear unobserved effects equation to remove additive heterogeneity, and then using variable selection methods, such as Lasso, to allow for correlation between unobserved heterogeneous trends and unknown functions of observed covariates – including the policy variable or

variables being studied. The approach seems promising. So far, such methods have been applied to linear models with relatively few sources of heterogeneity.

14.4 Prediction

Despite recent advances in identification and causality in big data settings, which we have just reviewed, it is fair to say that the literature in the field is mainly focused on prediction. Using the same notation as above, the problem consists in computing the conditional expectation

$$\mathrm{E}(Y|W, X)\,. \tag{14.15}$$

Forecasting has a long tradition in economics, especially in macroeconomics. Indeed, many economists in the private sector and policy institutions are employed for this task. In forecasting, robustness is typically tested in out-of-sample validation studies, a perspective typically ignored in empirical microeconomics. For desirable out-of-sample performance, models must respect the principle of parsimony (i.e., contain a rather small number of free parameters) to avoid overfitting. However, the curse of dimensionality problem naturally arises from lags, nonlinearities, and the presence of many potentially relevant predictors.

The recent literature has suggested methods to deal with the curse of dimensionality issue in dynamic models. Here we should mention dynamic factor models cited earlier and, more recently, large Bayesian vector autoregressive models. Following the work of De Mol et al. (2008), Banbura et al. (2010) have shown empirically how to set priors to estimate a vector autoregressive model with large datasets. The idea is to set the degree of 'shrinkage' in relation to the dimension of the data. Intuitively this implies to set priors so as to avoid overfitting, but still let the data be informative. Giannone et al. (2015) have developed a formal procedure to conduct inference for the degree of shrinkage. These models have many applications in economics beyond pure forecasting and can be used to design counterfactuals for policy analysis and identification of exogenous shocks and dynamic propagation mechanisms. Large data allow to better identify exogenous shocks since they can accommodate for realistic assumptions on agents' information set (for an analysis on this point, see Forni et al. (2009)).

One very successful application of the large models described above (if measured by impact on modelling in policy institutions and the financial industry) has been '*now-casting*'. Now-casting is the forecast of the present (present quarter or present month) based on data available at different frequencies (daily, weekly, monthly and quarterly). A now-cast produces a sequence of updates of predictions in relation to the publication calendar of the data. This allows to exploit the timeliness of certain data releases to obtain an early estimate of those

series which are published with a delay with respect to the reference quarter such as GDP or the reference month such as employment (see Giannone et al., 2008 and subsequent literature). Empirical results show that exploiting survey data, which are published earlier than hard data, allows to obtain an accurate early estimate at the beginning of the quarter and, as new data are released through time, the estimates become more accurate (see Banbura et al., 2013 for a review of the literature). In principle, nonstandard data such as Google queries or twitters, due to their timeliness, could be exploited in this context. However, once the details of the problem (mixed frequency, nonsynchronous data releases) are appropriately modelled and relevant timely indicators considered, there is no evidence that Google indexes used successfully in a simpler setup by Choi and Varian (2012) and Scott and Varian (2014) have any additional predictive value (see Li, 2016), but more research is needed on this topic.

It has to be noted that most of the applied work on the methods mentioned have concerned traditional time series (macroeconomic variables, possibly disaggregated by sectors or regions, financial variables and surveys) and rarely with dimension above 150. Empirical results show that, in general, forecasts of macroeconomic variables based on datasets of medium dimension (of the order of 20) are not outperformed by forecasts based on 100 or more variables although the dimension helps especially in now-casting where successful results rely on the use of timely information. Moreover, as mentioned in Section 14.2, Lasso regressions provide unstable variable selection due to the near-collinear feature of macroeconomic data. Important empirical issues are also related to robustness with respect to variable transformation such as deseasonalization or detrending as well as nonlinearity. Potentially, machine-learning type of techniques could be useful in this setup but this is open to future research. The literature is at too early stage to provide a definitive answer on the potentials of new data and new methods in this context but it is our view that any successful applications have to incorporate the detailed micro-structure of the problem. In now-casting, for example, this implies taking care of mixed frequency, the nonsynchronicity of releases and other details.

In microeconometrics the emphasis on predictions and out-of-sample is newer than in macro but studies using big data are more numerous. Predictions based on a large cross-section of data have been successfully obtained for various problems. Examples can be found in papers by Varian (2014), by Einav and Levin (2014) and by Kleinberg et al. (2015b), as well as in the references therein. The last paper discusses a problem of health economics, namely the prediction of whether replacement surgery for patients with osteoarthritis will be beneficial for a given patient, based on more than 3000 variables recorded for about 100,000 patients. Another policy decision based on prediction is studied by Kleinberg et al. (2015a), who show that machine-learning algorithms can be

more efficient than a judge in deciding who has to be released or go to jail while waiting for trial because of the danger of committing a crime in the meanwhile. Another application would be to predict the risk of unemployment for a given individual based on a detailed personal profile.

It should be remarked that machine-learning algorithms present several advantages: they focus on a best-fit function for prediction, possibly handling very rich functional forms, and have built-in safeguards against overfitting so that they can handle more variables than data points. Moreover, they do not require too many assumptions about the data generating process as it is the case in classical econometrics. We should be aware, however, that precisely because of their great generality and versatility, they may not be optimally tailored for the specificities of a given problem.

Another trend is to make use not only of more data, but also of new types of data. Many types of data are nowadays passively collected and are largely unexploited, such as those provided by social networks, scanner data, credit card records, web search queries, electronic medical records, insurance claim data, etc. They could complement more traditional and actively collected data or even be a substitute for them. The mining of language data, such as online customer reviews, is also a challenge and can be used for so-called '*sentiment analysis*' (see e.g., Pang et al., 2002).

Returning to the issue of causality discussed in the previous section, it should be noted that prediction algorithms also provide new ways to test theories. Indeed, we can see how well we can predict the output Y with all variables but X and/or how much the inclusion of a given variable (or a group of variables) X helps improving the prediction. We should be cautious, however, in drawing conclusions: the fact that a variable is not among the best predictors does not necessarily mean that it is not 'important'. For example, when Varian (2014) discusses differences in race in mortgage applications, saying that race is not among the best predictors is not the same as saying that evidence of discrimination does not exist.

In addition, the completeness of a given theory could be tested by confronting its prediction abilities against an atheoretical benchmark provided by machine learning.

14.5 Computational Issues

The collection and analysis of bigger and bigger datasets obviously pose methodological as well as computational challenges. Nevertheless, since there has been at the same time a tremendous increase in computing capabilities, researchers can handle larger and larger datasets using standard software and desktop computers. For example, up to the late 90s maximum-likelihood estimation of dynamic factor models could be performed only with a small set

of variables (Stock and Watson, 1989), while recent research has shown how these models can be easily estimated in a high-dimensional context (Doz et al., 2012, Jungbacker and Koopman, 2015). In parallel with this increase in computing power, significant progress has been made in the development of fast and reliable numerical algorithms which scale well with dimension. In particular, considerable research effort has been dedicated to improving the speed and performance of algorithms for Lasso regression.

In many situations, however, computational capability still represents an important constraint on our ability to handle and analyse big datasets. Methods that can handle thousands of variables may become inappropriate when moving to millions of variables. Moreover, some procedures can be particularly demanding in terms of computational complexity when applied to more than a handful of data. This is the case, for example, for complex latent variable models for which closed-form solutions are not available. In this context there is a demand for extra computing power. Unfortunately, the growth rate in computational capability of integrated circuits (CPU microchips) seems to be slowing down. However, thanks to technological progress driven by the video-game industry, new and fast growing computational power is coming from so-called graphics processing units (GPU), which allow for parallel computation and are easy to program. The general idea is that it is often possible to divide large problems into smaller independent tasks, which are then carried out simultaneously.

Splitting large problems into small ones is particularly natural in simulation-based Bayesian methods, which have recently attracted growing interest (see e.g., Hoogerheide et al., 2009, Lee et al., 2010, Durham and Geweke, 2013). In Bayesian methods, the reduction in dimensionality is made by assuming prior distributions for the unknown parameters to infer and, whereas the computation of the so-called MAP (Maximum a Posteriori) estimator requires solving an optimization problem, the computation of conditional means and covariances only requires integration, but in a high-dimensional space. For this task stochastic simulation methods and artificially generated random variables are used. Since the early days of Monte Carlo methods, there has been substantial development of new more sophisticated Sequential Monte Carlo and Particle Filter methods, allowing us to deal with complex posterior distributions and more flexible econometric models.

Examples of successful applications of simulation-based Bayesian methods are reported by Billio et al. (2013a,b) and Casarin et al. (2013, 2015). The paper by Casarin et al. (2015) deals with the problem of conducting inference on latent time-varying weights used to combine a large set of predictive densities for 3712 individual stock prices, quoted in NYSE and NASDAQ, using 2034 daily observations from 18 March 2002 to 31 December 2009. The authors

find substantial forecast and economic gains and also document improvement in computation time achieved by using parallel computing compared to traditional sequential computation. Another application to nowcasting is discussed by Aastveit et al. (2014), who show that a combined density now-cast model works particularly well in a situation of early data releases with relatively large data uncertainty and model incompleteness. Empirical results, based on US real-time data of 120 leading indicators, suggest that combined density nowcasting gives more accurate density now-casts of US GDP growth than a model selection strategy and other combination strategies throughout the quarter, with relatively large gains for the first two months of the quarter. The model also provides informative signals on model incompleteness during recent recessions and, by focusing on the tails, delivers probabilities of negative growth, that provide good signals for calling recessions and ending economic slumps in real time.

14.6 Conclusions

Data are essential for research and policy. Definitely there is a trend towards empirical economics, and from this perspective, the advent of big data offers an extraordinary opportunity to take advantage of the availability of unprecedented amounts of data, as well as of new types of data, provided that there is easy access to them, in particular for academic research.

In this chapter, we have focused on some methodological aspects of the analysis of large datasets. We have argued that many of the issues raised by big data are not entirely new and have their roots in ideas and work over the past decades. On the applied side, applications with truly big data are still rare in economics although in recent years more research has been devoted to the use of relatively large but traditional datasets.

While in many problems the focus is shifting from identification towards prediction, which is a more '*revolutionary trend*', causality is still considered important and this duality is a matter for interesting debates in econometrics.

As concerns algorithmic and computational issues, the field of '*machine learning*', a popular heading covering very different topics, is and will remain helpful in providing efficient methods for mining large datasets. However, we should be careful rather than blindly import methodologies from other fields, since economic data structures have their own specificities and need appropriately designed research tools.

Undoubtedly, this research area calls for a lot of new, exciting and perhaps unexpected developments within and outside the framework sketched here, and if the datasets are big, the challenges ahead are even bigger, in optimally exploiting the information they contain.

Notes

1. This chapter is based on the presentations given by the authors at the COEURE workshop on 'Developments in Data and Methods for Economic Research' held in Brussels in July 2015. The presentations took place in two sessions of the workshop: 'Big Data: Definition, challenges and opportunities', chaired by Christine De Mol, and 'How will Big Data change econometrics?' chaired by Domenico Giannone. Christine De Mol coordinated and integrated the authors' presentations to the chapter.
2. Since this implies the existence of a nontrivial null-space for $\mathbf{X}'\mathbf{X}$, with at least $p - n$ zero eigenvalues.
3. See http://www.nasonline.org/programs/sackler-colloquia/completed_colloquia/Big-data.html.

References

Aastveit, K. A., Ravazzolo, F., and van Dijk, H. K. 2014 (Dec). *Combined Density Nowcasting in an Uncertain Economic Environment*. Tinbergen Institute Discussion Papers 14-152/III. Tinbergen Institute. Accepted for publication in *Journal of Business Economics and Statistics*. http://tandfonline.com/doi/abs/10.1080/07350015.2015.1137760

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. 2014 (July). *Finite Population Causal Standard Errors*. Working Paper 20325. National Bureau of Economic Research.

Akaike, H. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **AC-19**(6), 716–723.

Athey, S., and Imbens, G. W. 2015. Machine Learning Methods for Estimating Heterogeneous Causal Effects. *ArXiv e-prints*, Apr.

Banbura, M., Giannone, D., and Reichlin, L. 2010. Large Bayesian Vector Auto Regressions. *Journal of Applied Econometrics*, **25**(1), 71–92.

Banbura, M., Giannone, D., Modugno, M., and Reichlin, L. 2013. Now-Casting and the Real-Time Data-Flow. In: Elliott, G., and Timmermann, A. (eds), *Handbook of Economic Forecasting, Volume 2, Part A*. Elsevier-North Holland, pp. 195–237.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. 2012. Sparse Models and Methods for Instrumental Regression, with an Application to Eminent Domain. *Econometrica*, **80**(6), 2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. 2013. Inference for High-Dimensional Sparse Econometric Modelling. In: Acemoglu, D., Arellano, M., and Dekel, E. (eds), *Advances in Economics and Econometrics, Tenth World Congress of the Econometric Society*, vol. 3. Cambridge University Press, pp. 245–295.

Belloni, A., Chernozhukov, V., and Hansen, C. 2014a. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, **28**(2), 29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. 2014b. Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, **81**(2), 608–650.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. 2013a. *Interactions between Eurozone and US Booms and Busts: A Bayesian Panel Markov-switching VAR Model*. Working Papers 2013:17. Department of Economics, University of Venice 'Ca' Foscari'.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. 2013b. Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, **177**(2), 213–232.

Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. K. 2013. *Parallel Sequential Monte Carlo for Efficient Density Combination: The DeCo Matlab Toolbox*. Working Papers 2013:08. Department of Economics, University of Venice 'Ca' Foscari'.

Casarin, R., Grassi, S., Ravazzolo, F., and van Dijk, H. K. 2015 (July). *Dynamic Predictive Density Combinations for Large Data Sets in Economics and Finance*. Tinbergen Institute Discussion Papers 15-084/III. Tinbergen Institute.

Chamberlain, G., and Rothschild, M. 1983. Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica*, **51**(5), 1281–304.

Choi, H., and Varian, H. 2012. Predicting the Present with Google Trends. *The Economic Record*, **88**(s1), 2–9.

De Mol, C., Giannone, D., and Reichlin, L. 2008. Forecasting Using a Large Number of Predictors: Is Bayesian Shrinkage a Valid Alternative to Principal Components? *Journal of Econometrics*, **146**(2), 318–328.

De Mol, C., Giannone, D., and Reichlin, L. 2015. Forecasting with High-dimensional Time Series. Oberwolfach Reports, No. 38/2015. European Mathematical Society.

Diebold, F. X. 2003. Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting: A Discussion of the Papers by Reichlin and Watson. In: Dewatripont, M., Hansen, L., and Turnovsky, S. (eds), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, vol. 3. Cambridge University Press, pp. 115–122.

Diebold, F. X. 2012 (Sep). *On the Origin(s) and Development of the Term 'Big Data'.* PIER Working Paper Archive 12-037. Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Doz, C., Giannone, D., and Reichlin, L. 2012. A Quasi-Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics*, **94**(4), 1014–1024.

Durham, G., and Geweke, J. 2013 (Apr.). *Adaptive Sequential Posterior Simulators for Massively Parallel Computing Environments*. Working Paper Series 9. Economics Discipline Group, UTS Business School, University of Technology, Sydney.

Einav, L., and Levin, J. 2014. Economics in the Age of Big Data. *Science*, **346**(6210).

Farrell, M. H. 2015. Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations. *Journal of Econometrics*, **189**(1), 1 – 23.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. 2000. The Generalized Dynamic Factor Model: Identification and Estimation. *Review of Economics and Statistics*, **82**, 540–554.

Forni, M., Giannone, D., Lippi, M., and Reichlin, L. 2009. Opening the Black Box: Structural Factor Models with Large Cross Sections. *Econometric Theory*, **25**(10), 1319–1347.

Frank, I. E., and Friedman, J. H. 1993. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**(2), pp. 109–135.

Gautier, E., and Tsybakov, A. 2011. High-dimensional Instrumental Variables Regression and Confidence Sets. *ArXiv e-prints*, May.

Giannone, D., Reichlin, L., and Small, D. 2008. Nowcasting: The Real-time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, **55**(4), 665–676.

Giannone, D., Lenza, M., and Primiceri, G. E. 2015. Prior Selection for Vector Autoregressions. *Review of Economics and Statistics*, **97**(2), 436–451.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning*. 2 edn. Springer-Verlag.

Hoerl, A. E., and Kennard, R. W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**(1), pp. 55–67.

Hoogerheide, L. F., van Dijk, H. K., and van Oest, R. D. 2009. Simulation-Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances. In: *Handbook of Computational Econometrics*. John Wiley and Sons, Ltd, pp. 215–280.

Jungbacker, B., and Koopman, S. J. 2015. Likelihood-based Dynamic Factor Analysis for Measurement and Forecasting. *Econometrics Journal*, **18**, C1–C21.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. 2015a. *Human Decisions and Machine Predictions*. Harvard mimeo.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. 2015b. Prediction Policy Problems. *American Economic Review*, **105**(5), 491–495.

Lee, A., Yau, C., Giles, M. B., Doucet, A., and Holmes, C. C. 2010. On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods. *Journal of Computational and Graphical Statistics*, **19**(4), 769–789.

Li, Xinyuan. 2016. *Nowcasting with Big Data: Is Google Useful in the Presence of Other Information?* London Business School Mimeo.

Lucas, R. J. 1976. Econometric Policy Evaluation: A Critique. *Carnegie-Rochester Conference Series on Public Policy*, **1**(1), 19–46.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In: *Proceedings of EMNLP*, pp. 79–86.

Reichlin, L. 2003. Factor Models in Large Cross Sections of Time Series. In: Dewatripont, M., Hansen, L., and Turnovsky, S. (eds), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, vol. 3. Cambridge University Press, pp. 47–86.

Sala-I-Martin, X. X. 1997. I Just Ran Two Million Regressions. *The American Economic Review*, **87**(2), pp. 178–183.

Schwarz, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, **6**(2), 461–464.

Scott, S. L., and Varian, H. R. 2014. Predicting the Present with Bayesian Structural Time Series. *International Journal of Mathematical Modelling and Numerical Optimisation*, **5**(1/2), 4–23.

Stein, C. 1956. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, pp. 197–206.

Stock, J. H., and Watson, M. W. 1989. New Indexes of Coincident and Leading Economic Indicators. In: *NBER Macroeconomics Annual 1989, Volume 4*. NBER Chapters. National Bureau of Economic Research, Inc, pp. 351–409.

Stock, J. H., and Watson, M. W. 2002. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, **97**, 147–162.

Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Varian, H. R. 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, **28**(2), 3–28.

Watson, M. W. 2003. Macroeconomic Forecasting Using Many Predictors. In: Dewatripont, M., Hansen, L., and Turnovsky, S. (eds), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress of the Econometric Society*, vol. 3. Cambridge University Press, pp. 87–115.