Essex Finance Centre
Working Paper Series

Working Paper No36: 07-2018

"Machine Learning Macroeconometrics: A Primer"

"Dimitris Korobilis"

Essex Business School, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ
Web site: http://www.essex.ac.uk/ebs/

Machine Learning Macroeconometrics: A Primer

Dimitris Korobilis
University of Essex

March 21, 2018

Abstract

This Chapter reviews econometric methods that can be used in order to deal with the challenges of inference in high-dimensional empirical macro models with possibly "more parameters than observations". These methods broadly include machine learning algorithms for Big Data, but also more traditional estimation algorithms for data with a short span of observations relative to the number of explanatory variables. While building mainly on a univariate linear regression setting, I show how machine learning ideas can be generalized to classes of models that are interesting to applied macroeconomists, such as time-varying parameter models and vector autoregressions.

Contents

1 Introduction and motivation

High-dimensional inference and machine learning methods currently shape research in various fields of human knowledge, but they are not as popular yet in mainstream economic research and decision-making. There are several reasons for this observation. Macroeconomics, in particular, is dominated by aggregate data sets that cannot compete in size with data sets that, for example, engineers or astronomers collect. Despite the fact that theoretical and empirical macro models have substantially increased in size since the global crisis of 2007-2009 (mainly to accommodate linkages with the financial sector and global interdependencies), such models are far from being considered models for Big Data. Finally, machine learning methods are focused on prediction, and may not be so reliable for parameter estimation and structural analysis (Mullainathan and Spiess, 2017). While many macroeconomic results rely on accurate out-of-sample forecasts from agents, a large part of traditional structural econometric inference focuses on identification of primitive underlying shocks and reliable parameter estimation of coefficients and elasticities in-sample. Such tasks cannot be supported by many existing machine learning algorithms which are so flexible, to the extend that it is hard for the economist to interpret underlying economic relationships in data, or for the econometrician to be able to prove asymptotic consistency of parameter estimates.

So is there a need for a Chapter in high-dimensional and machine learning inference in applied macroeconomics? Our current modeling trends in macroeconometrics are heavily defined by lessons learned during the first attempts to specify large-scale structural econometric models. The Cowles Commission for Research in Economics was established in 1932 and it consisted of economists like Tjalling Koopmans and Trygve Haavelmo who actively tried to link economic theory with statistics using large systems of equations (Christ, 1994). But unlike modern machine learning and Big Data methods that aim to learn from the data, their inference methods were characterized by numerous identification restrictions based on economic theory, or simply driven by the need to make estimation feasible. Unless such subjective restrictions are learned by the data, it is highly improbable that the data will support a large number of them. The "Minnesota revolution" of the 1970s consisting of economists such as Sargent, Sims, Geweke, Neftçi, Litterman and Doan -all associated with the University of Minnesota and Minneapolis Fed, hence the name- suggested statistical shrinkage and selection as an alternative to the unrealistic restrictions imposed by the Cowles Commission models. A major contribution was the famous Minnesota prior (Litterman, 1979), which is an empirical Bayes procedure that drives research in vector autoregressions (VARs) forty years later (see Koop and Korobilis, 2010, for an introduction to Bayesian VARs). Their second response to the unrealistic restrictions of large-scale macroeconometric models lies in the introduction of (dynamic) factor methods by Sargent and Sims (1977) in order to introduce time series analysis "without pretending to have too much a priori economic theory"; see also Geweke (1977). Almost 40 years later there are numerous studies relying on the traditional Minnesota prior or general factor methods; see the Chapter by Miranda-Agrippino and Ricco (2018) in this Encyclopedia.

The purpose of this Chapter is to review existing methods in econometrics and statistics that facilitate inference with large information sets, and at the same time examine new machine learning ideas that might prove to be useful to the applied macroeconomist. The primary focus is on the case of "fat" data, that is, data with more predictors/variables than observations. Given that macroeconomic time series (especially those outside the US) typically have short time spans and are measured at low frequencies (e.g. monthly or quarterly), many existing empirical models fall into this category. At the same time, as new disaggregated data sets become increasingly available[1], estimation methods must be *scalable* that is, be able to scale computationally to very large dimensions without causing a "computational bottleneck".

The main aim and challenge is to filter out and review only those machine learning methods that general macroeconomists will find accessible and readily applicable to various settings. For that reason, presentation of algorithms and methods builds exclusively on the familiar univariate regression model. Additionally, rather than focusing only on recent developments in the machine learning literature, I attempt to review existing methods in econometrics for high-dimensional data and assess how such methods have or can be extended using machine learning ideas. In one sense, this review is restrictive as it does not cover topics that gain increasing popularity, such as textual analysis (Hansen, McMahon and Prat, forthcoming), or algorithms for high-frequency financial data (Aït-Sahalia and Xiu, 2017). However, after analyzing the

[1] For example, Jiménez et al. (2014) use in their analysis the 32 million loan transactions obtained from the the Credit Register of the Banco de España.

simple regression model I subsequently show that many of the methods and algorithms introduced in univariate regression can easily be generalized to several complicated settings. The main idea is that of designing efficient estimation algorithms by casting complex high-dimensional problems into a simpler univariate linear regression form. I provide examples from time-varying parameter regressions with possibly many predictors, and large vector autoregressions. Thus, instead of listing advanced machine learning algorithms using computing science jargon that many economists won't find appealing, this Chapter presents a more accessible approach that tries to build on benchmark ideas from univariate regression and least squares estimation.

The next Section describes the simple linear regression problem with many predictors, and suggests four classes of methodologies that can be applied to deal with this problem. Section 3 describes ways an applied macroeconomist can simplify more complex high-dimensional models in order to bring them in a form that broadly resembles the simple linear regression model. Section 4 concludes this chapter.

2 The linear regression problem

The starting point is a generic time series[2] univariate regression problem for the scalar dependent variable y_t and for the $1 \times p$ vector of exogenous predictors x_t, observed for periods $i = 1, ..., T$. This can be written in the usual form

$$y_t = x_t \beta + \varepsilon_t, \tag{1}$$

where β is a $p \times 1$ vector of regression coefficients and $\varepsilon_t \sim N\left(0, \sigma^2\right)$ with σ^2 a scalar regression variance. The regressors x_t may include intercepts, dummies, exogenous predictors, lags of the dependent variable etc. This is the simplest benchmark for an economist, that is straightforward to estimate and communicate to clients and policy-makers. In that sense, it will be helpful if we view any extensions we introduce later in this chapter, as special cases of this basic regression problem.

Stock and Watson (1999, 2002), and others, have shown that it is empirically superior to augment small, traditional macroeconomic models, such as the New

[2] I adopt a time series notation which is appropriate for many problems in macroeconomics and finance, even though for most part of the analysis in this Section our data could be cross-sectional. The time series notation will prove useful in subsequent Sections, when generalizing high-dimensional inference to vector autoregressive and time varying parameter models.

Keynesian Phillips Curve (NKPC), with information contained in many predictors. What is now established as "many predictors" for US data is a set of possibly 100-200 variables available roughly from 1960; see the FRED-MD (and FRED-QD) data sets described in McCracken and Ng (2016). For other countries the number of available predictors might be smaller but time series observations are also typically shorter; see Nicoletti-Altimari (2001) for a key application in Euro-Area inflation using many monetary aggregates as predictors.

As argued in the Introduction, inference with such data sets can hardly be classified as "Big Data analytics", but it can cause headaches of its own. First, consider the case of selecting a handful of predictors out of a larger set of possible predictor variables, e.g. following economic theory. If the true model is of the form

$$y = x_1\beta_1 + x_2\beta_2 + \varepsilon, \tag{2}$$

where $x_i = \left(x'_{i,1}, ..., x'_{i,T}\right)'$ for $i = 1, 2$, but instead we estimate a regression using only predictor x_1 then the OLS estimator b_1 of β_1 is

$$\begin{aligned} b_1 &= (x'_1x_1)^{-1}x'_1y = (x'_1x_1)^{-1}x'_1\left(x_1\beta_1 + x_2\beta_2 + \varepsilon\right) & (3) \\ &= \beta_1 + (x'_1x_1)^{-1}x'_1x_2\beta_2 + (x'_1x_1)^{-1}x'_1x_2\varepsilon, & (4) \end{aligned}$$

and its bias is

$$E(b_1) = \beta_1 + (x'_1x_1)^{-1}x'_1x_2\beta_2, \tag{5}$$

that is, it is equal to β_1 in expectation only if $x'_1x_2 = 0$ (the omitted predictor is uncorrelated with x_1), or if $\beta_2 = 0$ (the omitted predictor was not generating y in the first instance). In any other case the OLS estimate b_1 is biased, a well-known result which is known as *omitted variable bias.* Such biases become significant, the more important variables we fail to include in the regression, especially as many macroeconomic time series are in general heavily correlated. More importantly for macroeconomists, in certain applications (e.g. vector autoregressions) omitted variable bias translates into lack of identification of structural shocks using data.

Second, consider the case of modelling using all available predictor variables. In many such cases (e.g. data measured quarterly) it will hold that $p \approx T$ or even $p > T$, which makes OLS inference numerically unstable or infeasible. Even when the number of observations is large enough to guarantee sufficient degrees of freedom for parameter estimation, using the full set of available predictors is not desirable because it leads to overfitting and overparametrization. If the econometrician estimates a very flexible model with many parameters that fits the data very well in-sample, there is the danger that this model will fail out-of-sample. Intuitively we can think of forecasting a crisis event, such as the Global Recession of 2007-2009, from the point of view of an economist just before its outbreak: information about such an abrupt break does not exist in observations preceding 2007, so an overparametrized model that follows past data closely is condemned to fail more so than a small model that does well on average (but doesn't fit extremely well all past data points). Finally, from an econometric point of view smaller models are easier to maintain and communicate to policy-makers and clients, compared to models with many variables and parameters.

So what kind of tools are at the econometrician's disposal? I review here four methodologies and related algorithms that can be used to deal with the high-dimensional regression problem.

2.1 Extreme Bounds Analysis

Leamer (1983) argued that certain features of an econometric model can be sensitive to the econometric specification used. In a regression setting with many possible predictors, this argument highlights the fact that different combinations of predictors might result in models with comparable explanatory ability for our dependent variable, but with conflicting economic interpretation. So how can we test for specification bias for a scalar predictor of interest x_t^S, when the number of potential explanatory variables, x_t^D, is large? Leamer (1983) suggested to estimate the following $j = 1, ..., M$ models

$$y_t = a_j + x_t^F \beta_{F,j} + x_t^S \beta_{S,j} + x_{j,t}^D \beta_{D,j}, \tag{6}$$

where x_t^F is a small set of variables always included in all models (free variables), x_t^S is the scalar predictor of interest, and $x_{j,t}^D$ denotes a subset of the large set of k variables in x_t^D. For example, if $k = 10$ and (for computational reasons) we want to consider all possible M models we can construct with up to five predictors in $x_{j,t}^D$ then $M = 637$.

We can now define the variable x_t^S to be a robust predictor if i) all estimates $\beta_{S,j}$ are of the same sign, and ii) all estimates of $\beta_{S,j}$ are statistically significant. Using the collection of least squares estimates $\widehat{\beta}_{S,j}$ and their associated standard deviations

$\widehat{\sigma}_{S,j}$, we can compute the extreme bounds of coefficient β_S as

$$\widehat{\beta}_S^{lower} = \min_{j \in 1,M} \left[\widehat{\beta}_{S,j} - q_{1-\alpha/2} \widehat{\sigma}_{S,j} \right], \tag{7}$$

$$\widehat{\beta}_S^{UPPER} = \max_{j \in 1,M} \left[\widehat{\beta}_{S,j} + q_{1-\alpha/2} \widehat{\sigma}_{S,j} \right]. \tag{8}$$

Therefore, $\widehat{\beta}_S^{lower}$ is the smallest lower bound among all M confidence intervals, and $\widehat{\beta}_S^{upper}$ is the largest upper bound of the collection of M confidence intervals, of size $100(1-\alpha)$.

Levine and Renelt (1992) provide a key application of EBA in the problem of finding robust predictors for growth regressions. Benson Durham (2001) applies EBA in order to test 23 anomalies for the cross-section of stock returns. The original extreme bounds analysis (EBA) would require both $\widehat{\beta}_S^{lower}$ and $\widehat{\beta}_S^{upper}$ to be of the same sign, a definition that can hardly be satisfied in practical situations. Granger and Uhlig (1990) introduced a "reasonable" version of EBA with less strict requirements. Sala-i-Martin (1997) introduced a weighted EBA algorithm that assigns some level of confidence to variables of interest, instead of labelling them only as "robust" or "non-robust".

2.2 Model averaging and variable/model selection

Algorithms such as EBA were quite useful in an era when the personal computer and specialized mathematical programming languages[3] were still in their prime. Notice how traditional applications of EBA would consider a large number of potential control variables x_t^D, but each of the M models discussed previously would only consider a small subset of these variables. However, in an era where computing power is so strong and more elaborate algorithms have been discovered since, it is feasible to consider all possible sources of uncertainty regarding the correct model specification. Controlling for model uncertainty can either take the form of model/variable selection or model averaging. Model selection selects the "best" predictors that might have generated our data y, where "best" can either be defined using statistical (e.g. goodness of fit) or economic (e.g. utility) criteria. Model averaging uses the

[3] MATLAB, GAUSS and RATS - three of the traditionally most popular matrix programming languages used by macroeconomists - all launched in the mid-1980s (with RATS based on Chris Sim's Fortran routines developed in the 1970s), that is, around the time that EBA was first developed.

information in all available variables, weighted by their probability of belonging to the "true" model that might have generated our data of interest.

If a researcher is faced with a finite number of predictor variables, say 20 or 30, variable selection is relatively straightforward.[4] Since each predictor is either included or excluded from the best model, then with p predictors the possible number of models is $K = 2^p$. Model 1 is the model with no predictors, Model K is the model with all p predictors, and the remaining $K-2$ models have combinations of two, three, or up to $p-1$ predictors. We can subsequently enumerate and construct all K models, estimate them (e.g. with OLS) and select the best model using some statistical criterion. In this case we have K models of the form

$$y = X^{(i)}\beta^{(i)} + \varepsilon, i = 1, ..., K, \tag{9}$$

where $X^{(i)}$ denotes the ith combination of columns of X. For example, Pesaran and Timmermann (1995) consider the stock return predictability problem, where they want to forecast excess stock returns with a set of nine, primarily macroeconomic, predictors.[5] Even with the computational resources available in early to mid-90s, Pesaran and Timmermann estimate, using simple OLS, all possible $2^9 = 512$ models at their disposal for forecasting excess stock returns. They calculate various statistical (BIC, adjuster R^2) and economic criteria (Sharpe ratio, wealth criterion) in order to select the forecasting model that leads to maximum wealth for an investor who can choose to allocate funds either to shares or bonds.

Instead of doing variable selection, one can also use the same criteria (BIC, R^2, Sharpe ratios) to do model averaging. The idea of model averaging is similar to diversification in portfolio allocation. An investor does not want to choose the best performing stock, as there is huge risk associated with such a stock. For the same reason, the investor does not want to choose an asset with the lowest risk (variance), since via diversification they can achieve much higher returns with almost the same amount of risk. The same arguments motivate model averaging, that is, a researcher should use information all models – good or bad – weighted by their "importance".

[4] That is, from a computational point of view. From an econometric point of view variable selection can be undermined by predictors that are correlated or persistent, parameters subject to breaks etc., in which case it becomes anything but a straightforward problem.

[5] See also the complete subset regression of Elliott, Gargano and Timmermann (2013) for a computationally efficient procedure that fits in the class of shrinkage estimators (rather than variable selection).

The idea is that even badly fitting models might be able to capture certain features of our y that the single best model cannot. In a forecasting example, where we want to predict y_{T+1} using the K combinations of available predictors, the model averaged forecast is defined as

$$y^{MA}_{T+1|T} = \sum_{i=1}^{K} y^{(i)}_{T+1|T} \times w^{(i)}, \tag{10}$$

where $y^{(i)}_{T+1|T}$ and $y^{MA}_{T+1|T}$ is the forecast of period $T+1$ using information at time T for the ith model and the model average, respectively. The quantity $w^{(i)}$ is the weight of the ith model, which following the analysis of Raftery (1995) can simply be calculated using the BIC of each model which is a first-order approximation to the marginal likelihood. Kapetanios, Labhard and Price (2008) apply these ideas[6] and construct weights as

$$w^{(i)} = \frac{exp(BIC^{(i)})}{\sum_{j=1}^{K} exp(BIC^{(j)})}. \tag{11}$$

Such weights are explicitly calculated model probabilities that are easy to interpret, plus they can be converted into variable-specific probabilities. That way, these weights are not only appropriate for model averaging, but can also lead to straightforward variable selection. For example, Barbieri and Berger (2004) show that the "(Bayesian) median probability model", that is the model with those variables that have probabilities larger than 0.5, is optimal for prediction. This convenience in interpretation explains why such a variable selection/averaging procedure is possibly preferable to conventional hypothesis testing. As Raftery (1995) argues, with traditional testing we use significance levels 1% or 5% only because Sir Ronald Fischer was using such values with samples of 30 or 200 observations; see for example Fischer (1925, Chapter 4). Hence there is no scientific reason why $\alpha = 0.05$ is a good choice to test hypotheses. Raftery (1995) notes that as the sample size increases p-values should be judged using smaller significance levels. By examining the full set of possible models interpretation and communication of regression results is easy, as the macroeconomist can explicitly test each predictor against economic theory using easy-to-interpret "probabilities of inclusion of each predictor". In contrast, a p-value is not the probability that the null hypothesis is true, which many times has lead to their misinterpretation in scientific and academic studies; see the American Statistical Association's statement on p-values in Wasserstein and Lazar (2016).

[6] See also Doppelhofer, Miller and Sala-i-Martin (2004).

The above procedure using OLS and BIC is trivial, and is accessible to all applied macroeconomists who are concerned about overfitting issues. However, notice that for $p \gg 30$ it is computationally impossible to enumerate and estimate the total number of models.[7] In such cases, variable selection can only be implemented using computationally efficient algorithms. There are two popular algorithmic strategies of sequentially exploring for, and selecting, good predictors: specific-to-general and general-to-specific. As the name suggests, the specific-to-general approach starts with a small regression model (possibly with few predictors suggested from economic theory) and then sequentially expands the size of the model to accommodate only those predictors that are important/significant. The general-to-specific approach begins with the full, overparametrized model using all available predictors and then sequentially drops unimportant predictors. Most modern statistical and machine learning approaches to variable selection and shrinkage (discussed in the next subsection) use a general-to-specific approach.[8] Such an approach is mostly useful when the amount of predictors is really large (Big Data) and/or economic theory is not available, so it doesn't make sense to begin from a specific small model and then expand the number of predictors. Put informally, the general-to-specific approach implies that a researcher can just "throw" in a regression all available predictors, and then rely on an algorithm to do the sorting of predictor variables into good or bad, according to some metric.

From a Bayesian point of view the variable selection problem is characterized by the use of a clear probabilistic framework (Bayes Theorem), where the researcher assigns prior model weights and uses the information in the data to update those into posterior model probabilities, $w^{(i)}$. At the same time posterior simulation algorithms, such as Markov chain Monte Carlo, make variable selection in high dimensional spaces feasible. There are many ways of implementing Bayesian variable selection, but the default/benchmark choice used when modeling large data sets in e.g. Biology, Engineering, Astronomy, is the "spike and slab" prior. For our regression problem in

[7] If we have just $p = 50$ then the number of possible models, $K = 2^{50}$, is so vast that even if it takes $1/1000$-th of a second to estimate each model by OLS and save its BIC value, it would take a total of 35702 years to estimate all models!

[8] The general-to-specific in the econometrics literature, also called the "LSE approach", dates at least to the work of Sargan (1964). A key example in the econometric literature is the Autometrics approach, see Hendry and Doornik (2014). Autometrics is an algorithm for automatic model discovery, focusing primarily on variable selection, but which also allows to accommodate other modeling/specification features, such as breaks.

equation (1) this prior takes the form

$$\begin{aligned} \beta_i|\gamma_i &\sim (1-\gamma_i)\delta_0(\beta) + \gamma_i N\left(0, \tau_i^2\right), & (12) \\ \gamma_i|\pi &\sim Bernoulli(\pi), \quad i = 1, .., p & (13) \\ \pi &\sim Beta(c_0, d_0). & (14) \end{aligned}$$

In the formula above $\delta_0(\beta)$ denotes the Dirac delta function evaluated at zero. While the parameter of interest is β, we also introduce two more parameters that have their own prior distributions: dummy indicators $\gamma = (\gamma_1, ..., \gamma_p)$ and probabilities π.[9] Conditional on $\gamma_i = 0$, then the prior for β_i is $\delta_0(\beta)$ which is a point mass at zero. In this case the posterior distribution of coefficient β_i will also be a point mass at zero and the ith predictor is removed from the regression. If $\gamma_i = 1$ then the prior for β_i is $N(0, \tau_i^2)$. For a sufficiently large (noninformative) value of τ_i^2 this prior is dominated by the likelihood, meaning the posterior estimate of β_i will be unrestricted and different from zero. The way γ is either zero or one is determined by the information in the data, since γ in equation (13) has its own prior and the data likelihood can update this prior into a well-defined posterior distribution.[10]. Therefore, the vector γ, taking values zero or one for each element i, $i = 1, ..., p$, has the ability to index all possible $K = 2^p$ regression models constructed using our predictor variables. Posterior computation can be implemented using simulation algorithms, such as MCMC; see George and McCulloch (1997). In this case, the Monte Carlo posterior algorithm has the ability to visit stochastically the most probable models, since estimating all possible models is impossible for large p. At each iteration of the MCMC sampler we can save a sample from the posterior distribution of γ which is going to be a vector of zeros and ones. Doing so for several thousand iterations, the posterior mean or median of all samples of γ will be a vector of *posterior inclusion probabilities* for each predictor in our regression.

While a Bayesian setting for variable selection takes advantage of MCMC simulation algorithms that are anyway needed for parameter estimation, the

[9] As explained in the following subsection, this is the case of a *hierarchical prior*.

[10] Regarding π this is the probability of a Bernoulli random variable, hence, many researchers set this to 0.5 which is the value of a Bernoulli experiment using a "fair coin". However, as George and McCulloch (1997) note, $\pi = 0.5$ implies that a-prior our expectation is that half of the elements in β are zero, which is a very informative choice. In this case, π can have its own prior distribution and update this also by the data likelihood.

frequentist approach to simulation-based variable selection can similarly take advantage of a similarly powerful simulation technique, the bootstrap. In a key and highly cited paper for the field of machine learning, Breiman (1996) introduced the bootstrap aggregation, a.k.a. bagging, algorithm. Bagging involves generating a large number of sub-samples of our data and training (pre-testing) the model on each generated sample. Then predictions can be made by averaging the predictions from all generated models. Even though the final outcome is prediction, such algorithms can be used for variable selection by enumerating the number of times a variable is selected in each sample (similarly to what we did above with Bayesian variable selection by averaging the indicators γ). Random forest regression can remarkably improve the performance of the bagging algorithm (Amit and Geman, 1997), and boosting (Breiman, 1998) can be more successful than bagging in reducing variance of estimators (see also next subsection). A good example of a bootstrap method that is designed primarily for variable selection is that of bumping (Tibshirani and Knight, 1999).

There are already several solid examples of model averaging in empirical macroeconomics and finance starting from the early study of Geisel (1973). In general macroeconomic forecasting settings the contributions of Min and Zellner (1993), Koop and Potter (2004), Wright (2008) are based on univariate regression models, while Garratt et al. (2009), George et al. (2008) and Korobilis (2008; 2013c; 2016) focus on variable selection and model averaging in vector autoregressive models. For the problem of stock return predictability in finance Avramov (2002) and Cremers (2002) are two early references. There is no shortage of variable selection and averaging applications in the field of empirical economic growth. Here we can mention, among many others, Chen et al. (2009), Durlauf et al. (2011), Fernández et al. (2001a, 2001b), Masanjala and Papageorgiou (2008) and Moral-Benito (2012) who propose various theoretical, algorithmic or empirical enhancements to standard model averaging procedures. Finally, Koop and Korobilis (2012, 2018) show how machine learning methods can be used to implement model averaging and selection in a dynamic fashion. Koop and Korobilis (2012) use variance discounting methods to estimate regressions with time-varying parameters (see next Section) and use a dynamic version of equation (11) that allows them to estimate time-varying probabilities for each of their predictor in their data set. More recently, Koop and Korobilis (2018) use variational Bayes methods -that are popular in computing science and engineering- to estimate a regression that features a time-varying version of the Spike and Slab prior in (13) and a large number of predictors.

2.3 Penalized and shrinkage estimators

When comparing the performance of different estimators under a square loss function, it is useful to define the mean squared error (MSE) of an estimator $\widehat{\beta}$ of a (assume for simplicity, scalar) parameter β. This is of the form

$$
\begin{aligned}
E\left[\left(\widehat{\beta}-\beta\right)^2\right] &= E\left[\left(\widehat{\beta}-E(\widehat{\beta})+E(\widehat{\beta})-\beta\right)^2\right] && (15)\\
&= E\left[\left(\widehat{\beta}-E(\widehat{\beta})\right)^2\right]+\left(E(\widehat{\beta})-\beta\right)^2 && (16)\\
&= var\left(\widehat{\beta}\right)+bias^2\left(\widehat{\beta}\right). && (17)
\end{aligned}
$$

This formula can be used to understand a well-known tradeoff in econometric estimation, that among bias and variance. Standard textbook analysis postulates that among all unbiased estimators the one with the lowest variance is preferable. When faced with many predictors and a small number of degrees of freedom then unbiased estimators, such as OLS, tend to have very large variance. In such cases a biased estimator that has much lower variance can achieve a lower MSE and is, in general, preferable.

This is a motivating example for introducing the class of shrinkage and penalized regression estimators that can be used to achieve *regularized* estimation. Regularization in machine learning and statistics is the process of introducing additional information in order to prevent overfitting. Such ideas are not new, as they precede our Big data era by several decades. For example, a striking result was presented by Stein (1956) who showed that a certain biased estimator of the mean of a multivariate Normal distribution dominates in terms of MSE the maximum likelihood (i.e. OLS) estimator associated with this problem. Efron and Morris (1975) provided further intuition by showing that the Stein estimator results from an empirical Bayes procedure that places a prior distribution on the parameter of interest, with its prior hyperparameters being functions of the data. For the remainder of the analysis, I make the argument that it is quite helpful and intuitive to view all shrinkage estimators as special cases of Bayesian estimators. To illustrate this point assume the regression model in (1) with only one predictor, a fixed regression variance $\sigma^2 = 1$, and a prior distribution on the regression coefficient of the form $\beta \sim N(0, \tau^2)$. Then the posterior mean/mode of β is of the form

$$\beta^{Bayes} = \left(1/\tau^2 + x'x\right)^{-1} (x'y). \tag{18}$$

When $\tau^2 \to \infty$ then the estimator becomes identical to OLS. However, the interesting case is when τ^2 is finite, that is, informative. Even when $x'x$ is not invertible[11], calculating $(1/\tau^2 + x'x)^{-1}$ in the Bayes estimator is possible for an appropriate choice of τ^2. Using jargon from computing science, τ^2 is then a *regularizer*. In the extreme case where $\tau^2 = 0$, then $\beta^{Bayes} = 0$, which implies full shrinkage of the coefficient towards zero. This can be seen by the fact that in this case the prior for β becomes a point mass at zero, and the prior will dominate any (possibly weak) information that the data likelihood might contain about this coefficient.

Perhaps the most popular example of a shrinkage estimator is that of the lasso (least absolute shrinkage and selection operator) introduced by Tibshirani (1996). The lasso solves the following optimization problem, which we can write in its Langrangian form (that should look more familiar to economists)

$$\min_{\beta \in \Re^p} \left\{ \frac{1}{T} \|y - x\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{19}$$

where λ is a free parameter that determines the amount of regularisation. The notation $\| \bullet \|_p$ denotes the ℓ^p norm, where $p = 2$ is the case of the usual Euclidean norm. Therefore, the first component in the above formula is the usual SSE of the regression and the second component is an ℓ^1 penalty term, hence the term "penalized (regression) estimator".[12] The minimization problem of the lasso can be solved using a wide variety of techniques from convex analysis and optimization theory. Nevertheless, as the dimension of the regression increases (more predictors), the complexity of the optimization problem also increases and

[11] One such case is when many predictors are present and in particular when $p >> T$, in which case this matrix is rank deficient, and the OLS estimator doesn't have a unique solution and will overfit the data; see Bühlmann and Van De Geer (2011, Section 2.2.1).

[12] For other penalties we can get other popular estimators in statistics, for instance, the Euclidean norm gives penalty $\lambda\|\beta\|_2$ which results in the famous ridge regression estimator. Note that other shrinkage estimators might use a different loss function for the residuals; see for example the Dantzig selector of Candes and Tao (2007).

the performance of estimation algorithms (such as the LARS algorithm of Efron et al., 2004) might deteriorate. For that reason an active field of research in machine learning is the development of approximate algorithms that solve the lasso problem[13]. However, many of these fast, approximate machine learning algorithms developed in the compressive sensing and related literatures, are not necessarily universally good algorithms that could be used with persistent and correlated macroeconomic time series data. There are, of course, many cases of fast algorithms with good theoretical guarantees for convergence. For example, Donoho, Maleki and Montanari (2009) develop an iterative algorithm for convex optimization called approximate message passing (AMP), and Mousavi, Maleki and Baraniuk (2017) show asymptotic consistency of this algorithm for lasso problems. Wang et al. (forthcoming) derive the frequentist consistency of the class of popular (in Bayesian statistics) variational Bayes approximate estimators. Nevertheless in some instances, fast approximate algorithms might either be based on simplifying assumptions (e.g. zero or low correlation in predictors) that might make their convergence troublesome and their application to economic data infeasible.

Consistent with our interpretation of Bayesian estimators as shrinkage estimators in equation (18), Tibshirani (1996, Section 5) noted that the solution to the lasso problem is equivalent to a Bayesian regression using a Laplace prior on the regression coefficients we want to penalize. Given that the Laplace distribution can be denoted as a scaled mixture of Normals representation (also known as hierarchical Bayes), Park and Casella (2008) show that Bayesian posterior computation is trivial using the Gibbs sampler, a Markov chain Monte Carlo technique that is familiar to many applied macroeconomists working with VAR and DSGE models. As a matter of fact, all popular extensions of the lasso, such as the elastic net of Zhou and Hastie (2006) and the fused lasso of Tibshirani et al. (2005), have an equivalent hierarchical Bayes representation; see Kyung et al. (2010) and Korobilis (2013b). While Bayesian hierarchical priors are very powerful, sometimes proofs of consistency of resulting posterior distributions are not readily available or they are simply ignored. In any case, an expanding recent literature is being devoted to the derivation of the asymptotic properties of Bayesian hierarchical shrinkage estimators; see for example Bhadra et al. (2016), Ghosh et al. (forthcoming) and Johnson and Rossell (2012).

[13] In the signal processing and compressive sensing jargon, the lasso optimzation problem is called "basis pursuit denoising (BPDN)".

2.4 Factor models and projection methods

The main idea behind factor models is that of finding a lower-dimensional representation of our large vector of data. That is, in typical applications a handful of unobserved factors can summarize the information in large data sets with minimal loss of information. The factors are unobserved exactly because they have to be estimated from the data. This is why this class of models have been a benchmark methodology in psychology, engineering, biology, marketing, economics and other fields. While there are many ways one can introduce factors in a macroeconometric model[14], the standard way to use them in a univariate regression setting is through the following formulation

$$\begin{aligned} x_t &= f_t \lambda + u_t, & (20)\\ y_t &= f_t \gamma + \varepsilon_t, & (21) \end{aligned}$$

where x_t is the $1 \times p$ vector of predictors with p "large", and f_t is the $1 \times k$ vector of factors with $k \ll p$. What these two equations describe is a situation where instead of inserting the large data x_t as predictors in the regression for our variable of interest y, we instead use a lower dimensional vector of k factors. As a consequence the matrix λ is $k \times p$ and the parameter vector γ is $k \times 1$. We can of course generalize and assume lags of factors in both equations, which is the case of the *dynamic* factor model; see for example Stock and Watson (2002). Depending on the exact specification and estimation method used, errors u_t typically have a $p \times p$ covariance matrix that is either diagonal (no correlation between the p variables in x_t) or it is characterized by weak correlation. In any case, the quantity $\chi_t = f_t \lambda$, also known as the "common component", is meant to model most of the covariation among the p series.

There are many benefits from specifying and using factor models. The obvious implication is that in our regression for y we have k parameters to estimate instead of p. In typical macro applications it will be the case that, say, $p = 120$ and $k = 3$, which will enhance greatly estimation accuracy. As already mentioned, another implication of factors is that they capture comovements among macro data. For simplicity take the example of three series that are highly correlated, namely GDP, employment and industrial production. These three series, that measure the output in

[14] There are numerous reviews of factor models in economics, but the recent work of Stock and Watson (2016) serves as a very thorough primer on factor models in macroeconometrics.

an economy, are subject to measurement error and frequent revisions from statistical offices. Additionally, they are only incomplete proxies of the more general notion of "economic activity" that macroeconomists have in mind when constructing theoretical models. In that respect, using the joint information (i.e. comovement) in these three variables via a single common factor is not only more parsimonious but it could also provide protection against data irregularities (Bernanke, Boivin and Eliasz, 2005). Another benefit is that the class of factor models has a representation which is similar to the state-space form of dynamic stochastic general equilibrium (DSGE) models with many observable variables and a smaller set of state variables (Forni and Gambetti, 2014). Finally, factor models have been used in other flexible settings that allow for our large data x_t to be unbalanced (Stock and Watson, 2002) or measured in various frequencies (Mariano and Murasawa, 2003); settings where factors enter a regression model as "soft" prior restrictions using Bayesian methods (Hahn, Carvalho and Mukherjee, 2013); settings where the factor model is used as a means of data-based forecast combinations (Chan, Stock and Watson, 1999), and in structural VAR analysis (Stock and Watson, 2005; Korobilis, 2013a).

There are several ways to estimate factor models. Parametric likelihood-based methods, such as maximum likelihood and Bayesian methods, are typically computationally cumbersome. This is due to the high latency of factor models, since equation (20) is a regression where both f_t and λ are latent. In such cases iterative estimation algorithms for latent data, such as the EM algorithm (Doz, Giannone and Reichlin, 2012) and the Gibbs sampler (Lopes and West, 2004), might be needed. An alternative approximate two-step approach that is asymptotically consistent is based on replacing factors with principal component estimates; see Stock and Watson (2016). Principal component analysis (PCA) provides nonparametric estimates that are based on eigenvalue decomposition of the covariance matrix (static approach) or the spectral density matrix (dynamic approach) of our data x_t, and not on the parametric likelihood of the factor model. Once factors are replaced by PCA estimates we can, in a second step, estimate equations (20) and (21) using simple OLS. The PCA approach is computationally simple, and due to the two-step approach it doesn't suffer from the same identification issue that occurs when estimating f_t and λ jointly in one step.[15] In that respect, it is not unreasonable that the PCA approach is the

[15] This issue is important, as demonstrated in Bernanke, Boivin and Eliasz (2005), which is possibly one of the very few studies in economics that compares results using both a simple two-step

most popular in economics despite several variants of factor models that have been proposed over the years.

Despite the popularity of principal components, in practical situations there might be several challenges that undermine their performance. First, principal components do not have any immediate economic interpretation[16] even if they summarize information in economic time series. When using large macroeconomic panels the first principal component can sometimes be thought of as a real activity factor (Stock and Watson, 2003, Section 3.3.2), but further factors cannot be labelled easily unless they are extracted in blocks of data releases as in Belviso and Milani (2006). Second, unlike likelihood-based factor models which can be used for nonstationary/persistent macro data (Peña and Poncela, 2004) principal components can only describe data that are fully or approximately stationary.[17] Additionally, unlike common belief among applied economists, principal component analysis of factor models is not a black-box procedure where we can simply add any number of explanatory variables and then be sure that factors will provide an optimal summary of these variables. Boivin and Ng (2006) formalize this argument by showing that just adding more data is not always better for factor analysis using principal components. Finally, if we want to go way beyond the standard large macro panels with 100-200 series and model with truly Big Data, then PCA is not computationally trivial any more. For example, implementing the eigenvalue decomposition of a $100,000 \times 100,000$ covariance matrix is anything but trivial.

In truly high-dimensional cases, there are alternative algorithms that can help us estimate factor models. In multi-dimensional analysis and machine learning in particular, there are several alternative methods with names such as probabilistic PCA, independent component analysis (ICA), linear discriminant analysis (LDA),

estimator (via PCA + OLS), and a Bayesian one-step estimator (via MCMC methods). Due to the unreasonable zero restrictions that ought to be imposed during estimation, structural impulse responses from the one-step estimation method degenerate to a zero median for many variables in their data set. In contrast, factors estimated with PCA help identify the same variables much better; see and compare Figures 2-5 in Bernanke, Boivin and Eliasz (2005).

[16] In addition to the naming/labelling issue, PCA is a nonparametric procedure that only approximates the true parametric factor model in equations (20) and (21). Therefore, PCA estimates of f_t are not optimized to convey maximum influence on y. That is, they provide a convenient, approximate two-step estimate of f_t based on a decomposition only of the data x_t without reference to our data y. In machine learning and artificial intelligence such methods are referred to as "unsupervised learning" procedures.

[17] The first principal component extracted from nonstationary data degenerates to be the simple mean of the p series, with the loadings vector λ being a vector of ones.

mixtures of factor analyzers, Principal Coordinates Analysis (PCoA) and other acronyms (Barber, 2012; Bishop, 2006). Of interest in high-dimensional inference is the method of random projections introduced in the field of compressive sensing (Donoho, 2006). As the name suggests linear projection methods can be used to project a high-dimensional matrix into a lower-dimensional vector or matrix (with the columns of the latter being a linear combination of the columns of the former). PCA is such a method that projects the data into lower-dimensional orthogonal vectors but, as already claimed, the optimization problem it solves (maximizing remaining variance of data explained by each component) can become computationally cumbersome in high dimensions. Random projection (RP), in contrast, is a "data-oblivious" method that simply requires to generate matrices that project our data x_t into a lower dimensional f_t using random numbers from our PC. Under certain conditions, the so-called Johnson-Lindenstrauss Lemma guarantees that a lower bound exists for the error from approximating our data with the randomly generated, lower-dimensional f_t. Guhaniyogi and Dunson (2015) present an interesting Bayesian application of this method where they combine RP with Bayesian model averaging in a regression problem and prove consistency of posterior predictive distributions. Maillard and Munos (2012) provide approximation error bounds in a least-squares regression problem involving RP compression of predictors. On a final note, there are other efficient algorithms related to PCA and general factor models that have also been used successfully in forecasting applications in economics, such as Partial Least Squares (PLS) and the three-Pass Regression Filter (3PRF) of Kelly and Pruitt (2015).

3 Extensions of the basic regression model

Up to this point, a large part of this review was devoted to summarizing methods that can be used in the univariate linear regression setting. Nevertheless, applied macroeconomists currently tell us[18] that for the problem of, say, modeling inflation, time-varying parameter regression models with stochastic volatility are empirically far superior than a linear regression. Similarly, modern macroeconomic problems need to be cast in a multivariate time series form, rather than a univariate one, in order to decompose and explain all static and dynamic linkages between variables.

[18] See for example Pettenuzzo and Timmerman (2017) and Stock and Watson (2007).

The purpose of this Section is to build further intuition by demonstrating various ways to approximate a high-dimensional inference problem in multivariate and time-varying parameter regression models. While there is always the option to solve such problems by relying on extensive derivations and state-of-the-art statistical algorithms, there are many cases where we can cast a possibly nonlinear or multivariate problem into (something that looks like) a linear regression. Or we can cast the linear multiple regression model with p predictors into a collection of regressions with one predictor. Once we do that, then we might be able to apply existing, simple estimation algorithms and adapt them to a much harder problem. The main idea for motivating such an approach is at the core of machine learning inference, and applied macroeconomists can learn a lot from this: instead of dealing with a difficult and hard to approximate problem, try to break it into smaller pieces that are easy to approximate quickly.[19] I explain what is meant by this procedure, using three distinct examples.

3.1 Variable elimination in regression

Variable elimination or marginalization is a machine learning procedure used in graphical models that, loosely speaking, allows (via certain rules) to break a high-dimensional inference problem into a series of smaller problems. We can use similar ideas in our standard regression setting in order to facilitate high-dimensional inference. Assume that we work again with a regression model setting with p predictors, but this time interest lies in the j-th predictor and its coefficient. We can rewrite the regression as

$$y = x_j \beta_j + x_{(-j)} \beta_{(-j)} + \varepsilon, \tag{22}$$

where y, x_j and ε are all $T \times 1$ vectors and $x_{(-j)}$ is a $T \times (p-1)$ predictor matrix with predictor j removed. It might be the case that we are interested only in parameter β_j because this is a policy parameter. A first useful result is the one of partitioned regression, or partial-time regression using the terminology of Frisch and Waugh (1933). Defining the $T \times T$ annihilator matrix $M_j = I_T - x_j \left(x_j' x_j \right)^{-1} x_j'$, it is easy to

[19] This logic is particularly important for high-dimensional machine learning inference because it allows to break a problem into multiple small steps that can be distributed into multiple processing cores, thus taking advantage of increased availability of multi-core CPUs, GPUs and High-Performance Clusters.

show using the algebra of partitioned matrices that $\widehat{\beta}_j$, the OLS estimates of β_j can be obtained as the solution of

$$\widehat{\beta}_j = \left(x_j' x_j\right)^{-1} x_j' \left(y - x_{(-j)} \widehat{\boldsymbol{\beta}}_{(-j)}\right) \tag{23}$$

where the sub-vector $\widehat{\beta}_{(-j)}$ is the solution of the following regression

$$\widehat{\beta}_{(-j)} = \left(x_{(-j)}^{\dagger\prime} x_{(-j)}^{\dagger}\right)^{-1} x_{(-j)}^{\dagger\prime} y^{\dagger} \tag{24}$$

with $x_{(-j)}^{\dagger} = M_j x_{(-j)}$ and $y^{\dagger} = M_j y$ denoting the projections of $x_{(-j)}$ and y on a space that is orthogonal to x_j.

This result provides very useful intuition about the relationships between our variables and coefficients in the OLS regression. Most importantly they can be generalized to efficient procedures for high-dimensional inference. Consider for example combining partitioned regression results with a penalized estimator instead of OLS. To demonstrate this point, I consider an alternative partition of the regression due to van den Boom, Reeves and Dunson (2015). Define the $T \times 1$ vector $q_j = x_j / \|x_j\|$, and generate randomly a matrix Q_j that is normalized as $Q_j Q_j' = I - q_j q_j'$. This means that the matrix $Q = [q_j, Q_j]$ is orthogonal, such that multiplying both sides of (22) by Q' gives

$$\begin{aligned}
Q'y &= Q'x_j\beta_j + Q'x_{(-j)}\beta_{(-j)} + Q'\varepsilon \Rightarrow & (25)\\
\begin{bmatrix} q_j' y \\ Q_j' y \end{bmatrix} &= \begin{bmatrix} q_j' x_j \\ Q_j' x_j \end{bmatrix} \beta_j + \begin{bmatrix} q_j' x_{(-j)} \\ Q_j' x_{(-j)} \end{bmatrix} \beta_{(-j)} + Q'\varepsilon \Rightarrow & (26)\\
\begin{bmatrix} y^* \\ y^+ \end{bmatrix} &= \begin{bmatrix} \|x_j\| \\ 0 \end{bmatrix} \beta_j + \begin{bmatrix} x_{(-j)}^* \\ x_{(-j)}^+ \end{bmatrix} \beta_{(-j)} + \widetilde{\varepsilon}, & (27)
\end{aligned}$$

where $y^* = q_j' y$, $y^+ = Q_j' y$, $x_{(-j)}^* = q_j' x_{(-j)}$, $x_{(-j)}^+ = Q_j' x_{(-j)}$ and $\widetilde{\varepsilon} = Q'\varepsilon$. In the derivation above we have used the fact that $Q_j' x_j = Q_j' q_j \|x_j\| = 0$ because Q_j and q_j are orthogonal. Additionally, $var(\widetilde{\varepsilon}) = \sigma^2 Q'Q = \sigma^2 = var(\varepsilon)$ because by construction $Q'Q = I$. The likelihood of the transformed regression model in equation (27) is multivariate Normal, which means we can use standard results for conditional Normal distributions to show that we can first estimate $\beta_{(-j)}, \sigma^2$ by regressing y^+ to $x_{(-j)}^+$, and then at a second stage obtain β_j by regressing y^* on $\|x_j\|$ conditional on

$\beta_{(-j)}, \sigma^2$ being known. This is a very useful result since now, conditional on obtaining in a first step some estimates of $\beta_{(-j)}, \sigma^2$, we can estimate β_j in a regression with known variance.[20] In a Bayesian context van den Boom, Reeves and Dunson (2015) use this result to derive analytically approximate marginal posteriors for β_j under a class of spike and slab priors; see equation (13). Korobilis and Pettenuzzo (2018) generalize this idea to high-dimensional VARs under a wider class of hierarchical shrinkage priors. Considering that the exact way of calculating marginal posteriors would involve solving numerically a $p-1$-dimensional integral for each j, doing this transformation and deriving the marginal posteriors analytically means large gains in computation.

3.2 Time-varying parameter models

Time-varying parameter models are a natural extension of the linear regression model analyzed in the previous Section. The standard form of the time-varying parameter (TVP) regression model used in economics is

$$\begin{aligned} y_t &= x_t\beta_t + \varepsilon_t, & (28)\\ \beta_t &= \beta_{t-1} + u_t, & (29)\\ \beta_0 &\sim N\left(\underline{\beta}, \underline{V}\right) & (30) \end{aligned}$$

where $u_t \sim N(0, Q)$ with Q a $p \times p$ covariance matrix, and for simplicity assume that $\varepsilon_t \sim N(0, \sigma^2)$ despite the fact that in practical situations one would also want the regression variance σ^2 to be time-varying. If we ignore the second equation for a moment, the first equation says that at each point in time a new regression coefficient holds, hence, the subscript t the coefficient vector. Granger (2008) quotes a very generic theorem suggested by Halbert White, stating that a general time-varying parameter specification can approximate any form of nonlinearity previously used in econometrics. The second and third equations of the system above make the time-varying regression model look like a state-space model that can be estimated using the Kalman filter algorithm and its variants (Kim and Nelson, 1999). The second equation can be viewed as a rule for the time series evolution of β_t, and the third equation is a necessary initial condition that has to be chosen as the data cannot

[20] Most importantly, we can do so in parallel for all predictors $j = 1, ..., p$.

provide any explicit information about period $t = 0$.

With a little bit of algebra we can immediately obtain further intuition about the TVP regression. We take only equation (28) and write it in a static regression form as

$$y = Z\beta + \varepsilon, \tag{31}$$

where $y = (y_1', ..., y_T')'$ and $\varepsilon = (\varepsilon_1', ..., \varepsilon_T')'$ are $T \times 1$ vectors, $\beta = (\beta_1, ..., \beta_T)$ is a $Tp \times 1$ vector of regression coefficients, and we define

$$Z = \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & x_T \end{bmatrix},$$

the $T \times Tp$ right-hand side matrix of predictors. Even though equations (28) and (31) are observationally equivalent, the second form allows us to see the TVP regression as a regression problem with many predictors. We originally have p predictors but by the time we have to estimate one time-varying parameter for each of these predictors, the static regression form of the TVP model has Tp "predictors".[21] This equation on its own cannot be estimated estimated with OLS because Z is of rank T when we have to estimate Tp coefficients. In this case some form of shrinkage along the lines of what we analyzed in the previous Section could allow estimation of the TVP regression. But this is exactly what equation (29) does in the original TVP model: instead of viewing this equation as a time-series autoregression for β_t, this equation can be viewed from a Bayesian perspective as a conditional prior of the form $p(\beta_t|\beta_{t-1}) \sim N(\beta_{t-1}, Q)$. As we argued in Section 2.3 such priors, if properly tuned and selected, lead to Bayesian shrinkage estimators that will allow estimation of the high-dimensional regression model in (31). However, even though use of the prior $p(\beta_t|\beta_{t-1})$ has been popular at least since the works of Cooley (1971) and Cooley and Sargent (1976), we can simply ignore this specific shrinkage prior and the resulting state-space methods. We can choose, instead, to estimate equation (31) using the lasso or some other penalized regression estimator of our choice and achieve similar results. Korobilis (2018) applies such ideas in a forecasting problem involving TVP regressions with many predictors. Shrinkage is implemented using hierarchical

[21] This form is basically the multiviarate form of a regression with time dummies for each predictor.

priors used widely in statistics, but computation relies on the *sum-product* algorithm developed in computing science (see Wand, 2017, for a review that is accessible to statisticians).

Even though we just showed how the autoregressive (random walk to be exact) evolution of β_t is not the only way to estimate the TVP regression problem, assume again the TVP regression in its original, popular form in equations (28) - (30). As shown in Früwirth-Schnatter and Wagner (2010), such state-space models can be written in an equivalent "non-centered" form:

$$\begin{aligned} y_t &= x_t\theta + x_t\theta_t + \varepsilon_t, & (32) \\ \theta_t &= \theta_{t-1} + u_t, & (33) \\ \theta_0 &\sim N(0,0) \equiv 0. & (34) \end{aligned}$$

When comparing to the original form it holds that $\beta_t = \theta + \theta_t$. Given the restriction that θ_t is initialized at a fixed point ($\theta_0 = 0$), means that in this formulation θ is the equivalent of the random initial condition we had for β_t. The major difference is that what was the initial condition can now be interpreted as the constant part of the regression model, and θ_t is the add-on time-varying part of the regression. Notice now how this new specification facilitates high-dimensional inference as we can now do shrinkage or variable selection (along the lines of the previous chapter) on the coefficient θ whilst in the original specification it wasn't obvious how to do the same with the initial condition β_0. Indeed, Früwirth-Schnatter and Wagner (2010) proceed in their analysis by using a spike and slab prior on θ.[22]

3.3 Vector autoregressions

Some of the most important quantitative exercises that policy-makers are interested in, involve the vector autoregressive (VAR) model and its variants. Economic theories can be tested reliably only in a multivariate econometric setting, and the same holds to a large degree for measuring the impact of shocks to the wider economy. While a large part of empirical analysis is done using VARs of say three or five variables, there is an expanding literature that acknowledges the benefits of large VARs (Ellahie and Ricco, 2017). In particular, small structural VARs might not be invertible (Forni

[22] The authors go one step further than that by also scaling equation and applying a shrinkage prior to θ_t; see Früwirth-Schnatter and Wagner (2010) for more details.

and Gambetti, 2014) meaning that their residuals will not span the same space as the structural shocks that macroeconomists want to identify (Bernanke, Boivin and Eliasz, 2005). From a non-structural point of view Bańbura, Giannone and Reichlin (2010) were the first to show that VARs with 130 endogenous variables and almost quarter of a million parameters can be used to forecast variables of interest. Since then, there is an expanding and lively literature on methods for estimating large VARs, see for example Koop, Korobilis and Pettenuzzo (forthcoming) and references therein.

A vector autoregression for an $1 \times n$ vector of variables of interest y_t can be written in the following form

$$y_t = B_0 + \sum_{i=1}^{p} y_{t-i} B_i + \varepsilon_t, \tag{35}$$

but we can write it in familiar multivariate regression form as

$$y_t = X_t B + \varepsilon_t, \tag{36}$$

where $X_t = (1, y_{t-1}, ..., y_{t-p})$, $A = [B_0, B_1, ..., B_p]$ and $\varepsilon_t \sim N(0, \Sigma)$ with Σ and $n \times n$ covariance matrix. Accumulation of parameters in VARs is quite different compared to univariate models. A VAR with $n = 3$ variables, intercept terms and $p = 1$ lag has 18 parameters. The same VAR with $n = 50$ variables has 3825 parameters. The last VAR with $p = 12$ has 31325 parameters. This gives an idea of the polynomial rate at which the number of parameters increases as n and/or p increase.

Many applied macroeconomists choose to shrink smaller or larger VARs using Bayesian methods, along the lines of variable selection or shrinkage priors discussed in the previous Section; see Miranda-Agrippino and Ricco (2018) in this Encyclopedia. Fortunately, as Bańbura, Giannone and Reichlin (2010) and Giannone, Lenza and Primiceri (2015) show, there is a way to use elegant Bayesian shrinkage priors in a VAR without having to rely on computationally expensive simulation methods to derive the posterior. The method relies on the so-called natural conjugate prior, that allows analytical derivations of (regularized) parameter posterior moments in the same way we derived equation (18). Nevertheless, such natural conjugate priors have the limitation that they treat VAR equations symmetrically. The implication of this is that we cannot impose or test restrictions from economic theory that suggest that only some independent variables might affect our endogenous variables. For example in VAR system with money, output and inflation, imposing money neutrality means that lags of money should not enter the equation for output. However, doing so means that lags of money should also affect inflation, even if we have reasons to believe that this is not correct. Allowing money to affect inflation means that we have to allow it to also affect GDP. Such a rigid situation is not ideal, especially in high dimensions when we might have hundreds of endogenous variables, and thousands of interactions among them.

Nevertheless, there are still simple ways to readily apply ideas from the univariate regression model. Carriero, Clark and Marcellino (2017) and Koop, Korobilis and Pettenuzzo (forthcoming), in the context of developing efficient estimation algorithms for large VARs, proposed to break the VAR into a collection of n univariate equations. Using ideas from estimation of simultaneous equation models (Hausman, 1983) we can transform the VAR in triangular form. Consider the Cholesky-like decomposition of the covariance matrix, $\Sigma = A^{-1} D \left(A^{-1}\right)'$ where D is a diangonal matrix for variances, and A^{-1} is a unitriangular matrix of the form

$$\boldsymbol{A}^{-1} = \begin{bmatrix} 1 & 0 & ... & 0 & 0 \\ \alpha_{2,1} & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ \alpha_{n-1,1} & ... & \alpha_{n-1,n-2} & 1 & 0 \\ \alpha_{n,1} & ... & \alpha_{n,n-2} & \alpha_{n,n-1} & 1 \end{bmatrix}. \tag{37}$$

Under this decomposition we can rewrite the VAR in equation (36) as

$$\begin{aligned} y_t &= X_t B + u_t \left(A^{-1} D^{\frac{1}{2}} \right)' \Rightarrow && (38) \\ y_t A &= X_t B A + u_t, D^{\frac{1}{2}} \Rightarrow && (39) \\ y_t + y_t \widetilde{A} &= X_t \Gamma + u_t, D^{\frac{1}{2}} \Rightarrow && (40) \\ y_t &= X_t \Gamma - y_t \widetilde{A} + u_t, D^{\frac{1}{2}}, && (41) \end{aligned}$$

where $u_t \sim N(0, I)$, $\Gamma = B \times A$ and $\widetilde{A} = A - I$ is a lower diagonal matrix created from A after we remove its unit diagonal elements. This is a so-called triangular VAR system due to the fact that $\widetilde{A}$ has a lower triangular structure. It cannot be estimated as a multivariate regression using standard linear estimators because y_t shows up both on the left-hand side and the right-hand side of the equation. However, due to the lower triangular structure of $\widetilde{A}$ and the fact that D is diagonal the system can be estimated equation-by-equation using simple OLS (Hausman, 1983). This means that in high dimensions we can essentially write the VAR in this form and apply any univariate regression estimator and algorithm we like.[23] More importantly, note that the last equation shows that all contemporaneous covariances among the n VAR equations can be written as RHS predictors $-y_t$. This is an important implication because it shows that $\widetilde{A}$ can be treated as a regression parameter and (given that we can estimate these equations recursively) we can readily apply methods of the previous section to impose shrinkage also on the VAR covariance matrix.

Finally, Carriero, Clark and Marcellino (2016) derive a similar triangular VAR that has slightly different representation and implications for estimation. Begin with equation (36) but now rewrite it in the form

$$\begin{aligned} y_t &= X_t B + u_t \left(A^{-1} D^{\frac{1}{2}} \right)' \Rightarrow & (42) \\ y_t &= X_t B + u_t \left(\left(\widetilde{A}^{-1} + I \right) D^{\frac{1}{2}} \right)' \Rightarrow & (43) \\ y_t &= X_t B + u_t \widetilde{A}^{-1} D^{\frac{1}{2}} + u_t D^{\frac{1}{2}} \Rightarrow & (44) \\ y_t &= X_t B + v_t \widetilde{A}^{-1} + v_t, & (45) \end{aligned}$$

where $v_t \sim N(0, D)$ and $\widetilde{A}^{-1} = A^{-1} - I$ is a triangular matrix created by removing the identity diagonal of A^{-1}. As Carriero, Clark and Marcellino (2016) show, the above system can also be estimated equation by equation, where in equation i we use residuals from the previous $i - 1$ equations. This form has different implications for designing estimation algorithms compared to the one in (41), even though they are observationally equivalent. Equation (45) allows direct estimation of the VAR matrices B and A^{-1}, while equation (41) estimates functions of those, i.e. Γ and A. Such examples show that high-dimensional inference can be approximated by efficient transformations of the VAR model that allow to readily apply univariate estimators which are simpler and possibly algorithmically faster.

[23] Of course, note that this flexibility comes at the cost of shrinkage or variable selection being dependent on the ordering of the variables in the VAR; see Koop, Korobilis and Pettenuzzo (forthcoming) for a discussion.

References

[1] Aït-Sahalia, Y. and Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics* 201(2), 384-399.

[2] Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation* 9(7), 1545-1588.

[3] Avramov, D. (2002.) Stock return predictability and model uncertainty. *Journal of Financial Economics* 64, 423-258.

[4] Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics* 3(2), 89-163.

[5] Bañbura, M., Giannone, D. and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics* 25(1), 71-92.

[6] Barber, D. (2012). *Bayesian reasoning and machine learning*, Cambridge University Press.

[7] Belviso F. and Milani, F. (2006). Structural factor-augmented VARs (SFAVARs) and the effects of monetary policy. *B.E. Journal of Macroeconomics* 6(3), 1-46.

[8] Benson Durham, J. (2001). Sensitivity analyses of anomalies in developed stock markets. *Journal of Banking and Finance* 25(8),1503-1541.

[9] Bernanke, B., Boivin, J. and Eliasz, P. (2005). Measuring the effects of monetary policy a factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics* 120, 387-422.

[10] Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika* 103(4), 955-969.

[11] Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer: New York.

[12] Boivin, J. and Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics* 132(1), 169-194.

[13] Box, G. and Jenkins, G. (1970). *Time series analysis: Forecasting and control.* San Francisco: Holden-Day.

[14] Breiman, L. (1996). Bagging predictors. *Machine Learning* 24(2), 123-140.

[15] Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *Annals of Statistics* 26(3), 801-849.

[16] Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data.* Springer Series in Statistics: 9, Springer-Verlag: Berlin Heidelberg.

[17] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* 35(6), 2313-2351.

[18] Carriero, A., Clark, T. and Marcellino, M. (2016). Large vector autoregressions with stochastic volatility and flexible priors. Working paper 16-17, Federal Reserve Bank of Cleveland.

[19] Chan, Y. L., Stock, J. H. and Watson, M. W. (1999). A dynamic factor model framework for forecast combination. *Spanish Economic Review* 1, 91-121.

[20] Chen, H., Mirestean A. and Tsangarides, C. (2009). Limited information Bayesian model averaging for dynamic panels with short time periods. IMF Working Paper 09/74.

[21] Christ, C. F. (1994). The Cowles Commission contributions to econometrics at Chicago: 1939-1955". *Journal of Economic Literature* 32(1), 30-59.

[22] Cooley, T. F. (1971). Estimation in the presence of sequential parameter variation. Ph.D Thesis. Department of Economics, University of Pennsylvania.

[23] Cooley, T. F. and Prescott, E. C. (1976). Estimation in the presence of stochastic parameter variation. *Econometrica* 44(1), 167-184.

[24] Cremers, K. J. M. (2002). Stock return predictability: A Bayesian model selection perspective. *The Review of Financial Studies* 15, 1223-1249.

[25] Doppelhofer, G., Miller, R. I. and Sala-i-Martin, X. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review* 94(4), 813-835.

[26] Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory* 52(4), 1289-1306.

[27] Donoho, D. L., Maleki, A. and Montanari, A. (2009). Message passing algorithms for compressed sensing. *Proceedings of National Academy of Sciences* 106(45), 18914-18919.

[28] Doz, C., Giannone, D. and Reichlin, L. (2012). A quasimaximum likelihood approach for large, approximate dynamic factor models. *Review of Economics and Statistics* 94(4), 1014-1024.

[29] Durlauf, S., Kourtellos, A. and Tan, C. (2011). Is God in the details? A reexamination of the role of religion in economic growth. *Journal of Applied Econometrics* 27, 1059-1075.

[30] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* 32(2), 407-499.

[31] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors - An empirical Bayes approach. *Journal of the American Statistical Association* 68(341), 117-130.

[32] Eicher, T., Papageorgiou, C. and Raftery, A. (2009). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics* 26, 30-55.

[33] Ellahie, A. and Ricco, G. (2017). Government purchases reloaded: Informational insufficiency and heterogeneity in fiscal VARs. *Journal of Monetary Economics* 90, 13-27.

[34] Elliott, G., Gargano, A. and Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics* 177(2), 357-373.

[35] Fernández, C., Ley, E. and Steel, M. (2001a). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100, 381-427.

[36] Fernández, C., Ley, E. and Steel, M. (2001b). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* 16, 563-576.

[37] Fisher, R.A. (1925). *Statistical methods for research workers.* Oliver and Boyd: Edinburgh. ISBN 0-05-002170-2.

[38] Forni, M. and Gambetti, L. (2014). Sufficient information in structural VARs. *Journal of Monetary Economics* 66, 124-136.

[39] Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica* 1, 387-401.

[40] Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics* 154(1), 85-100.

[41] Garratt, A., Koop, G., Mise, E. and Vahey, S. (2009). Real-time prediction with U.K. monetary aggregates in the presence of model uncertainty. *Journal of Business and Economic Statistics* 27, 480-491.

[42] Geisel, M. (1973). Bayesian comparisons of simple macroeconomic models. *Journal of Money, Credit and Banking* 5, 751-772.

[43] George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 88(423), 881-889.

[44] George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 7, 339-373.

[45] George, E. I., Sun, D. and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics* 142(1), 553-580.

[46] Geweke, J. (1977). The dynamic factor analysis of economic time series models. in Aigner, D. and Goldberger, A. (Eds.). *Latent variables in socioeconomic models*, 365-383. Amsterdam: North-Holland.

[47] Ghosh, S., Khare, K. and Michailidis, G. (forthcoming). High dimensional posterior consistency in Bayesian vector autoregressive models. *Journal of the American Statistical Association* DOI: https://doi.org/10.1080/01621459.2018.1437043.

[48] Giannone, D., Primiceri, G. E. and Lenza, M. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics* 97, 436-451.

[49] Granger, C. W. (2008). Non-linear models: Where do we go next - time varying parameter models? *Studies in Nonlinear Dynamics and Econometrics* 12(3), 1-11.

[50] Granger, C. W. J. and Uhlig, H. F. (1990). Reasonable extreme-bounds analysis. *Journal of Econometrics* 44, 159-170.

[51] Guhaniyogi R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association* 110(512), 1500-1514.

[52] Hahn, P. R., Carvalho, C. M. and Mukherjee, S. (2013). Partial factor modeling: Predictor-dependent shrinkage for linear regression. *Journal of the American Statistical Association* 108(503), 999-1008.

[53] Hansen, S., McMahon, M. and Prat, A. (forthcoming). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*, https://doi.org/10.1093/qje/qjx045.

[54] Hausman, J. A. (1983). Specification and estimation of simultaneous equation models. in Griliches, Z. and Intriligator, M. D. (Eds.). *Handbook of Econometrics* 1, Chapter 7, 391-448, Elsevier.

[55] Hendry, D. F. and Doornik, J. A. (2014). Empirical model discovery and theory evaluation: Automatic selection methods in econometrics. MIT Press.

[56] Hendry, D. F. and Krolzig, H.-M. (2005). The properties of automatic GETS modelling. *Economic Journal*, 115(502), C32-C61.

[57] Jiménez, G., Ongena, S., Peydro, J. and Saurina, J. (2014). Hazardous times for monetary policy: What do 23 million loans say about the impact of monetary policy on credit risk-taking? *Econometrica* 82(2), 463-505.

[58] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* 107(498), 649-660.

[59] Kapetanios, G., Labhard, V. and Price, S. G. (2008). Forecasting using Bayesian and information-theoretic model averaging: An application to U.K. inflation. *Journal of Business and Economic Statistics* 26(1), 33-41.

[60] Kelly, B. and Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics* 186(2), 294-316.

[61] Kim, C-J. and Nelson, C. (1999). *State-space models with regime switching: Classical and Gibbs-sampling approaches with applications.* The MIT Press,

[62] Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review* 53, 867-886.

[63] Koop, G. and Korobilis, D. (2018). Variational Bayes inference in high-dimensional time-varying parameter models. mimeo.

[64] Koop, G., Korobilis, D. and Pettenuzzo, D. (forthcoming). Bayesian compressed VARs. *Journal of Econometrics.*

[65] Korobilis, D. (2008). Forecasting in vector autoregressions with many predictors. *Advances in Econometrics* 23 (Bayesian Macroeconometrics), 403-431.

[66] Korobilis, D. (2013a). Assessing the transmission of monetary policy shocks using time-varying parameter dynamic factor models. *Oxford Bulletin of Economics and Statistics* 75, 157-179.

[67] Korobilis, D. (2013b). Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting* 29, 43-59.

[68] Korobilis, D. (2013c). VAR forecasting using Bayesian variable selection. Journal of Applied Econometrics 28, 204-230.

[69] Korobilis, D. (2016). Prior selection for panel vector autoregressions. *Computational Statistics and Data Analysis* 101, 110-120.

[70] Korobilis, D. (2018). Forecasting with many predictors using message passing algorithms. mimeo.

[71] Korobilis, D. and Pettenuzzo, D. (2018). Adaptive hierarchical priors for high-dimensional vector autoregressions. *Journal of Econometrics*, forthcoming.

[72] Kyung, M., Gill, J., Ghoshz, M. and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* 5, 369-412.

[73] Leamer, E. E., (1983). Lets take the con out of econometrics, *American Economic Review* 73, 31-43.

[74] Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *American Economic Review* 82(4), 942-963.

[75] Litterman, R. (1979). Techniques of forecasting using vector autoregressions. Federal Reserve Bank of Minneapolis Working Paper 115.

[76] Lopez, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* 14, 41-67.

[77] Maillard, O-A. and Munos, R. (2012). Linear regression with random projections. *Journal of Machine Learning Research* 13, 2735-2772.

[78] Mariano, R. S. and Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics* 18, 427-443.

[79] Masanjala, W. and Papageorgiou, C. (2008). Rough and lonely road to prosperity: a reexamination of the sources of growth in Africa using Bayesian model averaging. *Journal of Applied Econometrics* 23, 671-682.

[80] McAleer, M., Pagan, A.R. and Volker, P. A. (1985). What will take the con out of econometrics? *American Economic Review* 75, 293-307.

[81] McCracken, M. and Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business and Economic Statistics* 34(4), 574-589.

[82] Min, C.-K. and Zellner, A. (1993). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rates. *Journal of Econometrics* 56(1-2), 89-118.

[83] Moral-Benito, E. (2012) Determinants of economic growth: a Bayesian panel data approach. *The Review of Economics and Statistics* 94, 566-579.

[84] Mousavi, A., Maleki, A. and Baraniuk, R. G. (2017). Consistent parameter estimation for LASSO and approximate message passing. *Annals of Statistics* 45(6), 2427-2454.

[85] Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31(2), 87-106.

[86] Nicoletti-Altimari, S. (2001). Does money lead inflation in the Euro-Area?. ECB Working Paper No. 63.

[87] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* 103, 681-686.

[88] Peña, D. and Poncela, P. (2004). Forecasting with nonstationary dynamic factor models. *Journal of Econometrics* 119(2), 291-321.

[89] Pesaran, M. H. and Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance* 50(4), 1201-1228.

[90] Pettenuzzo, D. and Timmermann, A. (2017). Forecasting macroeconomic variables under model instability. *Journal of Business and Economic Statistics* 35, 183-201.

[91] Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology* 25, 111-163.

[92] Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review* 87(2), 178-183.

[93] Sargan, J. D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology. in Hart, P. E., Mills, G. and Whittaker, J. N. (Eds.) *Econometric Analysis for National Economic Planning*, 16, 25-54. London: Butterworths

[94] Sargent, T. J. and Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Papers 55, Federal Reserve Bank of Minneapolis.

[95] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proceedings of the Third Berkeley Symposium in Mathematical Statistics & Probability*, 1, 197-206.

[96] Stock, J. H. and Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics* 44(2), 293-335.

[97] 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(460), 1167-1179.

[98] Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. mimeo, available at: http://www.princeton.edu/ mwatson/papers/favar.pdf

[99] Stock, J. H. and Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking* 39, 333.

[100] Stock, J. H., and Watson, M. W. (2016). Factor models and structural vector autoregressions in macroeconomics. in Taylor, J. B. and Uhlig, H. (Eds). *Handbook of Macroeconomics*, Vol2A, Chapter 8, 415-526.

[101] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 58(1), 267-288.

[102] Tibshirani, R. and Knight, K. (1999). Model search by bootstrap "bumping". *Journal of Computational and Graphical Statistics* 8(4), 671-686.

[103] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67(1), 91-108.

[104] van den Boom, W., Reeves, G. and Dunson, D. B. (2015). Quantifying uncertainty in variable selection with arbitrary matrices. in *Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Cancun, Mexico.

[105] Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association* 112(517), 137-168.

[106] Wang, Y. and D. M. Blei (forthcoming). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, https://doi.org/10.1080/01621459.2018.1473776.

[107] Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 70(2), 129-133.

[108] Wright, J. H. (2008). Bayesian model averaging and exchange rate forecasting. *Journal of Econometrics* 146, 329-341.

[109] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67(2), 301-320.