Time-varying sparsity in dynamic regression models

Professor Jim Griffin
(joint work with Maria Kalli, Canterbury Christ Church University)

University of Kent
Often we are interested in making forecasts or predictions (e.g. inflation).

Regression models
Often we are interested in making forecasts or predictions (e.g. inflation).

One generic method for prediction is **regression** where we assume that there are observed predictors which can be used to help. The model for predicting \( j \) periods ahead is

\[
y_{t+j} = \alpha + \sum_{k=1}^{p} x_{t,k} \beta_k + \epsilon_t, \quad t = 1, 2, \ldots, T
\]

where

- \( y_t \) is the observation of the response variable at time \( t \).
- \( x_{t,k} \) is the value of the \( k \)-th predictor at time \( t \).
- \( \beta_k \) is the coefficient for the \( k \)-th predictor.
We used the data set of Goyal and Welch (2008)
We used the data set of Goyal and Welch (2008)

- The response variable is the value weighted monthly return of the S & P 500 obtained from the CRSP database.
We used the data set of Goyal and Welch (2008)

- The response variable is the value weighted monthly return of the S & P 500 obtained from the CRSP database.
- The sample period is May 1937 to December 2002.
We used the data set of Goyal and Welch (2008)

- The response variable is the value weighted monthly return of the S & P 500 obtained from the CRSP database.
- The sample period is May 1937 to December 2002.
- The set of twelve predictors includes variables relating to dividends, earnings, interest rates, bond yields and inflation.
Equity premium: regressions over decades

- **B/M**
- **TBL**
- **LTY**
- **NTIS**

- **40s 60s 80s**
- **INFL**
- **LTR**
- **SVAR**

- **40s 60s 80s**
- **D/Y**
- **EPR**
- **DE**

- **40s 60s 80s**
- **DFY**
The previous results show that the coefficient estimates are changing over time which suggests that the relationship between the predictors and the response is changing over time.
The previous results show that the coefficient estimates are changing over time which suggests that the relationship between the predictors and the response is changing over time.

The results also suggest that some predictors may not be important for predicting the response at all times or at some times.
The previous results show that the coefficient estimates are changing over time which suggests that the relationship between the predictors and the response is changing over time.

The results also suggest that some predictors may not be important for predicting the response at all times or at some times.

This is an explanation of why “static” regression models, where predictor effects are assumed constant over time often produce poor out-of-sample forecasts or predictions when fitted to different time periods (see Fisher and Statman (2006), Paye and Timmermann (2006) and Dangl and Halling (2012)).
One solution is the dynamic regression model.
One solution is the dynamic regression model. The model assumes that

\[ y_{t+j} = \alpha_t + \sum_{k=1}^{p} x_{t,k} \beta_{t,k} + \epsilon_t, \quad t = 1, 2, \ldots, T \]

where

- \( y_t \) is the observation of the response variable at time \( t \).
- \( x_{t,k} \) is the value of the \( k \)-th predictor at time \( t \).
- \( \beta_{t,k} \) is the coefficient for the \( k \)-th predictor at time \( t \).
- \( \epsilon_t \) is the error at time \( t \).
Estimation

The model cannot be directly estimated since there are \((p + 1)T\) coefficients and only \(T\) observations.
The model cannot be directly estimated since there are \((p + 1)T\) coefficients and only \(T\) observations.

One standard solution assumes that 
\[ \beta_{1,1}, \ldots, \beta_{1,p}, \ldots, \beta_{T,1}, \ldots, \beta_{T,p} \] follows a stochastic process such as a random walk

\[ \beta_{t,k} = \beta_{(t-1),k} + \nu_{t,k} \]

or vector autoregressive process

\[ \beta_t = \Lambda \beta_{t-1} + \nu_t \]

where \(\nu_{t,k}\) and \(\nu_t\) are random disturbances.
Our focus is on how to construct a stochastic process for $\beta_{t,k}$ which corresponds to the findings from the results.
Our focus is on how to construct a stochastic process for $\beta_{t,k}$ which corresponds to the findings from the results.

1. Many values of $\beta_{t,k}$ will be close to zero.
Dynamic regression models with many regressors

Our focus is on how to construct a stochastic process for $\beta_{t,k}$ which corresponds to the findings from the results

1. Many values of $\beta_{t,k}$ will be close to zero.
2. $\beta_{t,k}$ will be close to zero at all times or at some times for some predictors.
Dynamic regression models with many regressors

Our focus is on how to construct a stochastic process for $\beta_{t,k}$ which corresponds to the findings from the results

1. Many values of $\beta_{t,k}$ will be close to zero.
2. $\beta_{t,k}$ will be close to zero at all times or at some times for some predictors.
3. Some coefficients will have values of $\beta_{t,k}$ which are away from zero for all or most times.
Bayesian regularization

Returning to the static regression model, the coefficients $\beta$ need to be given a prior distribution.
Bayesian regularization

Returning to the static regression model, the coefficients $\beta$ need to be given a prior distribution.

If there are many predictors, it is natural to assume that some (perhaps, many) coefficients will have values close to zero.
Returning to the static regression model, the coefficients $\beta$ need to be given a prior distribution.

If there are many predictors, it is natural to assume that some (perhaps, many) coefficients will have values close to zero.

The proportion of such predictors is often called the sparsity of the regression problem.
Bayesian regularization

Returning to the static regression model, the coefficients $\beta$ need to be given a prior distribution.

If there are many predictors, it is natural to assume that some (perhaps, many) coefficients will have values close to zero.

The proportion of such predictors is often called the sparsity of the regression problem.

A prior distribution which can express different levels of sparsity is the normal-gamma, which has been suggested as a prior distribution in regression problems by Caron and Doucet (2008) and Griffin and Brown (2010).
The normal-gamma prior can be written as

$$\beta_k | \psi_k \sim N(0, \psi_k), \quad \psi_k \sim \text{Ga}(\lambda, \lambda/\mu)$$

where $\psi_k$ will be called the relevance parameter for the $k$-th predictor.
The normal-gamma prior can be written as

\[ \beta_k | \psi_k \sim N(0, \psi_k), \quad \psi_k \sim Ga(\lambda, \lambda/\mu) \]

where \( \psi_k \) will be called the relevance parameter for the \( k \)-th predictor.

The distribution of \( \psi_k \) expresses our prior belief about the relevance of the \( k \)-th predictor.
The normal-gamma prior can be written as

\[ \beta_k | \psi_k \sim N(0, \psi_k), \quad \psi_k \sim Ga(\lambda, \lambda/\mu) \]

where \( \psi_k \) will be called the relevance parameter for the \( k \)-th predictor.

The distribution of \( \psi_k \) expresses our prior belief about the relevance of the \( k \)-th predictor.

The choice of \( \lambda \) and \( \mu \) determines how much mass is placed close to or away from zero.
Normal-gamma distribution

\[ \log f(\beta) \]

\[ \log f(\Psi) \]

- \( \lambda = 0.1 \) (solid line)
- \( \lambda = 0.333 \) (dot-dashed line)
- \( \lambda = 1 \) (dashed line)
For dynamic regression models, we will construct a stochastic process for the coefficients whose marginal distribution is normal-gamma.
Normal-gamma autoregressive process

For dynamic regression models, we will construct a stochastic process for the coefficients whose marginal distribution is normal-gamma.

We write $\beta_{t,k} = \sqrt{\psi_{t,k}} \phi_{t,k}$ where

- $\phi_{1,k}, \ldots, \phi_{T,k}$ follows an AR(1) process with a standard normal stationary distribution and AR parameter $\varphi_k$.

- $\psi_{1,k}, \ldots, \psi_{T,k}$ follows an AR(1) process with a gamma marginal distribution with parameters $\lambda_k$ and $\mu_k$ and AR parameter $\rho_k$ (Pitt and Walker, 2005).
Normal-gamma autoregressive process

For dynamic regression models, we will construct a stochastic process for the coefficients whose marginal distribution is normal-gamma.

We write $\beta_{t,k} = \sqrt{\psi_{t,k}} \phi_{t,k}$ where

- $\phi_{1,k}, \ldots, \phi_{T,k}$ follows an AR(1) process with a standard normal stationary distribution and AR parameter $\varphi_k$.

- $\psi_{1,k}, \ldots, \psi_{T,k}$ follows an AR(1) process with a gamma marginal distribution with parameters $\lambda_k$ and $\mu_k$ and AR parameter $\rho_k$ (Pitt and Walker, 2005).

The processes $\phi_{1,k}, \ldots, \phi_{T,k}$ is independent of $\psi_{1,k}, \ldots, \psi_{T,k}$. 
NGAR process ($\lambda_k = 0.2$)

$\rho_k = 0.97$

$\varphi_k = 0.97$

$\varphi_k = 0.99$

$\rho_k = 0.99$
NGAR process ($\lambda_k = 1$)

\[ \varphi_k = 0.97 \]

\[ \rho_k = 0.97 \]

\[ \varphi_k = 0.99 \]

\[ \rho_k = 0.99 \]
• $\lambda_k$ controls the proportion of time that the process is close to zero.
• $\lambda_k$ controls the proportion of time that the process is close to zero.
• $\mu_k$ controls the scale of the process.
• \( \lambda_k \) controls the proportion of time that the process is close to zero.
• \( \mu_k \) controls the scale of the process.
• The parameters \( \varphi_k \) and \( \rho_k \) control the length of each period away from zero. Therefore, we choose priors for these parameters with most of their mass close to 1.
Further prior modelling

The NGAR process defines a process which spends some time close to zero and some time away from zero.
Further prior modelling

The NGAR process defines a process which spends some time close to zero and some time away from zero. However, “close” and “away” are relative to the parameter $\mu_k$ which controls scale.
Further prior modelling

The NGAR process defines a process which spends some time close to zero and some time away from zero. However, “close” and “away” are relative to the parameter $\mu_k$ which controls scale.

We define a second level of prior on $\mu_k$ so that

$$\mu_k \sim (\lambda^*, \lambda^*/\mu^*).$$

Smaller values of $\lambda^*$ suggests that more coefficients are close to zero at all times.
The NGAR process defines a process which spends some time close to zero and some time away from zero. However, “close” and “away” are relative to the parameter $\mu_k$ which controls scale.

We define a second level of prior on $\mu_k$ so that

$$
\mu_k \sim (\lambda^*, \lambda^*/\mu^*).
$$

Smaller values of $\lambda^*$ suggests that more coefficients are close to zero at all times.

The intercept $\alpha_t$ is assumed to follow a random walk.
The dynamic regression model is

\[ y_{t+j} = \alpha_t + \sum_{k=1}^{p} x_{t,k} \beta_{t,k} + \epsilon_t, \quad t = 1, 2, \ldots, T \]

which is completed by assuming that \( \epsilon_t \sim N(0, \sigma_t^2) \) where \( \sigma_t^2 \) is a given an AR(1) process with a gamma marginal distribution.
Simulated example

Intercept

X1

X2

X3

X4

X5

Intercept

X1

X2

X3

X4

X5

Intercept

X1

X2

X3

X4

X5

Intercept

X1

X2

X3

X4

X5
We used the data set of Goyal and Welch (2008)

- The response variable is the value weighted monthly return of the S & P 500 obtained from the CRSP database.
- The sample period is from May 1937 to December 2002.
- The set of twelve predictors includes variables relating to dividends, earnings, interest rates, bond yields and inflation.
Equity premium: coefficients
Equity premium: relevance

EPR

CSP

DE
• We constructed a data set using data series obtained from: FRED, the consumer survey database of the University of Michigan, the Federal Reserve Bank of Philadelphia, and the Institute of Supply Management.
We constructed a data set using data series obtained from: FRED, the consumer survey database of the University of Michigan, the Federal Reserve Bank of Philadelphia, and the Institute of Supply Management.

We predict the gross domestic product (GDP) deflator.
Inflation

- We constructed a data set using data series obtained from: FRED, the consumer survey database of the University of Michigan, the Federal Reserve Bank of Philadelphia, and the Institute of Supply Management.
- We predict the gross domestic product (GDP) deflator.
- The sample period is from Q2 of 1965 to Q1 of 2011.
We constructed a data set using data series obtained from: FRED, the consumer survey database of the University of Michigan, the Federal Reserve Bank of Philadelphia, and the Institute of Supply Management.

We predict the gross domestic product (GDP) deflator.

The sample period is from Q2 of 1965 to Q1 of 2011.

The data set includes 31 predictors, from activity and term structure variables to survey forecasts and previous lags.
GDP deflator: coefficients
GDP deflator: relevance
Out-of-sample prediction

We compare the DR model with the NGAR process prior to

- Time Varying Dimension (TVD) models (Chan et al, 2012),
- Dynamic Model Average (DMA) approach (Koop and Korobilis, 2011),
- Hierarchical shrinkage (HierShrink) (Belmonte et al, 2011).
- Rolling window Bayesian Model Averaging (BMA) using a $g$-prior for prediction.

$$\text{RMSE} = \sqrt{\frac{1}{T-s} \sum_{t=s+1}^{T} (y_t - \mathbb{E}[y_t|y_1, \ldots, y_{t-1}, x_1, \ldots x_t])^2}$$
## Out-of-sample prediction: RMSE

<table>
<thead>
<tr>
<th></th>
<th>Equity Premium</th>
<th>PCE Inflation</th>
<th>GDP Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RW</td>
<td>1.100</td>
<td>0.635</td>
<td>0.373</td>
</tr>
<tr>
<td>NGAR</td>
<td><strong>0.977</strong></td>
<td><strong>0.611</strong></td>
<td>0.410</td>
</tr>
<tr>
<td>DMA</td>
<td>1.01</td>
<td>0.660</td>
<td>0.422</td>
</tr>
<tr>
<td>TVD1</td>
<td>2.193</td>
<td>2.688</td>
<td>2.688</td>
</tr>
<tr>
<td>TVD2</td>
<td>0.986</td>
<td>0.623</td>
<td>0.481</td>
</tr>
<tr>
<td>TVD3</td>
<td>0.992</td>
<td>0.628</td>
<td>0.500</td>
</tr>
<tr>
<td>HierShrink</td>
<td>1.547</td>
<td>1.131</td>
<td>2.556</td>
</tr>
<tr>
<td>gprior¹</td>
<td>2.822</td>
<td>0.796</td>
<td>0.660</td>
</tr>
<tr>
<td>gprior²</td>
<td>1.648</td>
<td>0.712</td>
<td>0.588</td>
</tr>
<tr>
<td>gprior³</td>
<td>1.282</td>
<td>0.681</td>
<td>0.516</td>
</tr>
</tbody>
</table>

The window lengths for the three g-priors were 100 (gprior¹), 200 (gprior²) and 300 (gprior³) for the equity premium data and 50 (gprior¹), 70 (gprior²) and 90 (gprior³) for the inflation data.
• The DR model with NGAR process prior is the best performing approach for two data sets (equity premium and PCE inflation) and the second best performing for the GDP inflation data (with only the random walk giving better predictions).

• In general, the approaches which allow the complexity of the regression model to change over time (NGAR, TVD and DMA) outperform the other approaches (HierShrink and rolling window g-prior). This illustrates the importance of allowing time-variation in the relevance of regression coefficients.
In this talk, I’ve presented a method which allows the effects of predictors in a regression model to vary over time and which allows their value to be very close to zero for some proportion of time.
In this talk, I’ve presented a method which allows the effects of predictors in a regression model to vary over time and which allows their value to be very close to zero for some proportion of time.

Bayesian inference about this model allows this proportion to adapt to the data.
In this talk, I’ve presented a method which allows the effects of predictors in a regression model to vary over time and which allows their value to be very close to zero for some proportion of time.

Bayesian inference about this model allows this proportion to adapt to the data.

The method also allows some predictor to have their coefficients close to zero at all times (effectively removing the predictor).


