Forecasting with many predictors using message passing algorithms

Dimitris Korobilis

Department of Finance, Essex Business School, United Kingdom

University of Essex
ESSEX BUSINESS SCHOOL

Contribution of this paper

Purpose/contribution is two-fold:

1. Introduce to the econometrics literature machine learning methodologies for designing efficient Bayesian algorithms
   - Adopt the framework of "factor graphs", and use the Generalized Approximate Message Passing (GAMP) algorithm introduced in signal extraction and compressive sensing
   - Combine these algorithms with standard Bayesian "sparse learning" priors that induce shrinkage
2. Introduce to a novel interpretation and treatment of the time-varying parameter regression as a shrinkage problem.
   - Do not rely on state-space methods, rather use shrinkage to determine how "fast" or "slow" parameters should move.

Factor graphs

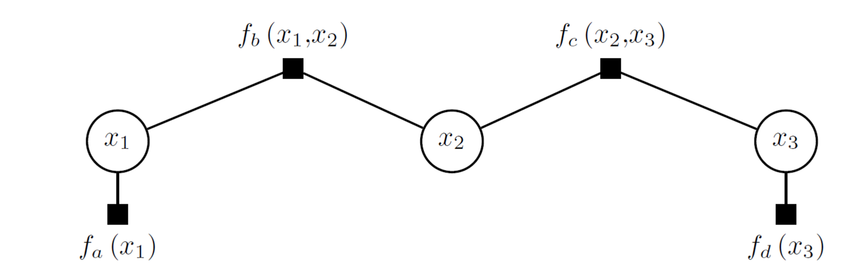
Starting point is factor graphs, message passing, and the sum-product algorithm

- ***Factor graph:*** *Bipartite graph that represents the way a global distribution of several random variables is decomposed into a product of simpler functions ("factors").*
- ***Message passing:*** *Dynamic programming solutions, where a node collects a result from a part of the graph and communicates it to the next neighboring node via a message.*
- ***Sum-product algorithm:*** *A rule specifying the way each node collects all messages in order to calculate the marginal distribution of that message.*

Simple example of factor graph:

Consider discrete variables $x = (x_1, x_2, x_3)$ and joint mass function p that can be decomposed as

$$p(x_1, x_2, x_3, x_4) = f_a(x_1) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_3), \quad (1)$$



Sum-product rule:

The message sent from variable x_i to factor node f_j is equal to the product of all messages arriving to node x_i except from the message coming from the target node f_j:

$$\mu_{x_i \to f_j} = \prod_{k \in N(x_i), k \neq j} \mu_{f_k \to x_i}, \quad (2)$$

where $N(x_i)$ is the set of neighboring (factor) nodes to x_i. Similarly, the message sent from factor node f_j to variable node x_i is given by the sum over the product of the factor function f_j itself and all the incoming messages, except the messages from the target variable node x_i:

$$\mu_{f_j \to x_i} = \sum_{x \setminus x_i} f_j(x) \prod_{l \in N(x_i), l \neq i} \mu_{x_l \to f_j}, \quad (3)$$

where $x \setminus x_i$ is the set x with the element x_i removed.

The marginal distribution of variable x_i is simply the product of all messages received only from factor nodes that are connected to x_i

$$p(x_i) \propto \prod_{m \in N(x_i)} \mu_{f_m \to x_i}. \quad (4)$$

Time-varying parameter regression as a high-dimensional problem

The starting point is the following time-varying parameter regression with stochastic volatility of the form

$$y_t = x_t \beta_t + \varepsilon_t \quad (5)$$

where y_t is variable of interest, $t = 1, ..., T$, x_t is a $1 \times p$ vector of predictors, β_t is a $p \times 1$ vector of coefficients, and $\varepsilon_t \sim N(0, \sigma_t^2)$. The "static regression" form of this model is

$$y = \mathfrak{X}\beta + \varepsilon, \quad (6)$$

where $y = [y_1, ..., y_T]'$ and $\varepsilon = [\varepsilon_1, ..., \varepsilon_T]'$ are column vectors stacking the observations y_t and ε_t respectively, $\beta = [\beta_1', ..., \beta_T']'$ is a $Tp \times 1$ vector, and $\mathfrak{X}$ is the following $T \times Tp$ matrix

$$\mathfrak{X} = \begin{bmatrix} x_1 & 0_{1\times p} & \dots & 0_{1\times p} & 0_{1\times p} \\ 0_{1\times p} & x_2 & \dots & 0_{1\times p} & 0_{1\times p} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0_{1\times p} & 0_{1\times p} & \dots & x_{T-1} & 0_{1\times p} \\ 0_{1\times p} & 0_{1\times p} & \dots & 0_{1\times p} & x_T \end{bmatrix}. \quad (7)$$

Estimation

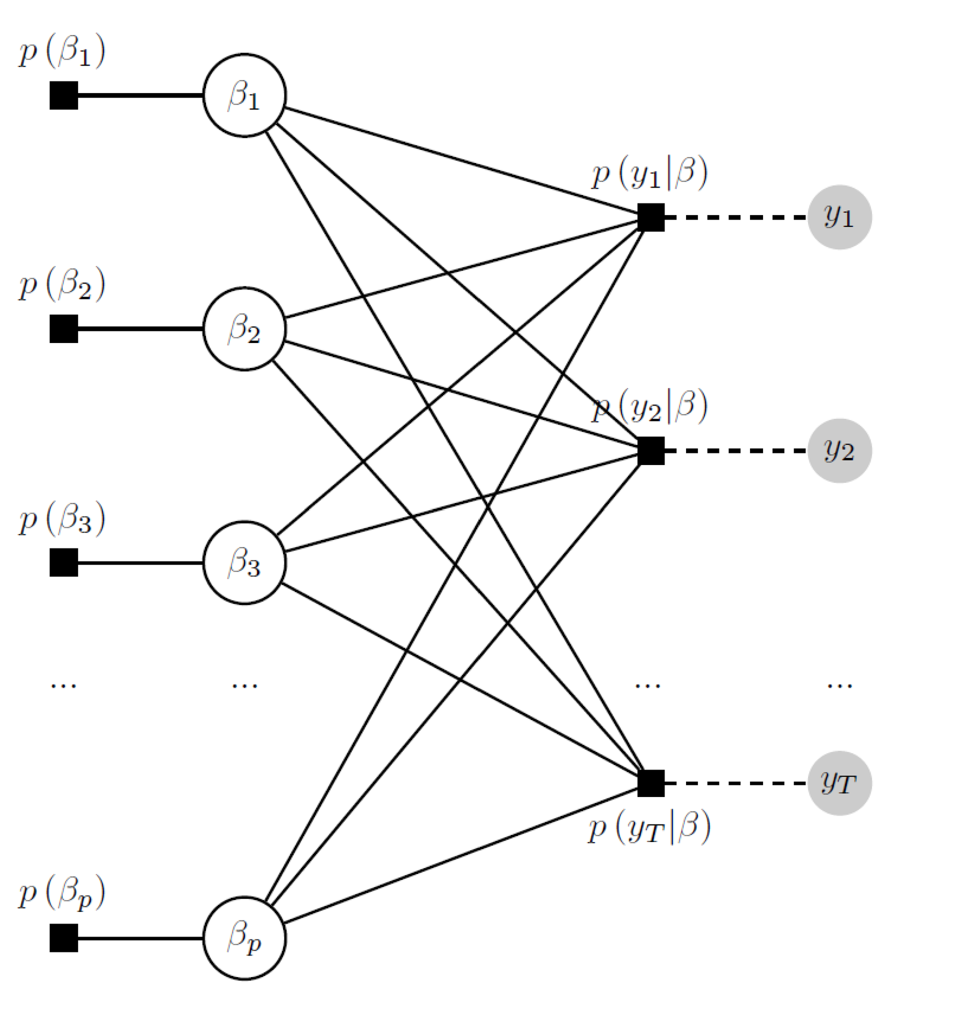
- The Gram matrix $(\mathfrak{X}'\mathfrak{X})$ is of rank $T \to$ OLS has not a unique solution
- **Standard approach:** Use 'hierarchical prior' $p(\beta_t|\beta_{t-1}) \sim N(\beta_{t-1}, Q)$
- **This paper argues:** estimate equation (6) using regularization/shrinkage!
- Number of predictors in $\mathfrak{X}$ grows both with p and T ($T = 700$ and $p = 50$ gives $q = 35000$ columns) $\to$ This is exactly where *message passing* inference comes handy.

Combine the "static regression" likelihood in (6) with the sparse Bayesian learning prior of Tipping (2001)

$$p(\beta_i|\alpha_i) = N\left(0, \alpha_i^{-1}\right), \quad (8)$$

$$p(\alpha_i) = Gamma\left(1e-10, 1e-10\right). \quad (9)$$

Figure 1: Factor graph for the posterior distribution of β



♠ We can now design the GAMP algorithm using the regression likelihood and the sparse Bayesian learning prior. Its output is the marginal posterior $p(\beta|y)$. Derivation of the algorithm is messy (see paper), but its worst case complexity is $\mathcal{O}(Tq)$ for q predictors!

Generic Form of Message Passing Algorithm

Initialize $\widehat{\beta}_j^{(0)} = 0$ and $\widehat{\tau}_j^{\beta,(0)} = 100 \; \forall j = 1, ..., q$, and set $\widehat{s}_t^{(0)} = 0 \; \forall t = 1, ..., T$.
$r = 1$
while $\|\widehat{\beta}^{(r)} - \widehat{\beta}^{(r-1)}\| \to 0$ **do**
  1) <u>OUTPUT MESSAGES STEP:</u>
  for $t = 1$ **to** T **do**
    $\widehat{c}_t^{(r)} = \sum_{j=1}^{q} \mathfrak{X}_{t,j} \widehat{\beta}_j^{(r-1)} - \widehat{s}_t^{(r-1)} * \widehat{\tau}_t^{c,(r)}$
    $\widehat{\tau}_t^{c,(r)} = \sum_{j=1}^{q} \mathfrak{X}_{t,j}^2 \widehat{\tau}_j^{\beta,(r-1)}$
    $\widehat{s}_t^{(r)} = g_{out}\left(\widehat{c}_t^{(r)}, \widehat{\tau}_t^{c,(r)}, y_t\right)$
    $\widehat{\tau}_t^{s,(r)} = -\frac{\partial}{\partial \widehat{c}} g_{out}\left(\widehat{c}_t^{(r)}, \widehat{\tau}_t^{c,(r)}, y_t\right)$
  end for
  2) <u>INPUT MESSAGES STEP:</u>
  for $j = 1$ **to** q **do**
    $\widehat{d}_j^{(r)} = \widehat{\beta}_j^{(r-1)} + \widehat{\tau}_j^{d,(r)} \sum_{t=1}^{T} \mathfrak{X}_{t,j} \widehat{s}_t^{(r)}$
    $\widehat{\tau}_j^{d,(r)} = \left(\sum_{t=1}^{T} \mathfrak{X}_{t,j}^2 \widehat{\tau}_t^{s,(r)}\right)^{-1}$
    $\widehat{\beta}_j^{(r)} = g_{in}\left(\widehat{d}_j^{(r)}, \widehat{\tau}_j^{d,(r)}\right)$
    $\widehat{\tau}_j^{\beta,(r)} = \widehat{\tau}_j^{d,(r)} \frac{\partial}{\partial \widehat{d}} g_{in}\left(\widehat{d}_j^{(r)}, \widehat{\tau}_j^{d,(r)}\right)$
  end for
  $r = r + 1$
end while
Obtain mean and variance of β as $\widehat{\beta} = \left(\widehat{\beta}_1^{(r)}, ..., \widehat{\beta}_q^{(r)}\right)$ and $\tau^{\beta} = \left(\widehat{\tau}_1^{\beta,(r)}, ..., \widehat{\tau}_q^{\beta,(r)}\right)$

Expressions for g_{out} and g_{in} depend on form of prior and likelihood, but are easy to derive.

Forecasting US Inflation

Forecasting model is of the form

$$\pi_{t+h}^h - \pi_t = \phi_{t,0} + f_t\theta_t(L) + \Delta\pi_t\gamma_t(L) + e_{t+h}, \quad (10)$$

using the FRED-MD data (i.e. forecast exercise a-la Stock and Watson (1999) JME).

Table 1: Forecast performance (MSFEs)

	CPI				PCE deflator			
	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
KP-AR	0.970	0.879	0.849***	0.834***	1.018	0.845***	0.806***	0.783***
GK-AR	0.999	1.008	1.009	1.005	0.999	0.996	1.005	0.999
TVP-AR	0.949	0.867***	0.828***	0.837***	1.010	0.793***	0.720***	0.732***
UCSV	1.027	0.970	0.911**	0.916*	1.064	0.841***	0.810***	0.761***
TVD	0.957	0.867***	0.862***	0.850***	1.015	0.787***	0.744***	0.742***
TVS	1.175	0.960	0.963	1.005	1.041	0.8578***	0.817***	0.814***
BMA	0.982*	0.588***	0.542***	0.531***	1.014	0.713***	0.663***	0.654***
TVP-BMA	1.090	0.770***	0.772**	0.629**	1.158	0.842**	0.798*	0.812
TVP-GAMP	0.923**	0.461***	0.421***	0.413***	0.982	0.614***	0.584***	0.565***

Model acronyms are as follows: **KP-AR:** Koop and Potter (2007) structural breaks AR(p) model; **GK-AR:** Giordani and Kohn (2008) structural breaks AR(p) model; **TVP-AR:** Pettenuzzo and Timmermann (2017) time-varying parameter AR(p) model; **UCSV:** Stock and Watson (2007) unobserved components stochastic volatility; **TVD:** Chan et al. (2012) time-varying dimension regression **TVS:** Kalli and Griffin (2014) time-varying sparsity regression **BMA:** George and McCulloch (1993) stochastic search variable selection regresison **TVP-BMA:** Groen et al. (2012) time-varying Bayesian model averaging model **TVP-GAMP:** Shrinkage representation of time-varying parameter regression, with message passing estimation. Next to MSFE values the results of the Diebold-Mariano statistic are presented, with * significance at the 10% level; ** at the 5% level; *** at the 1% level.

Table 2: Forecast performance (logPL)

	CPI				PCE deflator			
	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
KP-AR	0.060	0.135	-0.006	0.023	-0.033	0.071	0.044	0.016
GK-AR	-0.027	0.033	0.025	-0.027	-0.066	0.000	0.009	0.009
TVP-AR	0.216	0.095	0.045	0.071	0.068	0.157	0.116	0.118
UCSV	0.184	0.031	0.033	-0.002	0.051	0.065	0.062	0.081
TVD	-8.107	-2.665	-1.862	-1.859	-9.103	-2.887	-1.784	-1.559
TVS	0.032	0.154	0.100	0.058	0.004	0.149	0.167	0.103
BMA	0.019	0.303	0.279	0.292	-0.035	0.203	0.211	0.203
TVP-BMA	0.149	0.394	0.379	0.358	0.024	0.277	0.323	0.290
GAMP	0.017	0.528	0.422	0.381	0.046	0.258	0.279	0.266