Bayesian Analysis (2011) **6**, Number 4, pp. 639–664

Dynamic Financial Index Models: Modeling Conditional Dependencies via Graphs

Hao Wang*, Craig Reeson† and Carlos M. Carvalho‡

Abstract. We discuss the development and application of dynamic graphical models for multivariate financial time series in the context of Financial Index Models. The use of graphs generalizes the independence residual variation assumption of index models with a more complex yet still parsimonious alternative. Working with the dynamic matrix-variate graphical model framework, we develop general time-varying index models that are analytically tractable. In terms of methodology, we carefully explore strategies to deal with graph uncertainty and discuss the implementation of a novel computational tool to sequentially learn about the conditional independence relationships defining the model. Additionally, motivated by our applied context, we extend the DGM framework to accommodate random regressors. Finally, in a case study involving 100 stocks, we show that our proposed methodology is able to generate improvements in covariance forecasting and portfolio optimization problems.

Keywords: Bayesian forecasting; Covariance matrix forecasting; Dynamic matrix-variate graphical models; Index models, Factor models; Gaussian graphical models; Portfolio selection

1 Introduction

Since the seminal work of Sharpe (1964), Financial Index Models are in the core of asset pricing and portfolio allocation problems. These models assume that all systematic variation in the returns of financial securities can be explained linearly by one, or a set of market indices (factors). The central empirical implication of this assumption is a highly structured covariance matrix for the distribution of returns as, after conditioning on the chosen set of market indices, the residual covariance matrix is diagonal. The attractiveness of this approach is immediate as it offers a very simple, economically justifiable (Lintner 1965) and stable way to estimate potentially very large covariance

*Department of Statistics, University of South Carolina, Columbia, SC, wang345@mailbox.sc.edu
†Department of Statistics, Duke University, Durham, NC, craig.reeson@duke.edu
‡McCombs School of Business, The University of Texas, Austin, TX, carlos.carvalho@mccombs.utexas.edu

 DOI:10.1214/11-BA624

matrices. A vast body of literature is dedicated to testing the validity of Index Models and the selection of indices – we refer the reader to Cochrane (2001) for a detailed account of the area.

The covariance matrix of returns is a key input in building optimal portfolios and its estimation is often challenging as the number of parameters grows quadratically with the number of assets considered. Due to this high dimensionality of the parameter space (at times larger than the number of available observations) it is common to work with structured models that reduce the dimensionality of the problem and deliver more stable estimates and, in turn, better investment decisions. In this paper, we explore a generalization of Financial Index Models with more complex patterns of covariation between returns by allowing conditional dependencies via the introduction of graphical constraints. We work with the matrix-variate dynamic graphical model (DGM) framework of Carvalho and West (2007a,b) but, unlike their original work, graphs are used to increase complexity by adding non-zero elements to the off-diagonal part of the residual covariance matrix. Using graphical models as a tool for added complexity has the ability to do so in parsimonious ways as conditional independence constraints allow for covariance models of much lower dimensionality if compared to the otherwise unrestricted full covariance matrix.

This paper is intended as a case study exercise with contributions to empirical finance and statistics. From the finance point of view, our work shows that it is possible to improve upon traditional estimates of Index Models through dynamic matrix-variate graphical models. Our focus is not on the proposal of a new finance model but rather on the proper and effective implementation of important ideas of the field. In that sense, we show that DGMs provide a more flexible, efficient and still parsimonious strategy for estimating covariances that are fundamental to portfolio allocation problems.

As for the statistics contribution, we extend the DGM framework in two important ways: *(i)* we consider the problem of sequential inference about the graphical structure and, *(ii)* we define the sequential updating process in the presence of stochastic regressors.

The proposed forecasting model is tested on stock returns data in a portfolio selection exercise. Using 100 randomly selected domestic New York stock exchange (NYSE) monthly stock returns from 1989 through 2008, we find that our strategy yields better out-of-sample forecasts of realized covariance matrices and lower portfolio variances than the two traditional implementations of index models, the capital asset pricing model (CAPM) and the Fama-French (FF) model.

We start by describing Index Models in Section 2 along with their use in the dynamic linear model context. Sections 3 and 4 present the necessary background of dynamic matrix-variate graphical models. In Section 5 we discuss issues dealing with graph (model) uncertainty through time and a simulation study is presented in Section 6. Section 7 generalizes the DGM context to allow for random regressors. Finally, in Section 8 we explore the use of DGMs as a tool to improve the implementation of Financial Index Models.

2 Financial Index Models

A k-dimensional Index Model assumes that stock returns are generated by

$$Y_{it} = \alpha_i + \sum_{j=1}^{k} \theta_{ij,t} f_{jt} + \nu_{it}$$

where f_{jt} is the jth common factor at time t, and residuals ν_{it} are uncorrelated to index f_{jt} and to ν_{jl} for every j and every l. Under this class of models the covariance matrix of returns can be written as:

$$\mathbf{V}_t = \mathbf{\Theta}_t' \mathbf{\Psi}_t \mathbf{\Theta}_t + \mathbf{\Sigma}_t$$

where $\mathbf{\Theta}_t$ is the matrix of factor loadings of stocks, $\mathbf{\Psi}_t$ is the covariance matrix of the factors, and $\mathbf{\Sigma}_t$ is a diagonal matrix containing the residual return variances.

Some interesting Index Models include the single index model and three index model. The single index uses the excess return of the market as the single index. This model corresponds to the standard Capital Asset Pricing Model of Sharpe (1964). More recently, and perhaps the most commonly used approach is the three index model proposed by Fama and French (1993) where two new factors (besides the market) are added: value-weighted market index with size and book-to-market factors.

These models are usually estimated by running a set of independent regressions where the excess return of each stock is regressed against the indices for a certain window of time (Jagannathan and Wang 1996). Call $\hat{\boldsymbol{\theta}}_i$ the estimates of the regression coefficients for stock i and the $\hat{\sigma}_{ii}$ the residual variance estimate. This yields the following estimator for the covariance matrix of stock returns:

$$\hat{\mathbf{V}}_t = \hat{\mathbf{\Theta}}_t' \hat{\mathbf{\Psi}}_t \hat{\mathbf{\Theta}}_t + \hat{\mathbf{\Sigma}}_t,$$

where $\hat{\mathbf{\Psi}}$ is the sample covariance matrix of indices, $\hat{\mathbf{\Theta}} = [\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_p]$ is the matrix of regression coefficients for all p assets and $\hat{\mathbf{\Sigma}}$ is the diagonal matrix of residual variances.

This strategy usually defines the one-step forecast of the covariance matrix to be the current estimate of the covariance matrix $\hat{\mathbf{V}}$.

In our work, we recast the above strategy in a model based on a state-space or dynamic linear model (DLM) (West and Harrison 1997) representation. This follows the work of Zellner and Chetty (1965); Quintana and West (1987); Carvalho and West (2007a), to cite a few. We use a dynamic regression framework where, in its full generality, a $p \times 1$ vector time series of returns $\mathbf{Y}_t$ follows the dynamic linear model

$$\begin{aligned} \mathbf{Y}_t' &= \mathbf{F}_t'\boldsymbol{\Theta}_t + \boldsymbol{\nu}_t', & \boldsymbol{\nu}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{t}), & \quad (1) \\ \boldsymbol{\Theta}_t &= \boldsymbol{\Theta}_{t-1} + \boldsymbol{\Omega}_t & \boldsymbol{\Omega}_t \sim N(\mathbf{0}, \mathbf{W}_t, \boldsymbol{\Sigma}_\mathbf{t}), & \quad (2) \end{aligned}$$

for $t = 1, 2, \ldots$, with *(a)* $\mathbf{Y}_t = (Y_{ti})$, the $p \times 1$ observation vector; *(b)* $\boldsymbol{\Theta}_t = (\boldsymbol{\theta}_{ti})$, the $n \times p$ matrix of states; *(c)* $\boldsymbol{\Omega}_t = (\boldsymbol{\omega}_{ti})$, the $n \times p$ matrix of evolution innovations; *(d)* $\boldsymbol{\nu}_t = (\nu_{ti})$, the $p \times 1$ vector of observational innovations;*(e)* for all t, the $n \times 1$ regressor vector $\mathbf{F}_t$, is known. Also, $\boldsymbol{\Omega}_t$ follows a matrix-variate normal with mean 0, left and right covariance matrices $\mathbf{W}_t$ and $\boldsymbol{\Sigma}_\mathbf{t}$, respectively. In terms of scalar elements, we have p univariate models with individual n-vector state parameters, namely

$$\begin{aligned} \text{Observation:} \quad Y_{ti} &= \mathbf{F}_t'\boldsymbol{\theta}_{ti} + \nu_{ti}, & \nu_{ti} \sim N(0, \sigma_{ii,t}^2), & \quad (3) \\ \text{Evolution:} \quad \boldsymbol{\theta}_{ti} &= \boldsymbol{\theta}_{t-1,i} + \boldsymbol{\omega}_{ti}, & \boldsymbol{\omega}_{ti} \sim N(0, \mathbf{W}_t\sigma_{ii,t}^2), & \quad (4) \end{aligned}$$

for each i, t. Each of the scalar series shares the same $\mathbf{F}_t$ elements, and the reference to the model as one of exchangeable time series reflects these symmetries. This is a standard specification in which the correlation structures induced by $\boldsymbol{\Sigma}_t$ affect both the observation and evolution errors; for example, if $\sigma_{ij,t}$ is large and positive, vector series i and j will show concordant behavior in movement of their state vectors and in observational variation about their levels. Specification of the entire sequence of W_t in terms of discount factors (Harrison and Stevens 1976; Smith 1979; West and Harrison 1997) is also standard practice, typically using discount factors related to the state vector and their expected degrees of random change in time.

The above representation provides Kalman filter like (Kalman 1960) sequential, closed-form analytical updates of the one-step ahead forecast distributions of future returns and posterior distributions for states and parameters defining the model. This allows for proper accounting of the uncertainty associated with all necessary inputs in sequential investment decisions.

According to traditional Index Models, $\boldsymbol{\Sigma}_t$ is a diagonal matrix as all common variation between returns should be captured by the elements in $\boldsymbol{\Theta}_t$. We will depart from

this standard assumption and allow for a more flexible representation of the residual covariance matrix leading to potentially more complex forms of $\mathbf{V}_t$. This is done via the introduction of conditional independencies determined by graphical constraints in $\mathbf{\Sigma}_t$. The use of these models in sequential portfolio problems was originally proposed by Carvalho and West (2007a) and further analyzed by Quintana et al. (2009). In both references however, graphs were used to reduce the dimensionality of an otherwise fully unstructured covariance matrix of returns. Here, we come from a different direction and show that graphs can be successfully used to increase the complexity of an otherwise highly structured covariance matrix. Working with conditional independence constraints, we strike a parsimonious compromise between the diagonal matrix and the high-dimensional fully unconstrained covariance model. Before continuing, we need to define the necessary notation for the introduction of graphical models in DLMs.

3 Gaussian graphical model

Graphical model structuring characterizes conditional independencies via graphs (Lauritzen 1996; Jones et al. 2005), and provides methodologically useful decompositions of the sample space into subsets of variables (graph vertices) so that complex problems can be handled through the combination of simpler elements. In high-dimensional problems, graphical model structuring is a key approach to parameter dimension reduction and, hence, to scientific parsimony and statistical efficiency when appropriate graphical structures are identified.

In the context of a multivariate normal distribution, conditional independence restrictions are simply expressed through zeros in the off-diagonal elements of the precision (or concentration) matrix. Define a $p-$vector $\mathbf{x}$ with elements x_i and zero-mean multivariate normal distribution with $p \times p$ variance matrix $\mathbf{\Sigma}$ and precision $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ with elements ω_{ij}. Write $G = (V, E)$ as the undirected graph whose vertex set V corresponds to the set of p random variables in $\mathbf{x}$, and edge set E contains elements (i, j) for only those pairs of vertices $i, j \in V$ for which $\omega_{ij} \neq 0$. The canonical parameter $\mathbf{\Omega}$ belongs to $M(G)$, the set of all positive-definite symmetric matrices with elements equal to zero for all $(i, j) \notin E$.

The density of $\mathbf{x}$ factorizes as

$$p(\mathbf{x}|\mathbf{\Sigma}, G) = \frac{\prod_{P \in \mathcal{P}} p(\mathbf{x}_P|\mathbf{\Sigma}_P)}{\prod_{S \in \mathcal{S}} p(\mathbf{x}_S|\mathbf{\Sigma}_S)}, \tag{5}$$

a ratio of products of densities where $\mathbf{x}_P$ and $\mathbf{x}_S$ indicate subsets of variables in the

prime components (P) and separators (S) of G, respectively. Given G, this distribution is defined completely by the component-marginal covariance matrices $\mathbf{\Sigma}_P$, subject to the consistency condition that sub-matrices in the separating components are identical (Dawid and Lauritzen 1993). That is, if $S = P_1 \cap P_2$ the elements of $\mathbf{\Sigma}_S$ are common in $\mathbf{\Sigma}_{P_1}$ and $\mathbf{\Sigma}_{P_2}$.

A graph is said to be decomposable when all of its prime components are complete subgraphs of G, implying no conditional independence constraints within a prime component; we also refer to all prime components (as well as their separators) as cliques of the graph. Due to its mathematical and computational convenience, we will only consider decomposable graphs. In this context, the sequential updating and model assessment procedures remain tractable, especially in the high-dimensional settings. It is also our experience and belief that this restriction is not severe as the space of decomposable graphs is very large allowing for the necessary flexibility to our modeling goals. We now briefly review the theory of hyper-inverse Wishart distributions and its extensions to DLMs.

The fully conjugate Bayesian analysis of decomposable Gaussian graphical models is based on the family of *hyper-inverse Wishart* (HIW) distributions for structured variance matrices (Dawid and Lauritzen 1993). If $\mathbf{\Omega} = \mathbf{\Sigma}^{-1} \in M(G)$, the hyper-inverse Wishart

$$\mathbf{\Sigma} \sim HIW_G(b, \mathbf{D}) \tag{6}$$

has a degree-of-freedom parameter b and location matrix $\mathbf{D} \in M(G)$. This distribution is the unique hyper-Markov distribution for $\mathbf{\Sigma}$ with consistent clique-marginals that are inverse Wishart. Specifically, for each clique $P \in \mathcal{P}$, $\mathbf{\Sigma}_P \sim IW(b, \mathbf{D}_P)$ with density

$$p(\mathbf{\Sigma}_P | b, \mathbf{D}_P) \propto |\mathbf{\Sigma}_P|^{-(b+2|P|)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_P^{-1}\mathbf{D}_P)\right) \tag{7}$$

where $\mathbf{D}_P$ is the positive-definite symmetric diagonal block of $\mathbf{D}$ corresponding to $\mathbf{\Sigma}_P$. The full HIW is conjugate to the likelihood from a Gaussian sample with variance $\mathbf{\Sigma}$ on G, and the full HIW joint density factorizes over cliques and separators in the same way as (2); that is,

$$p(\mathbf{\Sigma}|b, \mathbf{D}) = \frac{\prod_{P\in\mathcal{P}} p(\mathbf{\Sigma}_P|b, \mathbf{D}_P)}{\prod_{S\in\mathcal{S}} p(\mathbf{\Sigma}_S|b, \mathbf{D}_S)},$$

where each component in the products of both numerator and denominator is IW as in equation (7). Finally, both the expected value of $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ can be obtained in closed form following the results in Rajaratnam et al. (2008) and Jones et al. (2005) respectively.

4 Dynamic matrix-variate graphical model

The matrix-variate graphical model framework combines HIW distributions together with matrix and multivariate normal distributions, in a direct and simple extension of the usual Gaussian-inverse Wishart distribution theory to the general framework of graphical models. The $n \times p$ random matrix $\mathbf{X}$ and $p \times p$ random variance matrix $\mathbf{\Sigma}$ have a joint matrix-normal, hyper-inverse Wishart (NHIW) distribution if $\mathbf{\Sigma} \sim HIW_G(b, \mathbf{D})$ on G and $(\mathbf{X}|\mathbf{\Sigma}) \sim N(\mathbf{m}, \mathrm{W}, \mathbf{\Sigma})$ for some $b, \mathbf{D}, \mathbf{m}, \mathrm{W}$. We denote this by $(\mathbf{X}, \mathbf{\Sigma}) \sim NHIW_G(\mathbf{m}, \mathrm{W}, b, \mathbf{D})$ with $\mathbf{X}$ marginally following a matrix hyper-T (as defined in Dawid and Lauritzen 1993) denoted by $HT_G(\mathbf{m}, \mathrm{W}, \mathbf{D}, b)$.

Back to the DGM context and given $\mathbf{\Sigma}_t$ constrained by any decomposable graph G, Carvalho and West (2007a,b) define the details of the full sequential and conjugate updating, filtering and forecasting for the dynamic regressions and time-varying $\mathbf{\Sigma}_t$. This approach incorporates graphical structuring into the traditional matrix-variate DLM context and provides a parsimonious yet tractable model for $\mathbf{\Sigma}_t$. Consider the matrix normal DLM described in equations (1) and (2). With the usual notation that $D_t = \{D_{t-1}, \mathbf{Y}_t\}$ is the data and information set upon any time t, assume the NHIW initial prior of the form

$$(\mathbf{\Theta}_0, \mathbf{\Sigma}_0 \mid D_0) \sim NHIW_G(\mathbf{m}_0, \mathbf{C}_0, b_0, \mathbf{S}_0). \tag{8}$$

In components, $(\mathbf{\Theta}_0 \mid \mathbf{\Sigma}_0, D_0) \sim N(\mathbf{m}_0, \mathbf{C}_0, \mathbf{\Sigma}_0)$ and $(\mathbf{\Sigma}_0 \mid D_0) \sim HIW_G(b_0, \mathbf{S}_0)$, which incorporates the conditional independence relationships from G into the prior. For now assume full knowledge of G defining the conditional independence relationships in $\mathbf{Y}$. Full sequential updating can be summarized in the following Theorem 1.

Theorem 1. *(Carvalho and West 2007a,b) Under the initial prior of equation (8) and with data observed sequentially to update information sets D_t the sequential updating for the matrix normal DGM on G is given as follows:*

(i) *Posterior at $t-1$:* $(\mathbf{\Theta}_{t-1}, \mathbf{\Sigma}_{t-1} \mid D_{t-1}) \sim NHIW_G(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}, b_{t-1}, \mathbf{S}_{t-1})$

(ii) *Prior at t:* $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid D_{t-1}) \sim NHIW_G(\mathbf{a}_t, \mathbf{R}_t, \delta b_{t-1}, \delta \mathbf{S}_{t-1})$ *where* $\mathbf{a}_t = \mathbf{m}_{t-1}$ *and* $\mathbf{R}_t = \mathbf{C}_{t-1} + \mathbf{W}_t$

(iii) *One-step forecast:* $(\mathbf{Y}_t \mid D_{t-1}) \sim HT_G(\mathbf{f}_t, q_t \delta \mathbf{S_{t-1}}, \delta b_{t-1})$ *where* $\mathbf{f}_t' = \mathbf{F}_t' \mathbf{a}_t$ *and* $q_t = \mathbf{F}_t' \mathbf{R}_t \mathbf{F}_t + 1$

(iv) *Posterior at t:* $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid D_t) \sim NHIW_G(\mathbf{m}_t, \mathbf{C}_t, b_t, \mathbf{S}_t)$ *with* $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t \mathbf{e}_t'$,

$\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{A}'q_t$, $b_t = \delta b_{t-1} + 1$, $\mathbf{S}_t = \delta\mathbf{S}_{t-1} + \mathbf{e}_t\mathbf{e}_t'/q_t$ *where* $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t/q_t$ *and* $\mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$.

The above derivation uses a "locally smooth" discount factor-based model to allow $\mathbf{\Sigma}_t$ to vary stochastically. This is a common approach in dynamic linear models (Quintana et al. 2003) where information is discounted through time by a pre-specified discount factor δ. This provides sequential estimates of $\mathbf{\Sigma}_t$ that keep adapting to new data while further discounting past observations. This is easily seen in the representation of the posterior harmonic mean that has the form of an exponentially weighted moving average estimate defined as

$$\hat{\mathbf{\Sigma}}_t \approx (1-\delta)\sum_{l=0}^{t-1}\delta^l\mathbf{e}_{t-l}\mathbf{e}_{t-l}'.$$

In practical terms the choice of δ represents a similar problem as the choice of the data window in the usual estimation of index models. Extensive discussion of choice of δ in dynamic variance models appears in Chapter 16 of West and Harrison (1997).

So far, G was assumed known and held fixed for all t. This is clearly a limitation of the framework of Carvalho and West (2007a) as it is not necessarily the case that the same set of conditional independence constraints remain fixed across time. Moreover, it is rarely the case that knowledge about G is available and data driven approaches to determine G are required which represents a non-trivial question in empirical applications. Carvalho and West (2007a) present one example where graphs were selected via the computationally intensive stochastic search ideas of Jones et al. (2005). Quintana et al. (2009) consider similar strategies and briefly explore the issue of time variation in G when modeling currencies. Before continuing in our exploration of the use of graphs in index models, we add to this discussion and consider alternatives to learn about the conditional independence relationships defining the models.

5 Graphical model uncertainty and search

5.1 Marginal likelihood over Graphs

In the standard static context, from a Bayesian perspective, model selection involves the posterior distribution of graphs, given by:

$$p(G|\mathbf{x}) \propto p(\mathbf{x}|G)p(G)$$

where $p(\mathbf{x}|G)$ is the marginal likelihood of G. The marginal likelihood function for any graph G is computed by integrating out the covariance matrix with respect to the prior

$$p(\mathbf{x}|G) = \int_{\mathbf{\Sigma}^{-1}\in M(G)} p(\mathbf{x}|\mathbf{\Sigma}, G)p(\mathbf{\Sigma}|G)\mathrm{d}\mathbf{\Sigma}$$

where $M(G)$, as before,indicates the set of all positive-definite symmetric matrices constrained by G.

Under a hyper-inverse Wishart prior for $\mathbf{\Sigma}$ and observed data $\mathbf{x}$ of sample size n, the above integration for a decomposable graph becomes a simple function of the prior and posterior normalizing constants, $H(b, \mathbf{D}, G)$ and $H(b+n, \mathbf{D}+\mathbf{S}_{\mathbf{x}}, G)$:

$$p(\mathbf{x}|G) = (2\pi)^{-np/2}\frac{H(b, \mathbf{D}, G)}{H(b+n, \mathbf{D}+\mathbf{S}_{\mathbf{x}}, G)}$$

where the normalizing constant $H(b, \mathbf{D}, G)$ is given by

$$H(b, \mathbf{D}, G) = \frac{\prod_{P\in\mathcal{P}} |\frac{\mathbf{D}_P}{2}|^{(\frac{b+|P|-1}{2})}\Gamma_{|P|}(\frac{b+|P|-1}{2})^{-1}}{\prod_{S\in\mathcal{S}} |\frac{\mathbf{D}_S}{2}|^{(\frac{b+|S|-1}{2})}\Gamma_{|S|}(\frac{b+|S|-1}{2})^{-1}}, \tag{9}$$

with $\Gamma_k(a)$ the multivariate gamma function.

In the dynamic set up, a fully Bayesian analysis will consider the graph predictive probability of $\pi(G \mid D_{t-1})$ over $\mathcal{G}$, the set of all decomposable graphs, and specify the unconditional predictive distribution $p(\mathbf{Y}_t \mid D_{t-1})$ as $\mathrm{E}_G\{p(\mathbf{Y}_t \mid D_{t-1}, G)\}$ with the expectation taken with respect to $p(G \mid D_{t-1})$, namely,

$$(\mathbf{Y}_t \mid D_{t-1}) \sim \sum_{G\in\mathcal{G}} \pi(G \mid D_{t-1})p(\mathbf{Y}_t \mid D_{t-1}, G). \tag{10}$$

Equation (10) indicates that the predictive probability $\pi(G \mid D_{t-1})$ is central to evaluating the predictive distribution $p(\mathbf{Y}_t \mid D_{t-1})$. The two possibilities for consideration of predicting G are as follows: *(i)* fixed graph for all t, that is for some $G \in \mathcal{G}$, DLM(G) holds for all t; *(ii)* time varying graphs where for some possible sequence of graphs $G_t \in \mathcal{G}$,$(t = 1, 2, \cdots)$, DLM(G_t) holds at time t.

For *(i)*, the predictive probability of graphs for time t is defined as

$$\pi(G \mid D_{t-1}) = p(G \mid D_{t-1}) \propto p(G)p(\mathbf{Y}_{1:t-1} \mid G) \tag{11}$$

where the marginal likelihood of a DLM on any graph G is

$$p(\mathbf{Y}_{1:t-1}|G) = p(\mathbf{Y}_{t-1}|D_{t-2}, G)p(\mathbf{Y}_{t-2}|D_{t-3}, G)\ldots p(\mathbf{Y}_1|D_0, G),$$

with each element in the product, $(\mathbf{Y}_t|D_{t-1}, G) \sim HT_G(\mathbf{f}_t, \mathbf{S_{t-1}}, b_{t-1})$ as defined in Theorem 1.

For *(ii)*, the time dependence is made explicit with time subscripts, so that a graph G_j at time t is $G_{t,j}$. Denote $\pi(G_{t,j} \mid D_{t-1})$ as the predictive probability at time $t-1$ for graph G_j. It is natural to dynamic modeling that, as time progresses, what occurred in the past becomes less and less relevant to inference made for the future. Applying this notion to graphs, past data loses relevance to current graphs as t increases. Once again, one practical possibility is to use a discount factor to reduce the impact of past information to current inferences, similarly to the discounting ideas used in modeling $\mathbf{\Sigma}_t$. The predicted probability of $G_{t,j}$ for time t at time $t-1$ could be written as

$$\pi(G_{t,j} \mid D_{t-1}) \propto \frac{H(b_0, \mathbf{S}_0, G_{t,j})}{H(\delta b_{t-1}, q_t \delta \mathbf{S}_{t-1}, G_{t,j})} \pi_0(G_{t,j}). \tag{12}$$

To provide insights into the nature of the predicted probability (12), suppose the graph has the prior (12) at time $t-1$. After observing $\mathbf{Y}_t$, this prior updates to the posterior via the usual updating equations:

$$\begin{aligned} \pi(G_{t,j} \mid D_t) &\propto p(\mathbf{Y}_t \mid D_{t-1}, G_{t,j}) \pi(G_{t,j} \mid D_{t-1}) \\ &\propto \frac{H(\delta b_{t-1}, q_t \delta \mathbf{S}_{t-1}, G_{t,j})}{H(b_t, q_{t+1} \mathbf{S}_t, G_{t,j})} \frac{H(b_0, \mathbf{S}_0, G_{t,j})}{H(\delta b_{t-1}, q_t \delta \mathbf{S}_{t-1}, G_{t,j})} \pi_0(G_{t,j}) \\ &= \frac{H(b_0, \mathbf{S}_0, G_{t,j})}{H(b_t, q_{t+1} \mathbf{S}_t, G_{t,j})} \pi_0(G_{t,j}). \end{aligned} \tag{13}$$

This has the same representation as equation (12), i.e. a ratio of two normalizing constants of hyper-inverse Wishart distributions, but updated location parameter and degrees of freedom, $\mathbf{S}_t = \delta \mathbf{S}_{t-1} + \mathbf{Y}_t \mathbf{Y}_t'$ and $b_t = \delta b_{t-1} + 1$. Substituting $t+1$ for t in (12) we obtain the prior probability for $\pi(G_{t+1,j} \mid D_t)$ at t as follows:

$$\pi(G_{t+1,j} \mid D_t) \propto \frac{H(b_0, \mathbf{S}_0, G_{t+1,j})}{H(\delta b_t, q_{t+1} \delta \mathbf{S}_t, G_{t+1,j})} \pi_0(G_{t+1,j}). \tag{14}$$

In comparison with equation (13), the above has a discount factor δ to model a decay of information between time t and $t+1$ in a way analogous to the standard use of discount factors in DLMs. The maintenance of the normalizing constant ratio prior and posterior probability at each time enables continued, easy sequential updating, with the minor modification that the degrees of freedom b_t are discounted successively.

This model implies that the most recent exponentially weighted residual covariance matrix $\mathbf{S}_{t-1}$ could predict both the one-step ahead residual graphical structure and the residual covariance matrix.

Let M_G be the graph predicting model that takes its value in the set $\{M_F, M_C\}$ where M_F represents the fixed graph predicting model as described by equation (11) and M_C represents the time-varying graph predicting model as described by equation (12). Let α be the pair of predicting model M_G and discount factor δ, that is, $\alpha = (M_G, \delta)$. We can choose α, the predicting model and discount factors, using the marginal likelihood across our candidate models:

$$p(\mathbf{Y}_t \mid D_{t-1}, \alpha) \quad = \quad \sum_{G_{t,j} \in \mathcal{G}} p(\mathbf{Y}_t \mid D_{t-1}, \alpha, G_{t,j}) \pi(G_{t,j} \mid D_{t-1}, \alpha). \tag{15}$$

One note regarding the identification of the underlying graphical structure: the DGM as defined by Carvalho and West (2007a) is identifiable as it is simply a multivariate dynamic regression model with a fixed set of restrictions in the innovation covariance. Our approach has the exact same structure and uses graphs as a regularization tool that allows for the mixing over a different set of constraints at each point in time. The discount framework effectively creates a "rolling window" of data which is then used to evaluate the marginal likelihood of each candidate graph. Note however that this approach cannot deliver precise inferences about G as, at each point in time, we only evaluate a small subset of elements in graph space.

5.2 Sequential stochastic search

Regardless of the choice of how to model G in time the model selection problem gets further complicated by the explosive combinatorial nature of the space of possible graphs. Without the restriction of decomposability there are $2^{\binom{p}{2}}$ elements in graph space, where p represents the number of vertices. Decomposability accounts for approximately 10% of this number (Jones et al. 2005) which is still impossible to enumerate for moderate size p. Any attempt to deal with these models requires the development of efficient computational tools to explore the model space. Here, we propose an extension to the shotgun stochastic search (SSS) of Jones et al. (2005) to sequentially learn $(G_{t,j} \mid D_{t-1})$. In a nutshell, our analysis generates multiple graphs at each time t from the predictive probability $\pi(G_{t,j} \mid D_{t-1})$, using SSS.

Suppose that, at time $t-1$, we have saved a sample of the top N graphs $G_{t-1,j}, j = 1, \cdots, N$ with highest predictive probabilities $\pi(G_{t,j} \mid D_{t-1})$. Proceeding to time t, we adopt the following search algorithm.

1. Evaluate the new predictive probability $\pi(G_{t+1,j} \mid D_t)$ of these N graphs from time $t-1$;

2. From among the N graphs, propose the ith graph as a new starting graph with probability proportional to $\pi(G_{t+1,j} \mid D_t)^c$, where c is an annealing parameter;

3. Start with $G_{t+1,j}$ and apply SSS. After each stage of SSS, compute the Bayesian model average (BMA) estimator of a predicted quantity of interest, e.g. predictive covariance matrix, using the current top N graphs;

4. Stop the search when certain distance between the last two BMA estimates is below a small number, set $t = t+1$ and return to (1).

The evaluation and resample steps of (1) and (2) are important because top graphs from the previous step still represent the majority of our knowledge and should be good starting points for a new SSS once a new data sample becomes available. We use the saved top N graphs to estimate the posterior graph probability and the covariance matrix. We do not consider the dilution effect caused by large model spaces, as we expect these saved high probability graphs to be sufficient to generate good covariance matrix estimates compared with those produced by empty graphs. In our two examples, we report the results based on $N = 1000$. We observed similar results when repeating the analysis with $N = 2000$ and 5000.

6 A simple example

To focus the idea of sequential learning in dynamic graphical models, we first consider a simple local trend DLM, namely

$$\begin{aligned} \mathbf{Y}_t &= \boldsymbol{\theta}_t + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim N(0, \boldsymbol{\Sigma}_t), \\ \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(0, W_t \boldsymbol{\Sigma}_t). \end{aligned}$$

This is a special case of the general DLMs presented in previous sections. We extend the example in Carvalho and West (2007a) where data from $p = 11$ international currency exchange rates relative to the US dollar is analyzed. In all models, we use fairly diffuse priors, $\mathbf{m}_0 = \mathbf{0}, \mathbf{C}_0 = 100, b_0 = 3$ and $\mathbf{S}_0 = 0.0003\mathbf{I_{11}}$, and annealing parameter $c = 1$. In addition, we use the 100 data points as a "training set" to define the prior at time zero.

We ran a set of parallel analyses differing through the value δ, with δ values of $0.93, 0.95, 0.97$ and 0.99, and graph predicting models, with $M_G = M_F$ as described by equation (11), and $M_G = M_C$ as described by equation (12). At each of the eight pairs of (M_G, δ), and time t, the marginal likelihood of equation (15) is approximated by summing over the top 1000 graphs at each time t, resulting in a full marginal likelihood function of (M_G, δ). Figure 1 displays the plots of log Bayes factors against the model $(M_F, 0.95)$ over time. When comparing Bayes factors within each δ, Figure 1 shows that all four time-varying graphs generate smaller Bayes factors than their fixed graph peers. Figure 2 highlights the change of the relative cumulative log Bayes factors of the top two models. Overall, the chosen maximum marginal likelihood estimation from such analysis is $(M_F, 0.97)$ over the period up to the end of 08/1992 and $(M_F, 0.95)$ over the period from then until the end of data at 06/1996. The change from $\delta = 0.97$ to $\delta = 0.95$ at the end of 08/1992 reflects a more adaptive model being favored since then. The occurrence of one or two rather marked changes of relative predictive density may be due to major economic changes and events. A key such event was Britain's withdrawal from the EU exchange rate agreement (ERM) in September 1992 and this led to the deviation from the steady behavior anticipated under a model with relatively high discount factor 0.97 to the more adaptive 0.95. A second period of change of structures occurred in early 1995 with major changes in Japanese interest rate policies as a response to a weakening Yen and a move toward financial restructuring in Japan. The more adaptive model $(M_F, 0.93)$ has lower likelihood than $(M_F, 0.95)$ and $(M_F, 0.97)$, because the corresponding one-step forecast distributions are too diffuse. Given the significant changes in likelihood for a small change of δ, it would be useful to draw fully Bayesian inference on δ. However, this additional sampling step is not trivial since the likelihood function of δ is not a standard form, and is beyond the scope of the current paper.

Each value in Figure 3 represents the number of times a given edge has inclusion probability greater than 0.5 out of four different time points. As can be seen, over time, graphs have several persistent signals - edges that are significant at all four time points.

This example serves to illustrate some features of inference with dynamic graphical models. In each of the DGMs (M_G, δ), and for any specified sequence of graphs $\{G_t\}$, the prior, posterior, and forecast distributions are all standard distributions of well-understood forms, whether they be hyper-inverse Wishart or hyper T. Forecasts that take into account graph uncertainties are easily calculated from the finite mixture of hyper T distributions of equation (15). If one is concerned about which are the best

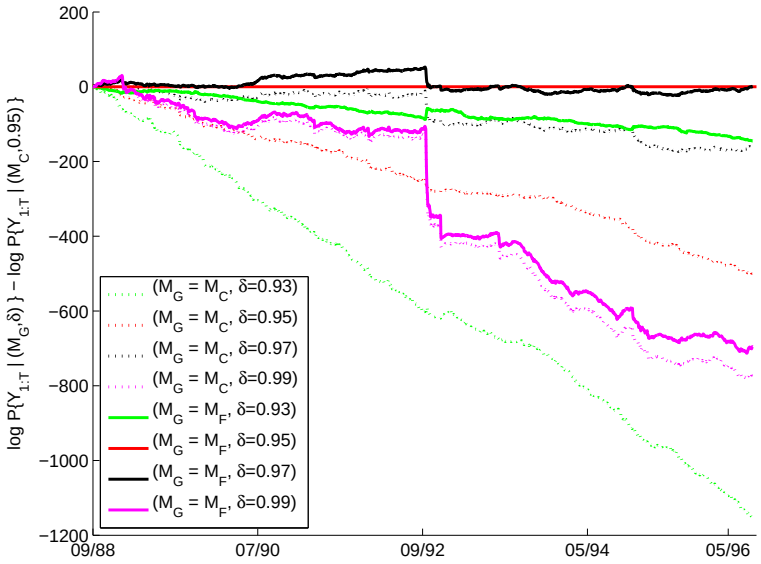

Figure 1: Relative cumulative log predictive density over time under the baseline model $(M_F, 0.95)$. The figure illustrates that the predictive density of time-varying graphs (dashed lines) are generally smaller than those of fixed graphs (solid lines).

graph predicting models or which discount factors to use, their corresponding cumulative marginal densities may be used to choose these specifications.

The two proposed graph predicting models together with the covariance matrix discount factors allow us to separately infer the dynamics of graphs and the dynamics of covariance matrices. In this particular example the marginal likelihoods favor static models M_F for all values of δ. This suggests that time-varying graphs inferred by a moving window may not produce consistently better predictions than fixed graphs with signals detected sequentially using all historical data. Although the fixed graphs are generally preferred over time-varying graphs for the same δ, the covariance matrix itself seems to be time-varying even when the graphs are fixed. This is because models with $(M_F, 0.95)$ and $(M_F, 0.97)$ are supported by data as is evident in Figure 1, indicating that Σ_t is time-varying. The dynamic of these time-varying Σ_t is specified as the matrix Beta-Bartlett HIW evolution by Carvalho and West (2007a).

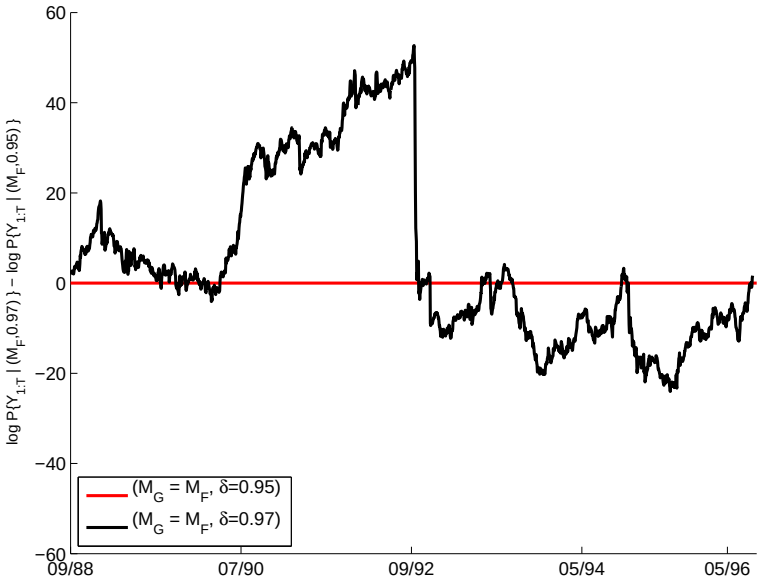

Figure 2: Relative cumulative log predictive density of model $(M_F, 0.97)$ under the baseline model $(M_F, 0.95)$.

7 Random regression vector DLM

Our applied and modeling interests are motivated by models where we attempt to predict $\mathbf{Y}_t$ with a regression vector $\mathbf{F}_t$ that is random and unknown before time t. For example, the simplest Index Model, the CAPM, has the market portfolio as $\mathbf{F}_t$, a variable that has to be predicted before the predictive covariance structure of stocks can be evaluated through our DGM framework. Now, let $I_t = \{\mathbf{Y}_1, \cdots, \mathbf{Y}_t, \mathbf{F}_1, \cdots, \mathbf{F}_t\}$ denote the data and information set. Assume $\mathbf{F}_t$ has a prior $p(\mathbf{F}_t \mid I_{t-1})$ at time t. Then under the assumption that the priors of $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t)$ and $\mathbf{F}_t$ are conditionally independent given I_{t-1}, namely, $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t) \perp\!\!\!\perp \mathbf{F}_t \mid I_{t-1}$, the following results apply.

Theorem 2. *Under the initial prior of equation (8) and with data observed sequentially to update information sets I_t the sequential updating for the matrix normal DLM on G is given as follows:*

(i) *Posterior at $t-1$:* $(\mathbf{\Theta}_{t-1}, \mathbf{\Sigma}_{t-1} \mid I_{t-1}) \sim NHIW_G(\mathbf{m}_{t-1}, \mathbf{C}_{t-1}, b_{t-1}, \mathbf{S}_{t-1})$.

(ii) *Prior at t:* $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid I_{t-1}) \sim NHIW_G(\mathbf{a}_t, \mathbf{R}_t, \delta b_{t-1}, \delta \mathbf{S}_{t-1})$ *where* $\mathbf{a}_t = \mathbf{m}_{t-1}$ *and* $\mathbf{R}_t = \mathbf{C}_{t-1} + \mathbf{W}_t$.

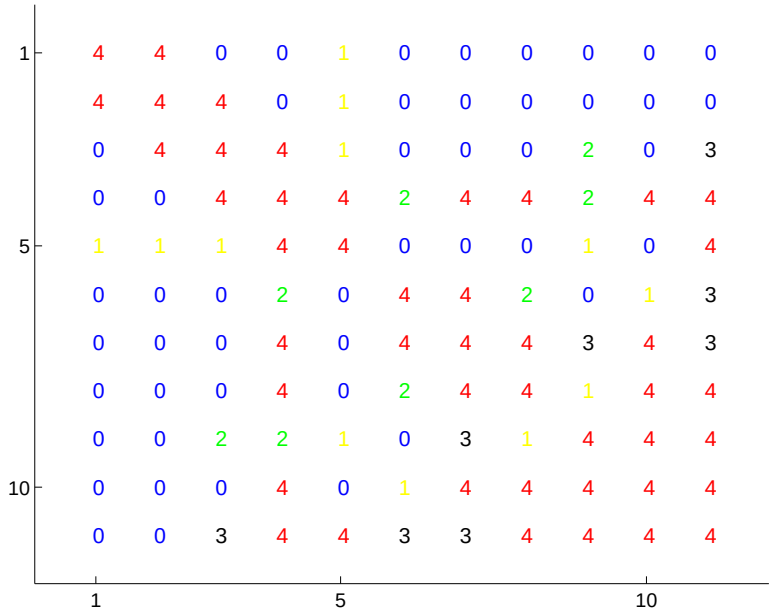

Figure 3: Persistent signals over time. Each value represents the number of times a given edge has inclusion probability greater than 0.5 out of four different time points at 09/1988, 09/1992, 05/1994 and 05/1996 respectively.

(iii) *One-step forecast:* $p(\mathbf{Y}_t \mid I_{t-1}) = \int HT_G(\mathbf{f}_t, q_t\delta\mathbf{S_{t-1}}, \delta b_{t-1})p(\mathbf{F}_t \mid I_{t-1})d\mathbf{F}_t$ *with first two moments:*

$$\begin{aligned}\mathbf{r}_{t,G} &\equiv E(\mathbf{Y}_t \mid I_{t-1}) = \mathbf{a}_t'\mu_{\mathbf{F}_t}\\ \mathbf{Q}_{t,G} &\equiv \mathrm{cov}(\mathbf{Y}_t \mid I_{t-1}) = \mathbf{a}_t'\mathbf{\Sigma}_{\mathbf{F}_t}\mathbf{a}_t + \{V_t + \mu_{\mathbf{F}_t}'\mathbf{R}_t\mu_{\mathbf{F}_t} + tr(\mathbf{R}_t\mathbf{\Sigma}_{\mathbf{F}_t})\}E(\mathbf{\Sigma}_t \mid I_{t-1})\end{aligned}$$

where $\mathbf{f}_t' = \mathbf{F}_t'\mathbf{a}_t$ *and* $q_t = \mathbf{F}_t'\mathbf{R}_t\mathbf{F}_t + V_t$*, the first and second moments of the predictive regression vector,* $\mu_{\mathbf{F}_t} = E(\mathbf{F}_t \mid I_{t-1})$ *and* $\mathbf{\Sigma}_{\mathbf{F}_t} = \mathrm{cov}(\mathbf{F}_t \mid I_{t-1})$.

(iv) *Posterior at* t: $(\mathbf{\Theta}_t, \mathbf{\Sigma}_t \mid I_t) \sim NHIW_G(\mathbf{m}_t, \mathbf{C}_t, b_t, \mathbf{S}_t)$ *with* $\mathbf{m}_t = \mathbf{a}_t + \mathbf{A}_t\mathbf{e}_t'$, $\mathbf{C}_t = \mathbf{R}_t - \mathbf{A}_t\mathbf{A}'q_t$, $b_t = \delta b_{t-1} + 1$, $\mathbf{S}_t = \delta\mathbf{S}_{t-1} + \mathbf{e}_t\mathbf{e}_t'/q_t$ *where* $\mathbf{A}_t = \mathbf{R}_t\mathbf{F}_t/q_t$ *and* $\mathbf{e}_t = \mathbf{Y}_t - \mathbf{f}_t$.

Proof. (i)(ii)(iv) follow directly from Theorem 1. (iii) results from the properties of conditional expectations applied to $p(\mathbf{Y}_t \mid I_{t-1})$, $E(\mathbf{Y}_t \mid I_{t-1})$ and $\mathrm{cov}(\mathbf{Y}_t \mid I_{t-1})$. □

The above theorem suggests a two stage model on the vector time series $\{\mathbf{Y}_t\}$: first, a model is fitted on low dimensional regression vectors $\{\mathbf{F}_t\}$; second, the fitted

model provides the necessary quantities for the dynamic graphical DLMs. Some specific contexts of $\{\mathbf{F}_t\}$ include:

- Pre-fixed regression vector in which the $\mathbf{F}_t$ values are specified in advance by design. This is the assumption made by the standard dynamic linear model, which yields a degenerated prior distribution $p(\mathbf{F}_t \mid I_{t-1})$ with $\mu_{\mathbf{F}_t} = \mathbf{F}_t$ and $\mathbf{\Sigma}_{\mathbf{F}_t} = 0$. In such cases, Theorem 1 applies as a special case of Theorem 2.

- Independent and identically distributed regression vector in which the $n-$vectors $\mathbf{F}_t$ are commonly assumed to be independent and identically distributed from a multivariate normal distribution with mean vector $\mu_{\mathbf{F}}$ and covariance matrix $\mathbf{\Sigma}_{\mathbf{F}}$.

- Dynamic regression vector in which another dynamic model structure could be imposed on vector process $\{\mathbf{F}_t\}$. For example, in asset pricing models, if $\mathbf{F}_t$ is the market excessive return, an AR-GARCH type of model could be applied.

It is important to note that G is pre-specified in Theorem 2. If G is chosen to be empty based on substantive prior information, then the model is essentially the dynamic version of Financial Index Models. If G is allowed to be uncertain, we may use the two practical graph predicting models of equations (11) and (12) to predict graphs in random regression vector DLMs. We may also use equation (15) to choose among different α representing different pairs of graph predicting models and discount factors. Finally, for a given α, the predicted covariance matrix of return $\mathbf{Y}_t$ at time $t-1$ is given by:

$$\text{cov}(\mathbf{Y}_t \mid I_{t-1}) \quad = \quad \sum_{G_{t,j} \in \mathcal{G}} \mathbf{Q}_{t,G_{t,j}} \pi(G_{t,j} \mid I_{t-1}, \alpha) \tag{16}$$

where $\mathbf{Q}_{t,G_{t,j}}$ is defined in (iii) of Theorem 2.

8 Example: portfolio allocation in stocks

To demonstrate the use of DGMs in the Index Model context we work with 100 stocks randomly selected from the population of domestic commonly traded stocks in the New York Stock Exchange. By selecting a random sample of 100 we hope to reduce potential selection biases. The sample period is from January 1989 to December 2008 in a total of 240 monthly returns. The first 60 months are used as a training set to set up the prior at time zero and therefore, the analysis starts at observation 61. Monthly US Treasury bill returns are used as the risk-free rate in the computation of the excess returns. Excess

Model	Mean	Std	Min	25th	Median	75th	Max
Sample	0.158	0.166	−0.530	0.046	0.159	0.269	0.836
CAPM	0.040	0.170	−0.594	−0.075	0.036	0.150	0.825
FF	0.014	0.154	−0.557	−0.092	0.011	0.114	0.816

Table 1: Summary statistics of correlations among sampled stocks. First row, summary of sample correlations; second and third row report summaries of residual correlations after fitting CAPM and FF models respectively. For each case, at the end of April of each year from 1994 to 2008, pairwise correlations are calculated based on the monthly excess returns over the prior 60 months. Summary statistics are based on the estimated values pooled over all years.

returns from a market weighted basket of all stocks in the AMEX, NYSE and NASDAQ were used as the *market* returns. This index along with the Fama-French three factor return data were obtained from the data library of Professor Kenneth R. French [1]. Summary statistics for the excess returns series are given in the first row in Table 1. The median pairwise correlation is 0.159, indicating that there were potentially large payoffs to portfolio diversification.

In an initial exploration of the data we fitted OLS regressions to the returns using the capital asset pricing model (CAPM) and Fama-French (FF) models. The second and third rows in Table 1 show summary statistics of cross-sectional residual correlations. The generally lower correlations compared with the sample correlation suggest that the indexes capture most of the common variation among the securities under consideration. However, there are remaining signals in the residuals as indicated by the maximum and minimum correlations, and these are precisely the quantities we are aiming to explore by relaxing the independence assumption with the inclusion of graphs.

To appreciate the importance and contribution of the use of graphical models, we consider the following alternatives: (1) sample covariance model; (2) Standard dynamic CAPM: $\mathbf{F}_t$ is the market returns and $\text{cov}(\mathbf{Y}_t \mid I_{t-1})$ is the $\mathbf{Q}_{t,G}$ of Theorem 2 when G is empty; (3)Dynamic CAPM with graphs: $\mathbf{F}_t$ is the market returns and $\text{cov}(\mathbf{Y}_t \mid I_{t-1})$ is given by equation (16); (4) Standard dynamic FF: $\mathbf{F}_t$ is the FF three factors and $\text{cov}(\mathbf{Y}_t \mid I_{t-1})$ is the $\mathbf{Q}_{t,G}$ of Theorem 2 when G is empty; (5) Dynamic FF with graphs: $\mathbf{F}_t$ is the FF three factors and $\text{cov}(\mathbf{Y}_t \mid I_{t-1})$ is given by equation (16); and (6) mixtures of (3) and (5) where mixture weights are based on equation (15).

[1] see, http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

In model (1), at each month t, the one-step ahead covariance matrix is based on the data from the preceding 60 months as the *in-sample* period. For models (2)-(6), we use weak priors, $m_0 = 0, C_0 = 10000I, b_0 = 3, S_0 = 0.0003I_{100}$, and $\delta = 0.983$ corresponding to a rolling window of about 60 months. For the random regression vector $\mathbf{F}_t$, we use the sample mean and covariance matrix of the past 60 months as forecasts of the first and second predictive moments, $\mu_{\mathbf{F}_t}$ and $\mathbf{\Sigma}_{\mathbf{F}_t}$. Furthermore, based on simulation experiments in Section 6, we chose to model graph uncertainty with the predictive model of equation (11) for alternatives (3) and (5). Our current code is in a serial version. On a dual-cpu 2.4GHz desktop running CentOS 5.0 unix, the cpu benchmarks for this example run to around 2 to 15 minutes for each month t; the time depends on the change of graphical spaces from $\pi(G_{t,j} \mid I_{t-1})$ to $\pi(G_{t+1,j} \mid I_t)$. Parallel implementations can be expected to run an order of magnitude faster (Jones et al. 2005). We conduct two additional runs of the sequential stochastic search algorithm starting from randomly chosen initials. The computing time and performances are persistent. We report the results from one run. In (6), CAPM and FF models are compared with each other and then averaged based upon their marginal likelihood of equation (15). The resulting posterior probabilities of FF model reach 1 after a short period of time. This should not be surprising as most of the current literature points to the use of a multi-factor model as opposeed to the traditional single factor CAPM. Due to this fact, the overall performances of models (5) and (6) are close so we only report results from model (5) hereafter.

Figure 4 displays the estimated expected number of edges over time starting from January 1994 under models (3) and (5). Three results are worth noting here. First, all graphs are sparse relative to the total 4950 possible edges. The inclusion of graphs provides the necessary flexibility to capture the remaining signals from the residual covariance matrix and the data is responsible to inform which of these non-zero entries are relevant. Second, when comparing with each other, the CAPM model has more edges than FF – once again no surprises here: FF imposes a richer structure for $\mathbf{\Sigma}$ so we should expect more non-zero elements in the residual covariation of assets when the market returns are the only covariate. Third, as more information becomes available, more signals in the residuals are detected.

We now evaluate these forecasting models in two ways: forecasting ability of future correlation matrices and in the construction of optimal portfolios. This is a predictive test in the sense that our investment strategy does not require any hindsight.

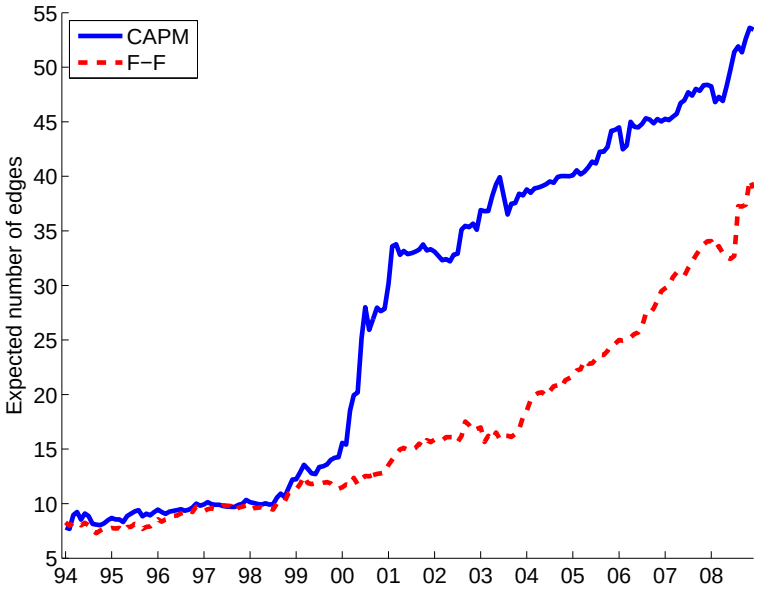

Figure 4: Estimated expectation of the numbers of edges across each month.

8.1 Out-of-sample covariance forecasts

At the end of every month, the correlations forecasts from each model are compared to the sample correlations realized over a subsequent 12 month period in the first experiment, and 36 months in the second experiment. Forecast performance is evaluated in terms of the absolute difference between the realized and forecasted values. Table 2 provides summary statistics on the absolute differences from these two experiments. When the performances evaluated using subsequent 12 months data are compared with those from subsequent 36 months, the average absolute forecast errors are reduced in the 36 month case. The drop in forecast errors suggests that there is a lot of noise in covariance matrices measured over a period as short as 12 months. Nevertheless, in both experiments, the relative performance of each model is generally the same.

The full sample covariance model, which is the most complex model in terms of the number of free parameters, has the highest median absolute error and root mean square error. All other models are better than the full covariance model. More complex models do not necessarily offer smaller forecast errors. This message is consistent with many empirical studies on correlation matrix forecasts of stock returns.

Comparing model (2) with (3) in the CAPM family, and model (4) with (5) in the

	12 month				36 month			
Model	Median	Std	95th	$\sqrt{\text{MSE}}$	Median	Std	95th	$\sqrt{\text{MSE}}$
(1) Full covariance	0.238	0.200	0.654	0.340	0.160	0.141	0.460	0.235
(2) CAPM Empty	0.234	0.186	0.612	0.323	0.146	0.127	0.413	0.212
(3) CAPM Graph	0.230	0.184	0.605	0.319	0.143	0.123	0.402	0.207
(4) FF Empty	0.230	0.185	0.609	0.321	0.143	0.124	0.405	0.208
(5) FF Graph	0.230	0.185	0.607	0.320	0.143	0.123	0.404	0.208

Table 2: Performance of correlation forecasting models. Forecasts of monthly return correlation matrices are generated from different models, based on the prior 60 months of data for model (1) and based on discount factor $\delta = 0.983$ for models (2)-(5). Forecasts are then compared against the realized sample covariance estimated over the subsequent 12 months (first four columns) and 36 months (last four columns). The last estimation period ends in December 2005. Summary statistics are provided for both the distribution of the absolute difference between realized and forecasted value of pairwise correlations: Std, standard deviation of absolute differences; 95th, 95th quantile of absolute difference, and $\sqrt{\text{MSE}}$, the root mean square errors of forecasts.

FF family, we see that models with graphs are better than their empty graph peers. This is more evident in the CAPM family. Model (3) has reduced the median of the absolute differences and the root mean square errors relative to model (2), while model (5) has almost the same absolute differences as model (4). The clearer advantage in the CAPM family is because there are more structures in the residuals left unexplained by only the market index than by the FF three indexes. In general, the improvement of out-of-sample covariance forecasts is minor. This is actually as expected, since the signals are very sparse which indicates the covariance matrices based on the graphical models do not differ much from those based on the traditional index models. This is actually as expected, since the empty and the graphical models generate different second moments, and the only difference between these second moments is that graphical models have additional sparse signals in their residual covariance matrices. However, as the experiments in the following section will show, these signals, though sparse, are influential when the forecasted covariance matrices are used to build optimal portfolios.

8.2 Portfolio optimization

From a practical point of view, the optimization experiments provide perhaps more important metrics for evaluating forecasting models. The setup of our portfolio optimization experiments is as follows. To highlight the role of the second predictive moment, we first form the global minimum variance portfolio. At the end of April of each year starting from 1994, we use the different models to predict the one-step ahead covariance matrix for the 100 stocks. These predictions are the input to a quadratic programming routine that defines the minimum variance portfolio (Markowitz 1959). Short sales are allowed so that the weights are only required to be summed up to 1. These weights are then applied to buy-and-hold portfolio returns until the next April, when the forecasting and optimization procedures are repeated. The resulting time series of monthly returns of portfolios allow us to characterize the performance of the optimized portfolio based on each model. We also form a mean-variance portfolio using the first two moment forecast $\{\mathbf{r}_t, \mathbf{Q}_t\}$ from Theorem 2 with a target annualized excessive mean return of 15%.

Table 3 summarizes these optimization results. These are all expressed on an annualized basis. In comparison within each group, it is clear that the introduction of the graphical structure helps. The annualized standard deviation of the optimized portfolio based on the graphical CAPM model is 10.7%, yielding a Sharpe ratio of 0.688, compared to a Sharpe ratio of 0.533 for the standard CAPM portfolio. The same advantage of using graphs can be found in the two models within FF class. The conclusion from this example is simple: it pays to allow for a more flexible residual covariance structure in the implementation of Index Models.

9 Further comments

By allowing more flexible models for the residual covariance matrix, Financial Index models can be improved in their abilities to build more effective optimal portfolios. In this paper we take advantage of the DGMs framework of Carvalho and West (2007a) and show that graphical models can also be used to identify sparse signals in the residual covariance matrices and thereby obtain a more complex representation of the distribution of asset returns. Unlike Carvalho and West (2007a) and Quintana et al. (2009), in the Index Model framework, graphs are used to increase the complexity of an otherwise very restrictive model. In that sense, it is our hope that our work complements the widely used tool box of dynamic linear models for the analysis of asset returns. Our

	Minimum variance portfolio			Mean variance portfolio		
Model	Rate	Std	Sharpe	Rate	Std	Sharpe
(1) Full covariance	-	-	-	-	-	-
(2) CAPM Empty	0.064	0.120	0.533	0.064	0.119	0.535
(3) CAPM Graph	0.074	0.107	0.688	0.075	0.107	0.700
(4) FF Empty	0.062	0.109	0.569	0.069	0.109	0.627
(5) FF Graph	0.070	0.105	0.661	0.072	0.106	0.678

Table 3: Performance of portfolios based on forecasting models. Summary statistics are presented: Rate, the annualized excessive returns $r - r_T$, where the annualized portfolio return r is determined by $(1+r)^{14} = \prod_{i=1}^{168}(1+r_i)$, and annualized risk-free return r_T is determined by $(1+r_T)^{14} = \prod_{i=1}^{168}(1+r_{T,i})$ with r_i and $r_{T,i}$ denoting the monthly return of portfolio and risk-free asset; Std, the annualized standard deviation of excess returns $r_i - r_{T,i}$; and Sharpe ratio, the annualized excessive return divided by the annualized standard deviation.

first example helps illustrate the model implementation and highlight the issue of specifying discount factors and graph predicting models. The second example discusses and explores aspects of random regression vectors and variable selection. This analysis confirmed that the CAPM and FF models generally do well in explaining the variation of stock returns, but identifying relevant non-zero entries in the unexplained covariation is of real practical value: the resulting covariance matrix forecast has lower out-of-sample forecast errors, and the corresponding portfolios achieve a lower level of realized risk in terms of variance and higher realized returns.

In addition to our case studies, we have also provided a fully Bayesian framework of two-stage forecasts of covariance matrices, a mechanism of graph evolution, and the use of sequential stochastic search for high-dimensional graphical model spaces.

In regards to the modeling of graphical structure through time, alternative approaches include the use of first-order Markov probabilities in which the graph obtained at time t depends on the graphs obtained at time $t-1$, but not on what happened prior to $t-1$, and higher-order Markov probabilities that extend the dependence to graphs at time $t-2, t-3, \cdots$, etc. These alternatives require the learning of a higher-dimensional transition matrix between graphs. Even a sparse representation of the transition matrix, such as each graph only moves to its neighbors between two time points, is limited in the sense that the sparse pattern would restrict the evolution of graphs between time.

The sequential stochastic search algorithm combines the sequential Monte Carlo idea and the shotgun stochastic search algorithm. Exploration of a static model space to find high posterior probability graphs can be successfully carried out using direct search such as the shotgun stochastic search method, certainly up to 100 vertices or so while traditional MCMC is competitive only for relatively small graphs (Jones et al. 2005). However, fast searching a sequence of large model space is more challenging. This problem can be eased by noticing that from one step to the next we do not expect large changes in the mass of the distribution. Therefore, we could use the high probability graphs from the previous step as starting points to initiate a new search and rapidly traverse the graphical model space around these promising models.

References

Carvalho, C. M. and West, M. (2007a). "Dynamic Matrix-Variate Graphical Models." *Bayesian Analysis*, 2: 69–98.
URL http://ba.stat.cmu.edu/vol02is01.php

— (2007b). "Dynamic Matrix-Variate Graphical Models - A synopsis." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics VIII*, 585–590. Oxford University Press.

Cochrane, J. (2001). *Asset Pricing*. Princeton University Press.

Dawid, A. P. and Lauritzen, S. L. (1993). "Hyper-Markov laws in the statistical analysis of decomposable graphical models." *Annals of Statistics*, 21: 1272–1317.

Fama, E. F. and French, K. R. (1993). "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, 33(1): 3–56.
URL http://ideas.repec.org/a/eee/jfinec/v33y1993i1p3-56.html

Harrison, P. J. and Stevens, C. F. (1976). "Bayesian Forecasting." *Journal of the Royal Statistical Society. Series B*, 38: 205–247.

Jagannathan, R. and Wang, Z. (1996). "The Conditional CAPM and the Cross-Section of Expected Returns." *Journal of Finance*, 51(1): 3–53.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in stochastic computation for high-dimensional graphical models." *Statistical Science*, 20: 388–400.

Kalman, R. (1960). "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering*, 82: 35–45.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

Lintner, J. (1965). "The Valuation of Risky Assets and the Selection of Risky Investment in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics*, 47: 13–37.

Markowitz, H. (1959). *Portfolio Selection:Efficient Diversification of investments*. New York: Wiley.

Quintana, J., Lourdes, V., Aguilar, O., and Liu, J. (2003). "Global gambling." In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M. (eds.), *Bayesian Statistics VII*, 349–368. Oxford University Press.

Quintana, J. M., Carvalho, C. M., Scott, J., and Costigliola, T. (2009). "Futures Markets, Bayesian Forecasting and Risk Modeling." In O'Hagan, T. and West, M. (eds.), *The Handbook of Applied Bayesian Analysis*. Oxford University Press.

Quintana, J. M. and West, M. (1987). "Multivariate time series analysis: New techniques applied to international exchange rate data." *The Statistician*, 36: 275–281.

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). "Flexible Covariance Estimation in Graphical Gaussian Models." *Annals of Statistics*, 36: 2818–49.

Sharpe, W. F. (1964). "Capital asset prices: a theory of market equilibrium under conditions of risk." *Journal of Finance*, 19: 425–442.

Smith, J. (1979). "A Generalization of the Bayesian Steady Forecasting Model." *Journal of the Royal Statistical Society. Series B*, 41: 375–387.

West, M. and Harrison, P. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

Zellner, A. and Chetty, V. K. (1965). "Prediction and Decision Problems in Regression Models from the Bayesian Point of View." *Journal of the American Statistical Association*, 60(310): 608–616.
URL http://www.jstor.org/stable/2282695