BAYESIAN STATISTICS 9
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Dynamic Stock Selection Strategies: A Structured Factor Model Framework

Carlos M. Carvalho
The University of Chicago and *The University of Texas at Austin, USA*
carlos.carvalho@chicagobooth.edu
carlos.carvalho@mccombs.utexas.edu

Hedibert F. Lopes
The University of Chicago, USA
hlopes@chicagobooth.edu

Omar Aguilar
Financial Engines, USA
o_aguilar@ymail.com

Summary

We propose a novel framework for estimating the time-varying covariation among stocks. Our work is inspired by asset pricing theory and associated developments in Financial Index Models. We work with a family of highly structured dynamic factor models that seek the extraction of the latent structure responsible for the cross-sectional covariation in a large set of financial securities. Our models incorporate stock specific information in the estimation of commonalities and deliver economically interpretable factors that are used both, as a vehicle to estimate large time-varying covariance matrix, and as a potential tool for stock selection in portfolio allocation problems. In an empirically oriented, high-dimensional case study, we showcase the use of our methodology and highlight the flexibility and power of the dynamic factor model framework in financial econometrics.

Keywords and Phrases: Dynamic factor models; Financial index models; Portfolio selection; Sparse factor models; Structured loadings.

Carlos M. Carvalho is Assistant Professor of Econometrics and Statistics, University of Chicago Booth School of Business. Hedibert F. Lopes is Associate Professor of Econometrics and Statistics, University of Chicago Booth School of Business. Omar Aguilar is Head of Portfolio Management at Financial Engines. The authors would like to thank Robert McCulloch for the helpful discussions throughout this project. Carvalho would like to acknowledge the support of the Donald D. Harrington Fellowship Program and the IROM department at The University of Texas at Austin.

1. INTRODUCTION

The understanding of co-movements among stock returns is a central element in asset pricing research. Knowledge of this covariation is required both to academics seeking to explain the economic nature and sources of risk and to practitioners involved in the development of trading strategies and asset portfolios. This leads to a vast literature dedicated to the estimation of the covariance matrix of stock returns; a challenging problem due to complex dynamic patterns and to the rapid growth of parameters as more assets are considered.

Since the proposal of the *Capital Asset Pricing Model* (CAPM) by Sharpe (1964) and the *Arbitrage Pricing Theory* (APT) of Ross (1976), Financial Index Models became a popular tool for asset pricing. These models assume that all systematic variation in the return of financial securities can be explained linearly by a set of market indices, or risk factors, leading to a highly structured covariance matrix. In financial terms, the implication is that equity risk is multidimensional but priced efficiently through a set of indices so that the only source of additional expected return is a higher exposure to one of these risk factors.

The appeal of index models is two-fold: (*i*) it leads to tractable and parsimonious estimates of the covariances and (*ii*) it is economically interpretable and theoretically justified. It follows that the task of estimating a large covariance matrix got simplified to the task of identifying a set of relevant risk factor. This is an empirical question usually guided by economic arguments leading to factors that represent macro-economic conditions, industry participation, etc. A very large body of literature is dedicated to selecting and testing the indices - we refer the reader to Cochrane (2001) and Tsay (2005).

In a series of papers, Fama and French (FF) identified a significant effect of market capitalization and book-to-price ratio into expected returns. This has lead to the now famous *Fama-French* 3 factor model where, besides the market, two indices are built as portfolios selected on the basis of firms' size and book-to-price ratio. This is perhaps the most used asset pricing model in modern finance research and it relates to many trading strategies based on "growth" and "value" stocks. An additional index based on past performance information (momentum) was proposed by Cahart (1997) and can also be considered a "default" factor these days.

The fact that size, book-to-price and momentum are relevant to explain covariation among stocks is exploited in two common ways:

- as individual regressors in a multivariate linear model;
- as ranking variables used to construct portfolios that are used as indices.

The first approach follows the ideas of Rosenberg and McKibben (1973) and it is known as the BARRA strategy (after the company BARRA, Inc. founded by Barr Rosenberg). The second is initially proposed by Fama and French (1993).

Taking the view that Financial Index Models are an appropriate choice for the purpose of covariance estimation and asset allocation, we develop a dynamic factor model framework that contextualizes the current ideas behind these 4 aforementioned factors. Our approach will encompass both the BARRA and Fama-French strategies in a simple yet flexible modeling set up. Part of the innovation is to propose a framework where variable specific information can be used in modeling the latent structure responsible for common variation. From a methodological viewpoint, our models can be seen as a "structured" extension of current factor model ideas as developed in Aguilar and West (2000), West (2003), Lopes and West (2004),

Lopes, Salazar and Gamerman, (2008) and Carvalho, *et al.*, (2008). On the applied side our goal is to propose a model-based strategy that creates better Financial Index Models, help deliver better estimates of time-varying covariances and lead to more effective portfolios.

We start in section 2 by introducing the general modeling framework. In Section 3 we define the specific choices defining the different index models. Section 4 explores a case study where the different specifications are put to the test in financial terms. Finally, in Section 5 we discuss the connections of our approach with the current factor model literature and explore future uses of the ideas presented here.

2. GENERAL FRAMEWORK

The general form of an Index Model assumes that stock returns are generated following:

$$r_t = \alpha_t + \mathbf{B}_t f_t + \epsilon_t \tag{1}$$

where f_t is a vector of common factors at time t, $\mathbf{B}_t$ is a matrix of factor loadings (or exposures) and ϵ_t is a vector of idiosyncratic residuals. If $Var(f_t) = \mathbf{\Theta}_t$ and $Var(\epsilon_t) = \mathbf{\Phi}_t$ the model in (1) implies that

$$Var(r_t) = \mathbf{B}_t \mathbf{\Theta}_t \mathbf{B}_t' + \mathbf{\Phi}_t.$$

When the number of factors is much smaller than the number of stocks, the above form for the covariance matrix of returns is represented by a relatively small set of parameters as the only source of systematic variation are the chosen indices. Assuming further that the factors are observable quantities the problem is essentially over as one is only left with a simple dynamic regression model and in fact, most of the literature will follow a "rolling window" approach based on OLS estimates (see Tsay, 2005, chapter 9).

In our work, we take a dynamic, model-based perspective and assume that at time t we observe the vector (r_t, x_t, Z_t) where:

- r_t is a p-dimensional vector of stock returns;
- Z_t is a $p \times k$ matrix of firm specific information; and
- x_t is the market return (or some equivalent measure).

We represent Index Models as defined by the dynamic factor model framework:

$$r_t = \alpha_t + \beta_t x_t + \mathcal{Z}_t f_t + \epsilon_t \tag{2}$$

where β_t is a p-dimensional vector of market loadings, ϵ_t is the vector of idiosyncratic residuals, and f_t is a k-dimensional vector of common factors. Our notation clearly separates the one factor that is observed (the market) from the rest of the factors that are latent (f_t). In all model specifications, we assume that each element of both α_t and β_t follow a first-order dynamic linear model (West and Harrison, 1997) and that ϵ_t is defined by a set of independent stochastic volatility models (Jacquier, Polson and Rossi, 1994; Kim, Shephard and Chib, 1998). Finally, we assume that

$$f_t \sim N(0, \mathbf{\Theta}_t)$$

where $\mathbf{\Theta}_t$ is diagonal with dynamics driven by univariate stochastic volatility models.

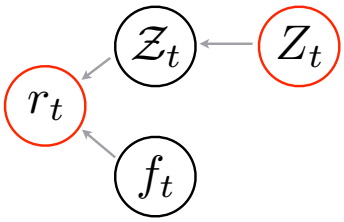

Figure 1: *Illustration summarizing the idea of structuring the loadings with observed variables present in our proposed framework. The red circles represent observable variables.*

Defining the factor loadings. One last element remains to be defined and it is in the core of the different model specifications considered: the $(p \times k)$ matrix of factor loadings $\mathcal{Z}_t$. Through $\mathcal{Z}_t$, company specific information will be used to help uncover relevant latent structures representing the risk factors. Before getting to the specific definitions of $\mathcal{Z}_t$ it is worth noting that many previously proposed models are nested in the form of (2). For example, taking $\beta_t = 0$ and fixing the loading through time gets us to the factor stochastic volatility models of Aguilar and West (2000) and Pitt and Shephard (1999). Letting the loadings vary in time with a DLM leads to the model considered in Lopes, Aguilar and West (2000) and Lopes and Carvalho (2007).

3. MODEL SPECIFICATIONS

3.1. *Dynamic CAPM*

We start with the simplest alternative in the proposed framework. Let $\mathcal{Z}_t = 0$ for all t and the dynamic CAPM follows:

$$\begin{aligned} r_t &= \alpha_t + \beta_t x_t + \epsilon_t \\ \alpha_{i,t} &\sim N(\alpha_{i,t-1}, \tau^2_{\alpha_i}) \\ \beta_{i,t} &\sim N(\beta_{i,t-1}, \tau^2_{\beta_i}) \\ \epsilon_{i,t} &\sim \text{SV Model} \end{aligned}$$

with independent dynamics for α_t, β_t and ϵ_t across i, for $i = 1, \ldots, p$. This is also the model with a very simple implementation strategy where conditional on the market, all the estimation is done in parallel for all p components in the vector of returns. Due to its historical relevance, this dynamic version of the CAPM will serve as the benchmark for comparing the alternative specifications.

3.2. *Dynamic BARRA*

If we now set $\mathcal{Z}_t = Z_t$ we get a dynamic version of the BARRA approach where the loadings are deterministically specified by the company-specific variables Z_t. Following the ideas of Fama and French (1996) and Carhart (1997), Z_t would have 3 columns with measures of market capitalization (size), book-to-price ratio and

momentum. The model follows:

$$\begin{aligned} r_t &= \alpha_t + \beta_t x_t + Z_t f_t + \epsilon_t \\ \alpha_{i,t} &\sim N(\alpha_{i,t-1}, \tau^2_{\alpha_i}) \\ \beta_{i,t} &\sim N(\beta_{i,t-1}, \tau^2_{\beta_i}) \\ f_t &\sim N(0, \Theta_t) \\ \epsilon_{i,t} &\sim \text{SV Model} \end{aligned}$$

This model is jointly estimated as the common factors f_t are now latent. This is still a somewhat standard model as it is a version of the models in Aguilar and West (2000) and Lopes and Carvalho (2007) where some factors are given (x_t) and their loadings have to be estimated and some time-varying loadings are given (Z_t) and their factor scores are unknown. It is important to highlight that by fixing the loadings at Z_t we force the latent factors to embed the information in the firm specific characteristics leading to set of latent factors with a direct economic interpretation as "size", "book-to-market" and "momentum" factors.

3.3. *Sparse Dynamic BARRA*

Having the different firm-specific characteristics directly defining the factors might be problematic due to potentially large amount of noise contained in these variables. The use of portfolios suggested by Fama and French (1993) was originally an attempt to filter out the relevant information contained in firm specific information about the underlying risk factors defining the covariation of equity returns. In our proposed framework this problem could be mitigated by additional structure in $\mathcal{Z}_t$. For example, we can take the view that due to excessive noise, some elements of Z_t should not play a role at a given time so that the corresponding element in $\mathcal{Z}_t$ would be set to zero. The introduction of sparsity in the loadings matrix of a factor model, as an attempt to regularize the estimation of factors in large dimensional problems, first appears in West (2003) and got further explored in Carvalho *et al.* (2008) and Frühwirth-Schnatter and Lopes (2010). We extend their approach to the time-varying loadings set-up of the dynamic BARRA by modeling the loadings of factor j at time t as:

$$\mathcal{Z}_{ij,t} = \begin{cases} Z_{ij,t} & \text{w.p.} \quad \pi_{j,t} \\ 0 & \text{w.p.} \quad 1 - \pi_{j,t} \end{cases}$$

where $\pi_{j,t}$ are the inclusion probabilities associated with factor j and are usually modeled with a beta prior. Again, this is a fairly straightforward model to estimate. Given $\mathcal{Z}_t$ we are back to a dynamic stochastic volatility factor model whereas, conditional on all remaining unknowns, each elements of $\mathcal{Z}_{j,t}$ requires a draw from a simple discrete mixture. Altought simple, the reader should be reminded that fitting such models to high-dimensional problems is computationally intensive and require careful coding as standard statistical packages are not up to the tasks. As an example, in the $p = 350$ dimensional case study presented below, each MCMC iteration requires, among other things, 703 filter-forward backward-sampling steps and sampling 1,050 elements of $\mathcal{Z}_t$. As a side note, given the conditionally Gaussian structure of the models, efficient sequential Monte Carlo algorithms are available and very attractive for the on-line sequential application of the proposed framework (see Aguilar and West, 2000 and Carvalho, Johannes, Lopes and Polson, 2010).

3.4. *Dynamic Fama-French*

Fama and French (1996) and Carhart (1997) define factors as portfolios built by sorting stocks based on their individual characteristics. The implied 4 factor model (3 factors plus the market) is by far the most successful empirical asset pricing model in modern finance. More specifically, the SMB (small minus big) factor is defined by ranking the stocks according to their market capitalization and building a value weighted portfolio with the returns of the firms below the median market cap, minus the returns of the firms above the median. The idea behind this construction is motivated by the observation that small firms seem to earn larger average returns relative to the prediction of the CAPM (also know as "growth" effect).

The HML (high minus low) factor is defined by ranking the stocks according to their book-to-price ratio and building a value weighted portfolio with the returns of the highest 30% book-to-price firms minus the returns on the lowest 30%. The intuition here is that "value" stocks have market value that are small relative to their accounting value and therefore tend to present higher than expected (by the CAPM) returns.

Finally, Carhart's momentum factor (MOM) starts by ranking stocks according to some measure of past performance and building equal weighted portfolios with the returns of the 30% top performers minus the returns on the 30% bottom past performers. Again, the idea arises from the observation that stock prices are mean reverting and therefore past losers with present higher than expected returns (see Jegadeesh and Titman, 1993).

We borrow these ideas and adapt their construction to our dynamic factor framework. To this end we use the dynamic BARRA set up of Section 3.2 and define $\mathcal{Z}_t$ following the directions above. This means that, at each time point, the loadings matrix takes values defined by the sorting variables size, book-to-price and momentum. In detail, the first column of $\mathcal{Z}_t$ takes values "+ market value" for small companies and "- market value" for large companies (as defined by the median at time t). The second column takes values "+ market value" for companies in the top 30% of book-to-price, "- market value" for companies in the bottom 30% and 0 otherwise. The final column is defined with +1 for the top 30% past performers, -1 for the bottom 30% and 0 otherwise.

Extending the specification of Section 3.3 is immediate and would serve the similar purpose of regularization. In addition it is a model-based alternative to sorts and ad-hoc cut-offs for inclusion in each factor. In that spirit, we could define the *Sparse Dynamic Fama-French* model in the same manner as in Section 3.3 but with the potential values of $\mathcal{Z}_t$ defined according to the instructions of Fama, French and Carhart.

3.5. *Probit-Sparse Dynamic Factor Models*

In this final specification we modify the sparse specification (either BARRA or Fama-French) so that to model the inclusion probabilities as a function of individual firm characteristics. By doing so we allow for different relationship forms between firm characteristics and their association with a latent risk factor. Once again, let

$$\mathcal{Z}_{ij,t} = \begin{cases} \theta_{ij,t} & \text{w.p.} \quad \pi_{ij,t} \\ 0 & \text{w.p.} \quad 1 - \pi_{ij,t}, \end{cases}$$

but now,

$$\pi_{ij,t} = probit(\gamma_j + \phi_j W_{ij,t}).$$

In the above, $\theta_{ij,t}$ is whatever chosen value to the loadings when variable i is involved with factor j. In the BARRA set up that could be the stock specific information Z_t or the simple transformations in the Fama-French context. $W_{ij,t}$ is the variable that carries information of whether or not stock i and factor j are related. This definition provides yet additional flexibility in using firm specific information in building systematic risk factors. Instead of using sorts or assuming that inclusion in a factor is exchangeable a priori across firms, this model is more informative and allow for more complex relationships to be uncovered. This is also a very useful context for the use of informative priors in relating variables to factors and for exploring non-linear relationships with polynomials and related transformations inside the probit link. One example, that relates directly to the Fama-French sorting, takes W_j to be a measure of distance from the median size company and assume that it is believed a priori that $\phi_j > 0$. That would imply that the larger (or smaller) a company is the more likely it is to participate in the associated factor.

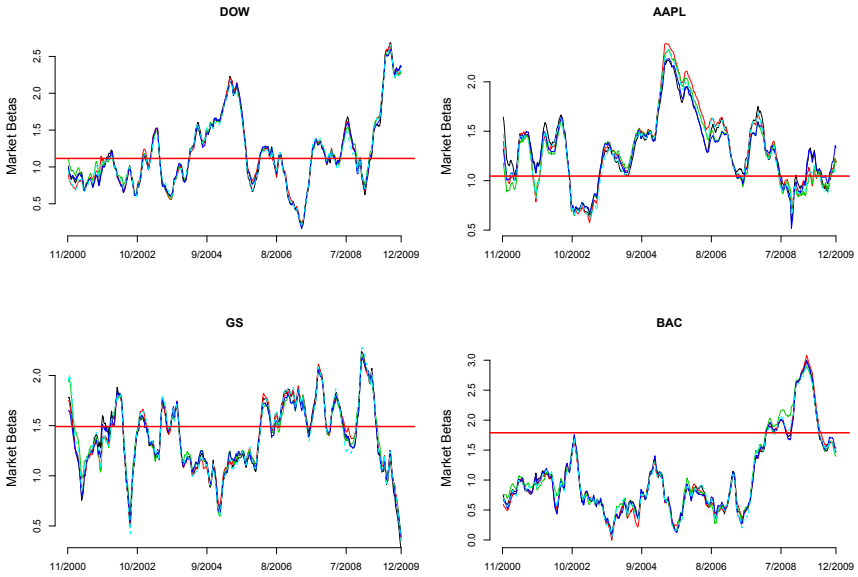

Figure 2: Case Study. *Market β's of Dow Chemical, Apple, Goldman Sacks and Bank of America for all models. The horizontal red line represents the OLS estimate of β in a simple linear regression.*

Although very appropriate to the applied context discussed here it is important to notice that the idea of using additional information in modeling factor loadings is much more general and widely applicable. Our ideas are inspired by the work of Lopes, Salazar and Gamerman (2008) where priors for factor loadings were informed by spatial locations. In section 4 a simulated example showcases the potential relevance of this approach in uncovering important latent structures responsible for common variation.

4. EXAMPLES

4.1. *Case Study: 350 stocks*

Our case study focuses on a set of 350 stocks in the U.S. market (part of the Russel 1000 index). From October 2000 to December 2009 we work with weekly returns and use size, book-to-price and momentum as stock specific information. An overall value-weighted index (from CRSP) is used as market returns. Due to the preliminary nature of this work we selected our variables to avoid missing data problems. This example serves as a test ground for the models and we hope to extend this analysis to the entire population of stocks in the near future.

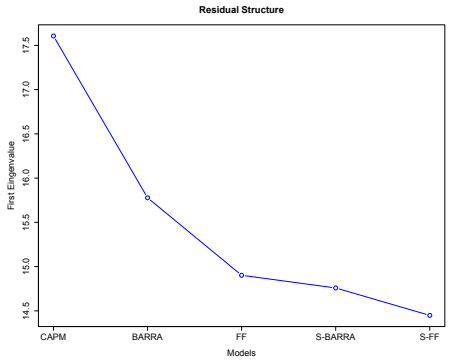

Figure 3: *Case Study. Eigenvalues of the covariance matrix of standardized residuals from each model: Dynamic CAPM, Dynamic BARRA, Dynamic Fama-French, Sparse Dynamic BARRA and Sparse Dynamic Fama-French. Absence of residual covariation would imply a eigenvalue of 1.*

Five models were considered in the initial analysis: (*i*) Dynamic CAPM, (*ii*) Dynamic BARRA, (*iii*) Dynamic Fama-French, (*iv*) Sparse Dynamic BARRA and (*v*) Sparse Dynamic Fama-French. Figure 2 shows the posterior means of the market β_t's for four companies in all models. The first thing to take notice is the clear dynamic nature of β – a fact that is ignored in a variety of empirical and theoretical work where OLS estimates (like the one presented in the figure) are used. It is also interesting to notice that the path of β's is very similar in all models leading to the conclusion that the market information is essentially orthogonal to the information contained in individual firm characteristics (at least in relation to the factors they create). This empirical fact has been observed in several articles in the finance literature and is discussed in detail by Cochrane (2001). In other words, our different factor models are seeking to uncover the latent structure left after the CAPM does its job.

A summary of the remaining unexplained linear "structure" in the residuals appears in Figure 3 where we compare the first eigenvalue of the standardized residual covariance matrix of each model. No residual structure would imply an eigenvalue of 1. It is important to remember that all models other than the Dynamic CAPM are of the same complexity and try to explain covariation with 4 factors. As expected, the simplest model, i.e., the Dynamic CAPM, leaves the most structure

behind while the Sparse Dynamic Fama-French picks up the most common variation among stocks. This is the first indication that our initial conjecture that not all stocks should be playing a role in determining the underlying factor associated with firm characteristics might be a relevant one. By simply zeroing out some elements of $\mathcal{Z}_t$ we ended up extracting factors better able to explain common variation, at least under this simple measure.

Table 1: *Bayes Factors in relation to the benchmark Dynamic CAPM.*

Model	*log(BF)*
Dynamic BARRA	-267.59
Dynamic Fama-French	-102.55
Sparse Dynamic BARRA	**343.50**
Sparse Dynamic Fama-French	**473.44**

A more relevant overall comparison of the performance of the models is presented in Table 1 where an approximate measure of the *log* Bayes Factor in relation to the Dynamic CAPM is presented (See Lopes and West, 2004). The evidence in favor of the Sparse BARRA and Sparse Fama-French specification is overwhelming while the simple Dynamic CAPM seems to be a better alternative than both the Dynamic BARRA and Dynamic Fama-French. Once again, this indicates that firm specific information can be helpful in uncovering relevant underlying structure but a simple *ad-hoc* definition of the loadings is not sufficient. The Sparse Dynamic BARRA and Sparse Dynamic Fama-French are our first attempt in trying to improve the modeling of the loadings and their results are so far promising.

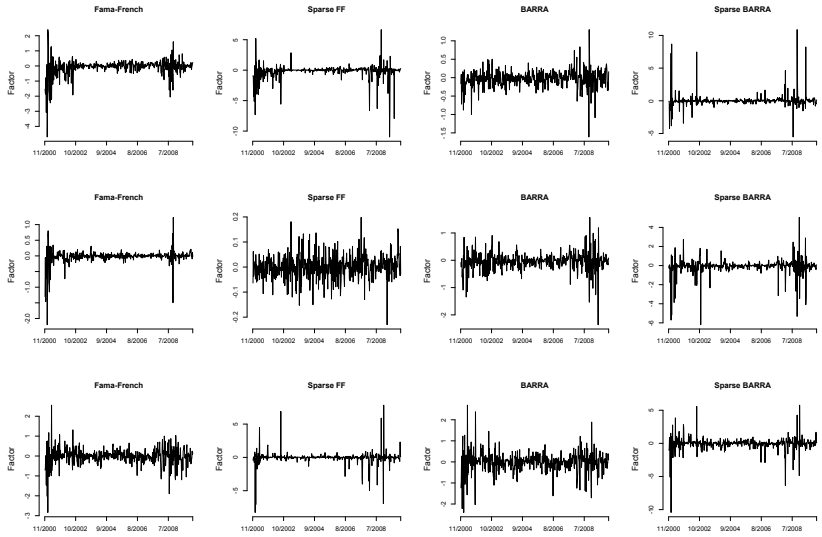

Figure 4: *Case Study. Posterior means of the factor scores. The rows represent the "size", "book-to-price" and "momentum" factors.*

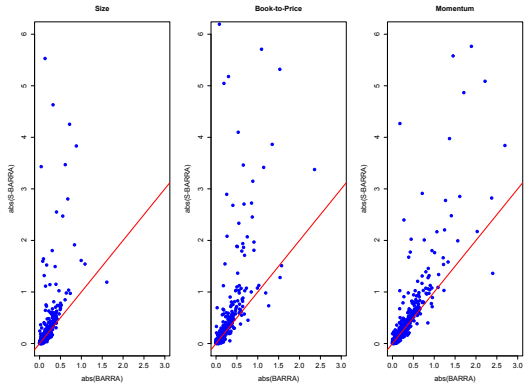

Figure 5: *Case Study. Scatter plots of factor scores from the Dynamic BARRA and Sparse Dynamic BARRA model specifications. In red, the 0-1 line.*

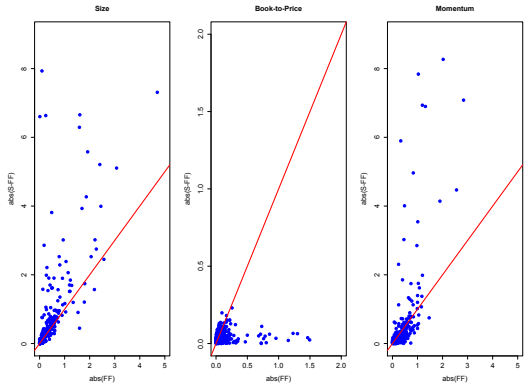

Figure 6: *Case Study. Scatter plots of factor scores from the Dynamic Fama-French and Sparse Dynamic Fama-French model specifications.*

To better understand the results in the different specifications it is worth examining the factor scores a little closer. Figure 4 shows the posterior means for all 3 latent factors in all models. It is clear that the f_t's are very different at a first glance as different values of $\mathcal{Z}_t$ have a tremendous impact in the estimation of f_t. This is indeed the case when comparing the factors from the Dynamic BARRA and Dynamic Fama-French. A second look however, shows that the results from the Dynamic BARRA and Dynamic Fama-French are quite related to their sparse counterparts. Figures 5 and 6 display scatter plots of the absolute value of each of the 3 factor scores in both sparse and non-sparse models. They are clearly linearly related but the results from the Dynamic BARRA are overly shrunk towards zero due to excessive noise in the loadings. The regularization exert by the sparse representation is able to better identify time periods where just a subset of stocks are really associated with the size, book-to-price and momentum effects leading to risk factors that are better able to explain covariation.

Table 2: *Inclusion Probabilities: "Overall" stands for the overall average of the posterior means of $\pi_{j,t}$ for each factor j. "Peak Dates" refer to the average for the time periods when we identify a big disparity between the factor scores obtained in the sparse versus non-sparse model specifications. In the Sparse Fama-French model, we don't observe the shrinkage effect in the Book-to-Market factor hence the N/A values.*

	Overall	*"Peak Dates"*
Size (BARRA)	0.5890	0.2501
Book-to-Market (BARRA)	0.5789	0.3718
Momentum (BARRA)	0.5971	0.3816
Size (FF)	0.5952	0.4025
Book-to-Market (FF)	N/A	N/A
Momentum (FF)	0.5886	0.2697

This point is emphasized by Table 2 where we summarize and compare the overall estimates of the inclusion probabilities $\pi_{j,t}$ relative to their values when factors scores are overly shrunk by the non-sparse models. The clear reduction in the probabilities implies that only a smaller subset of stocks share covariation through the characteristics based factors. Recall that the differences in the Bayes Factor between the Dynamic BARRA and Fama-French and their sparse versions are enormous even though the difference in their latent factor scores is somewhat subtle.

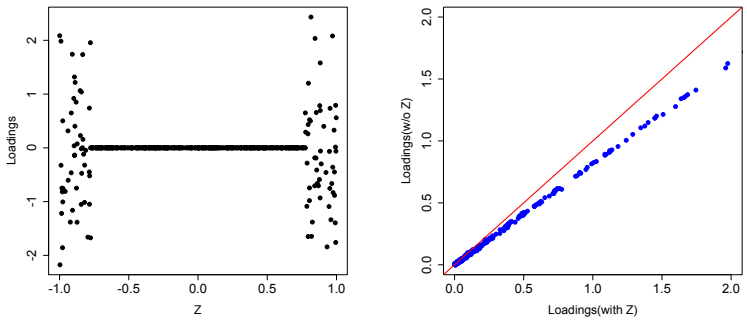

Figure 7: *Illustrative example. The left panel shows the relationship of the loadings in factor 2 with the explanatory variable Z. The right panel plots the estimates of the loadings with or without the information in Z.*

Finally, Figure 11 shows the growth in estimation risk as a function of dimension (p) and the conclusion is simple: the larger the problem, the higher the importance of appropriately using the information in Z.

To explore the financial effects of the different models, we build minimum variance portfolios based on the sequence of estimates of the covariance matrices of returns. This comparison is useful as it isolates the impact of the covariance matrix in investment decisions as the optimization solution only involves its inverse.

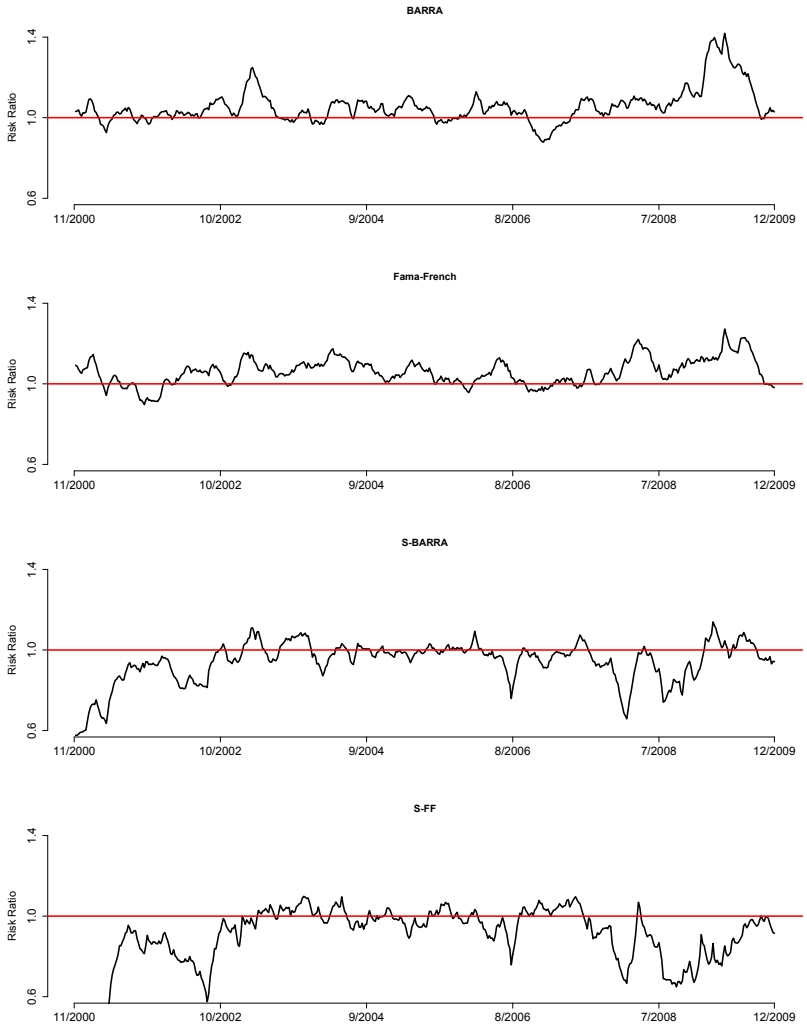

Figure 8: *Case Study. The estimated risk ratio of the returns obtained from minimum variance portfolios from the different models relative to the Dynamic CAPM. The volatility of the returns associated with each strategy was estimated via a stochastic volatility model.*

Figure 8 displays the series of risk ratios of each portfolio *vis-a-vis* the benchmark portfolio constructed by the Dynamic CAPM. Once again the observation is that the Sparse Dynamic BARRA and Sparse Dynamic Fama-French provide a significant improvement over the Dynamic CAPM as, for most time points, it results in a less volatile investment option.

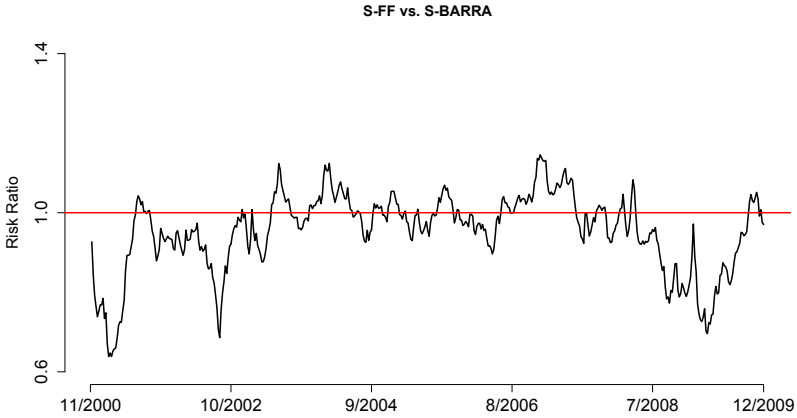

Figure 9: *Case Study. The estimated risk ratio of the returns obtained from minimum variance portfolios in the Sparse Dynamic Fama-French relative to the Sparse Dynamic Barra. The volatility of the returns associated with each strategy was estimated via a stochastic volatility model.*

4.2. *An Illustration*

We close this example with an illustration of the overall improvement of the proposed models relative to what we commonly see in many asset pricing articles. Figure 9 presents boxplots of the percentage of variation explained by the models (essentially a R^2 like measure) for each return series. The red boxplots refer to the standard regression-based CAPM, BARRA and Fama-French while their green counterparts are obtained from our proposed models. It is clear that the time-varying framework provides potentially relevant improvements and, once again, their sparse versions appear on top.

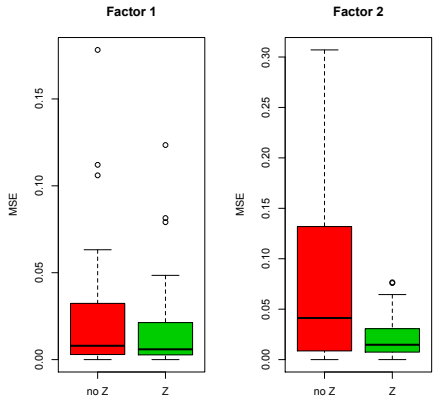

Figure 10: *Illustrative example. Errors in the estimation of factor scores over 100 simulations.*

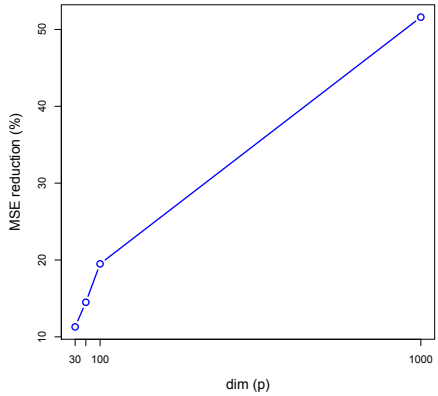

Figure 11: *Illustrative example. Estimation risk as a function of dimension. The y-axis represents the reduction in mean squared error of factor scores when the information about Z is used relative to a simple sparse factor model.*

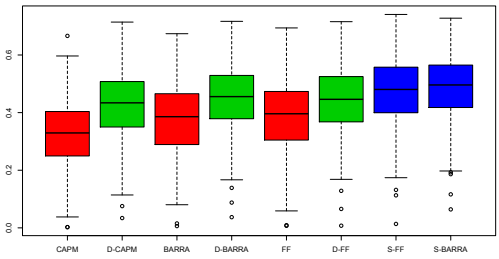

Figure 12: *Case Study. Boxplots of the percentage of variation explained by each model for all stocks. The red plots are based on simple linear regressions whereas the green represent the proposed model-based strategy. The blue plot refers to the better performing model, i.e, the Sparse-Dynamic BARRA.*

Our initial conjecture is somewhat validated by the performance of the sparse version of the BARRA and Fama-French specifications in the case study. At this point we have not been able to make use of the more complex and potentially interesting *Probit Sparse Factor Model*, as presented in Section 3.5, in modeling stock returns. To illustrate its potential, we now present a simulation exercise where we make use of a non-linear, non-dynamic version of the probit model.

We simulate data with different dimensions ($p = 30, 50, 100$ and 1000) using the loadings structure depicted in Figure 7 and sample size fixed at $n = 50$. In all examples, one "external" variable Z is associated in a non-linear fashion with the probability of inclusion in factor 2 (all models are defined with 2 factors) and a polynomial linear predictor was used in the probit model. The structure in Figure 7 leads to the conclusion that the probability association of a variable with factor 2 is

a non-linear function of Z as the values of the loadings are only significantly away from zero for variables with a large absolute value of Z.

Posterior means of the estimated loadings in a $p = 30$ dimensional example are also displayed in Figure 7. It is clear that trying to estimate the loadings without the information in Z is possible but leads to over shrinkage of the large elements of the loadings. This is a simple consequence of having only one parameter defining the inclusion probability which promotes an "averaging" effect to the baseline of inclusion. Small changes in the loadings may imply big changes in the estimation of factor scores and significant differences in the practical use of the model (as evidenced by the case study presented above). A summary of the estimation error associated with the factor scores appears in Figure 10 where it can be seen that the errors are much larger relatively when the information about Z is ignored.

5. CONCLUSIONS

We have focused on the use of a general dynamic factor model framework for the estimation of Financial Index Models where firm specific information is used to help uncover the relevant latent structure responsible for stock co-movements. Our conclusions are still preliminary but the case study demonstrates that small modeling modifications can lead to significant differences in the practical output of the models. This is our first attempt in exploring more carefully, from a statistical point of view, the very influential ideas related to the work of Fama and French. Building on this framework we hope to study additional, more complex, specifications that will hopefully lead to better performing covariance estimates and improved trading strategies. Moreover, by extending our approach to the entire set of stocks in the market we will be able to deliver more relevant factor scores that can be used as a tool in asset pricing models.

Finally, it is our view that the framework introduced here is more general than the financial problems discussed. Factor models are common place in many areas of scientific exploration and the ability to incorporate "external" information in the estimation of the latent structure can lead to more precise models of covariation.

REFERENCES

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation, *J. Business Econ. Studies* **18**, 338–357.

Carhart, M. (1997). On Persistence in mutual fund performance. *The Journal of Finance* **52**, 57–82.

Carvalho, C. M., Chang, J., Lucas, J., Wang, Q., Nevins, J. and West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103**, 1438–1456.

Carvalho, C.M., Johannes, M., Lopes, H.F and Polson, N. (2010). Particle Learning and Smoothing. *Statistical Science* (to appear).

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

Fama, E. (1970). Efficient capital markets: a review of theory and empirical work. *The Journal of Finance* **25**, 383–417.

Fama, E. and French, K. (1992). The cross-section of expected stock returns. *The Journal of Finance* **47**, 427–465.

Fama, E. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3–56.

Fama, E. and French, K. (1996). Multifactor explanations of asset pricing anomalies *The Journal of Finance* **51**, 55–84.

Frühwirth-Schnatter, S. and Lopes, H. F. (2010). Parsimonious Bayesian factor analysis when the number of factors is unknown. *Technical Report*. The University of Chicago Booth School of Business.

Jacquier, E., Polson, N. and Rossi, P. (1994). Bayesian analysis of stochastic volatility models. *J. Business Econ. Studies* **12**, 371–388.

Jegadeesh, N and Titman, S. (1993). Returns to buying winners and selling losers: implications for stock market efficiency. *The Journal of Finance* **48**, 65–91.

Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies* **65**, 361–393.

Lopes, H. F., Aguilar, O. and West, M. (2000). Time-varying covariance structures in currency markets. In Proceedings of the XXII Brazilian Meeting of Econometrics.

Lopes, H. F. and Carvalho, C. M. (2007). Factor stochastic volatility with time-varying loadings and markov switching regimes. *J. Statist. Planning and Inference* **137**, 3082–3091.

Lopes, H. F., Salazar E. and Gamerman, D. (2008). Spatial dynamic factor analysis. *Bayesian Anal.* **3**, 759–92.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14**, 41–67.

Pitt, M. and Shephard, N. (1999). Time varying covariances: a factor stochastic volatility approach. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 547–570.

Rosenberg, B. and McKibben, W. (1973). The prediction of systematic and specific risk in common stocks *The Journal of Financial and Quantitative Analysis* **8**, 317–333.

Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory* **13**, 341–360.

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

Tsay, R. (2005). Analysis of Financial Time Series. Chichester: Wiley

West, M. (2003). Bayesian factor regression models in the"large p, small n" paradigm. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press , 723–732.

DISCUSSION

MENDOZA, MANUEL (*Instituto Tecnológico Autónomo de México, Mexico*)

Let me start by thanking the authors for a nice and readable paper. They have taken us one step further along a fascinating road which started more than 40 years ago with Sharpe's paper. Explaining how the returns of assets in a financial market behave, as the authors have reminded us, is not only a matter of academic interest but also has enormous practical relevance since it is the basis for portfolio selection and, hence, the design of investment strategies. In this sense, research leading to a sound, adaptative and feasible model, able to accurately forecast returns within a reasonable time horizon, may well cause the authors to become not only prominent scholars, but also very wealthy. Under these circumstances, my first comment is that, just in case, we should keep an eye on these colleagues.

On a more technical note, I would like to recall that Sharpe's *Capital Asset Pricing Model* (CAPM), as well as other similar models, were originally proposed as *theoretical* explanations of a financial phenomenon rather than statistical tools for prediction. In fact, the CAPM asserts that –under *equilibrium* conditions– for each risky asset the expected return in excess over the risk-free asset must be proportional

to the expected return in excess over the same risk-free asset for the *market portfolio*. Thus,

$$(E(r) - r_0) = \beta\,(E(r_M) - r_0)$$

where the coefficient β may change with the risky asset. The above mentioned equilibrium conditions involve, for example, the existence of a common pure rate of interest available for all investors as well as the homogeneity of expectations among investors. With respect to these hypotheses, we may quote Sharpe (1964): *"Needless to say, these are highly restrictive and undoubtedly unrealistic assumptions"*. Despite this, when there are p risky assets in the market, and taking the CAPM for granted, a multivariate regression model has been adopted to explain the p-dimensional vector of returns in terms of the univariate return r_M,

$$r = \alpha + \beta\, r_M + \epsilon\,; \quad r^t = (r_1, ..., r_p).$$

Consequently, the vector of expected returns ($E(r) = \mu$) takes the form $\alpha + \beta\, E(r_M)$, and the corresponding $p \times p$ covariance matrix V is given by $\beta\beta^t\sigma_M^2 + \Sigma$, where $\Sigma = diag(\sigma_1^2, ..., \sigma_p^2)$. Thus, the problem of estimating the $p(p-1)/2$ different elements of V is reduced to the estimation of β, σ_M^2 and $\sigma_1^2, ..., \sigma_p^2$. This simplification is highly relevant since, for a portfolio $Q = \sum_{i=1}^{p} k_i a_i$, the expected return is given by $E(r_Q) = k^t\mu$, whereas the variance (risk) is $Var(r_Q) = k^tVk$ and, in accordance with Markowitz (1952), the portfolio selection problem is solved if the investor minimizes k^tVk for a fixed $k^t\mu$, or maximizes $k^t\mu$ for a fixed k^tVk. In any case, an estimate or forecast for the covariance matrix V of the future returns is required, and thus reduction from $p(p-2)/2$ to $2p$ parameters is essential.

Empirical applications of the regression model associated to the CAPM have shown that it is a rather poor statistical model (see Fama and French, 2004, for a recent discussion on this issue) and, as a natural consequence, some other models have been suggested as alternatives. In the regression setting, Rosenberg and McKibben (1973) explored the improvement of CAPM when other explanatory variables, apart from the market portfolio return, are included. Specifically, they used information from the firm associated to each asset. This approach may be reasonable in terms of prediction accuracy but it is not appropriate if the objective is to keep a low dimensional structure for the covariance matrix of returns (covariances among the p firms must be taken into account). Alternatively, in a number of papers, Fama and French, used the information from the firms to create *ad hoc* portfolios whose returns were then used as additional factors in a modified CAPM (see Fama and French 1993, 1996a and 1996b, for instance). They showed, with real data examples, that their model provided better forecasts than CAPM and, more importantly, allowed them to estimate the returns covariance matrix through a rather small number of parameters. The basic structure of this model is

$$r = \alpha + \beta\, r_M + \gamma\, r_A + \delta\, r_B + \epsilon$$

where A and B are *ad hoc* portfolios explaining variations in the returns that CAPM is unable to describe. Only a few years later, Pitt and Shephard (1999) and Aguilar and West (2000) introduced a Bayesian Dynamic Factor Model,

$$y_t = \theta_t + \boldsymbol{X}_t f_t + \epsilon_t,$$

where y_t is the vector of returns, f_t is a q-dimensional vector ($q << p$) of latent factors and $\boldsymbol{X}_t$ is a $p \times q$ unknown matrix of loadings. In particular, Aguilar and

West (2000) illustrate the model with some examples where the loadings matrix does not change with time. The innovation in this model is twofold. First, instead of defining some specific portfolios as factors explaining the common variation of the returns in the market, a set of latent factors is included. Second, the linear structure is assumed to be dynamic. I am not an economist but I might guess that the latter is, by far, the most relevant generalization from a theoretical point of view, since it allows the model to recognize that the equilibrium condition may not be reached in the market. Lopes and Carvalho (2007) explored this model in a more general situation with time varying loadings and jumps in the autoregressive model they used for the log-volatilities of the latent factors. In particular, for the loadings, those authors propose a first-order autoregressive evolution structure.

Now, in the paper we discuss here, Carvalho, Lopes and Aguilar introduce an even more general structure,

$$r_t = \alpha_t + \beta_t x_t + \boldsymbol{Z}_t f_t + \epsilon_t,$$

where r_t is a p-vector of returns, x_t is the market return, f_t is a q-vector of latent factors, and $\boldsymbol{Z}_t$ is a $p \times q$ time-varying matrix of loadings which is assumed to be given and defined as a function of observable data (the information used by Fama and French to build their *ad hoc* portfolios, for example). In addition, this model includes a random mechanism to decide, at each period of time, which factors have zero loadings. This is the idea of sparsity as introduced by West (2003) in connection to gene expression analysis. This an interesting model. Instead of replacing the market portfolio by a set of latent factors, it takes both sources of information into account. It is worth noticing that the dynamic nature of β_t, while introducing flexibility in the relationship between the return r_t and x_t, does not change the structure of the market portfolio (the relative weights in the linear combination of assets defining x_t remain fixed). On the other hand, the dynamic loadings matrix $\boldsymbol{Z}_t$ allows the relative weights for the factors to change over time. Moreover, the sparsity mechanism makes it possible to temporarily suppress the influence of a factor on a particular asset. This is a very general structure and includes as particular instances, among many others, dynamic counterparts of the CAPM (DCAPM), the model by Rosenberg and McKibben (DRM) and the three-factor model of Fama and French (DFF), as well as sparse versions of both, DRM and DFF (SDRM and SDFF).

One of the issues that deserves special attention when an elaborated structure like this is considered is that of identifiability. This topic has been addressed in the past (Aguilar and West 2000 and West 2003, for example) for some models of this type, but none of them involves simultaneously explanatory variables and latent factors. In addition, the sparse specification, specially when the inclusion probabilities are assumed to be function of the individual firm characteristics, might also require some constraints. It would be very helpful to see an extensive discussion of these topics.

The authors present a particular case study to show the type of results that can be obtained with their model. A real data set with $p = 350$ assets is analyzed and five models are considered (DCAPM, DRM, DFF, SDRM and SDFF) where the number of factors is $q = 1, 4, 4, 4$ and 4, respectively. There are several aspects of the analysis which are not completely clear to me. For instance, what are the specific prior distributions used in this example? For the sparse models, what is the prior used for $\pi_{j,t}$? In close relation with this, how is this prior updated? More specifically, what is the conditional independence structure of the posterior distribution? Are these probabilities related to other parameters in the model *a posteriori*?

The authors show the evolution of the β parameters over time for four companies in all models (Figure 2). I wonder, is this a general pattern in this example? If the dynamic version of the CAPM is better than DRM and DFF (although with a penalized criterion), I would expect more evidence against market equilibrium. In particular, I would expect to see something similar to the trends found in Lopes and Carvalho (2007) for the exchange rates example, where clearly the equilibrium condition for the market is not reached. In any case, is the pattern shown in Figure 2 shared, in general, by the other 346 firms? Do you have an interpretation for the scatter plot you get for the book-to-price factor scores in the case of the FF model (sparse vs non-sparse)? It is rather intriguing.

According to the specific model comparison procedure used in this example, the sparse models SDBM and SDFF are the best ranked models but, what can the authors tell us about their predictive abilities? This is a basic question if the results are to be used to design an investment strategy.

In a more general setting, although related to the results in the case study, I would like to know how an investment strategy could be developed on the basis of this model if the time-varying loadings, $\{\boldsymbol{Z}_t\}$, are treated as given. More specifically, how is the covariance matrix of a future vector of returns, r_{T+1}, estimated if it depends on $\boldsymbol{Z}_{T+1}$, which is assumed to be given but depends on the *future* firms information for which the model does not include an evolution component?

Let us recall again that CAPM was proposed as a theoretical explanation for the way financial markets behave, whereas the model proposed by the authors, as well as many of its predecessors, is an empirical structure whose aim is to accurately forecast the returns within a reasonable period of time for investment purposes. In this sense, I think this paper clearly illustrates the existence of two approaches to the portfolio selection problem. One uses an asset pricing model and thus involves some elements of financial theory. On the other hand, we have what Pástor (2000) calls the 'data-based' approach. Basically, this paper follows the second approach, and although the proposed model is rather general, it could be interesting to explore even more general and robust alternatives. In this direction, there is a huge amount of literature showing that returns as well as other financial data do not follow a normal distribution and several heavy-tails alternatives have been considered. In relation to this, is it possible to use another, more general distribution for the returns in this model (elliptical, for example)? See Hamada and Valdes (2008) for a related discussion. More in accordance with the new times, could this model be generalized to a semiparametric version?

Finally, it is worth noting that the CAPM has been extended in many ways. Some of these extensions remove the assumptions of a common pure rate of interest available for all investors and the homogeneity of expectations among investors. It so happens, however, that for most of these extensions no single portfolio of risky assets is optimal for every investor (see Perold 2004, for a related discussion). Maybe these ideas from financial theory could be used to propose more powerful statistical models for portfolio selection.

REPLY TO THE DISCUSSION

First we would to thank Prof. Mendoza for his kind works, encouraging comments and for clearly placing out work in the context of the financial literature regarding the CAPM and related models. One of our main goals with this paper was to translate the empirical versions of a few widely used asset-pricing models into an overarching statistical framework. We can only agree with your closing statements

and say that this is only the beginning of our efforts in tackling this problem, and in that sense, your suggestions are much appreciated.

It follows the reply to a some of your specific comments:

Identifiability. You are absolutely correct that identifiability is a potential issue in factor analysis. The decomposition of common variation into a matrix of factor loadings and a vector of factor scores allows many solutions and identifiability constraints are generally applied to the loadings matrix. In our set up, however, we are fixing the values of the elements in $\mathbf{Z}_t$ (for all t) and therefore we avoid any potential problem. To be sure, modifications of our approach might require additional identifiability conditions and we point the reader to the solutions proposed in Aguilar and West (2000) and Lopes and Carvalho (2007).

Priors. In all models we have used conditionally conjugate priors for all parameters. They are inverse-gamma for variances, betas for the inclusion probabilities and normals for all other coefficients. Whenever possible we used standard, weak-informative priors and made sure to access the sensitivity of our analysis to these choices. A few parameters, however, require more informative priors – in particular the variances in the evolution of the log volatilities are known to require informative priors (see for example Kim, Shephard and Chib1998).

Inclusion Probabilities. The update of the factor inclusion probabilities $\pi_{j,t}$ are very simple due the form of the model. Conditionally on the indicators of whether or not a variable is associated with a factor, i.e., if the factor loading in not zero, the posterior for $\pi_{j,t}$ is simple a beta distribution updated as usual. This step is exactly as it appears in West (2003).

Dynamic β's. We do observe that the β's for all 346 firms seem to have a dynamic nature. It is hard to illustrate this point in some many dimensions and it is perhaps harder to, in our framework, formally test the market-equilibrium hypothesis. This point is very relevant and we will attempt to address this question as we move forward with our research.

Figure 6. Yes, the second panel of Figure 6 is indeed puzzling! Our best guess for this result (which is very robust and holds with difference choices of priors) is that, by following the FF strategy, the values of factor loadings for both the "size" and "book-to-price" (when they are not zero) are the same. Therefore, by not imposing the zeros and trying to find its configurations we believe that these two factors are almost redundant. That would explain the clustering of factor scores near the origin.

ADDITIONAL REFERENCES

Aguilar, O. and West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business and Economic Statistics.* **18**, 338–357.

Fama, E. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* **33**, 3-56.

Fama, E. and French, K. (1996a). Multifactor explanations of asset pricing anomalies. *The Journal of Finance* **51**, 55–84.

Fama, E. and French, K. (1996b). The CAPM is wanted, dead or alive. *The Journal of Finance* **51**, 1974–1958.

Fama, E. and French, K. (2004). The Capital asset pricing model: theory and evidence. *Journal of Economic Perspectives* **18**, 25–46.

Hamada, M. and Valdez, E. A. (2008). CAPM and option pricing with elliptically countored distributions. *The Journal of Risk and Insurance* **75**, 387–409.

Lopes, H. F. and Carvalho, C. M. (2007). Factor stochastic volatility with time-varying loadings and markov switching regimes. *J. Statist. Planning and Inference* **137**, 3082–3091.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance* **7**, 77–91.

Pástor, L.(2000). Portfolio selection and asset pricing models. *The Journal of Finance* **55**, 179–223.

Perold, A. F. (2004). The Capital asset pricing model. *Journal of Economic Perspectives* **18**, 3–24.

Pitt, M. and Shephard, N. (1999). Time varying covariances: a factor stochastic volatility approach. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 547–570.

Rosenberg, B. and McKibben, W. (1973). The prediction of systematic and specific risk in common stocks. *The Journal of Financial and Quantitative Analysis*. **8**, 317–333.

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

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press , 723–732.