A Bayesian Multivariate Functional Dynamic Linear Model

Daniel R. Kowal\textsuperscript{1}, David S. Matteson\textsuperscript{2}, and David Ruppert\textsuperscript{3}

\textsuperscript{1}Cornell University, Department of Statistical Science, 301 Malott Hall, Ithaca, NY, 14853. Email: drk92@cornell.edu.

\textsuperscript{2}Cornell University, Department of Statistical Science and Department of Social Statistics, Ithaca, NY, 14853.

\textsuperscript{3}Cornell University, Department of Statistical Science and School of Operations Research and Information Engineering, Ithaca, NY, 14853.

August 21, 2014

Abstract

We present a Bayesian approach for modeling multivariate, dependent functional data. To account for the three dominant structural features in the data—functional, time dependent, and multivariate components—we extend hierarchical dynamic linear models for multivariate time series to the functional data setting. We also develop Bayesian spline theory in a more general constrained optimization framework. The resulting estimates are smooth and interpretable, and can be made common across multivariate observations for additional information sharing. The Bayesian framework permits joint estimation of these components, provides exact inference (up to MCMC error) on specific parameters, and allows generalized dependence structures. Sampling from the posterior distribution is accomplished with an efficient Gibbs sampling algorithm. We illustrate the proposed framework by modeling multi-economy yield curve data from the recent global financial crisis and discuss the flexibility, accuracy, and interpretability of the proposed methods.
1 Introduction

We consider a multivariate time series of functional data. Functional data analysis (FDA) methods are widely applicable, including diverse fields such as economics and finance (e.g., Hays, Shen, and Huang [2012]; brain imaging (e.g., Staicu, Crainiceanu, Reich, and Ruppert [2012]); chemometric analysis, speech recognition, and electricity consumption (Ferraty and Vieu [2006]); and growth curves and environmental monitoring (Ramsay and Silverman [2005]). Methodology for independent and identically distributed (iid) functional data has been well-developed, but in the case of dependent functional data, the iid methods are not appropriate. Such dependence is common, and can arise via multiple responses, temporal and spatial effects, repeated measurements, missing covariates, or simply because of some natural grouping in the data (e.g., Horváth and Kokoszka [2012]). Here, we consider two distinct sources of dependence: time-dependence for time-ordered functional observations and contemporaneous dependence for multivariate functional observations.

Suppose we observe multiple functions $Y_c(t)(\tau)$, $c = 1, \ldots, C$ at time points $t = 1, \ldots, T$. Such observations have three dominant features:

(a) For each $c$ and $t$, $Y_c(t)(\tau)$ is a function of $\tau \in \mathcal{T}$;

(b) For each $c$ and $\tau$, $Y_c(t)(\tau)$ is a time series for $t = 1, \ldots, T$; and

(c) For each $t$ and $\tau$, $Y_c(t)(\tau)$ is a multivariate observation with outcomes $c = 1, \ldots, C$.

To handle the ensuing model complexities, we develop and apply hierarchical Bayesian methods. In the Bayesian framework, we can explore convenient decompositions of the above structures using appropriate conditioning; for example, we may estimate the functional component (a) conditional on the other two dependence structures (b) and (c), and so on. Such decompositions can provide insight for incorporating appropriate techniques and concepts from relevant fields including FDA, nonparametric regression, and (multivariate) time series. In addition, we can include prior information based on the specific application, such as the
underlying theory and long history of recorded data available in fields like economics and finance. Then, using a Gibbs sampler (e.g., Casella and George [1992]), we can combine these elements and efficiently sample from the joint conditional distribution. All of the model components will be estimated simultaneously—without the need for a complex optimization algorithm. And by adopting the common assumption of conditional independence between levels of the hierarchy, we may substitute appropriate model structures at any level without substantially affecting the other levels. Such flexibility is relatively unique to the Bayesian approach, and is very useful in exploring the various structures encountered in applications, such as covariates, stochastic volatility, and change points.

As our primary contribution is in the FDA setting, we propose a novel method for estimating the functional structure of the data. We extend Bayesian spline theory to a more general constrained optimization framework, which in our case is necessary for parameter identifiability. We make our constraints explicit in the posterior distribution via an exponential tilt of the usual Bayesian spline posterior distribution, and demonstrate that the corresponding posterior mean is the solution to an appropriate optimization problem. Because we nest our approach within nonparametric spline theory and distribution-free constrained optimization theory, the proposed methods are widely applicable.

2 Motivation: Multi-Economy Yield Curves

An interesting example of multivariate, time-dependent functional data is given by yield curves \( Y_t^{(c)}(\tau) \) observed at dates \( t = 1, \ldots, T \) for economies \( c = 1, \ldots, C \). Generally speaking, for a given currency and level of risk of a debt, the yield curve describes the relationship between the interest rate and the length of the borrowing period (or time to maturity, \( \tau \)). For example, for U.S. Treasury bonds—which are approximately risk-free—the yield can be interpreted as the interest earned per year from purchasing a bond and then holding it until its maturity (Jarrow [2002]). Yield curves are extremely important in a variety of economic
and financial applications, such as evaluating economic and monetary conditions, pricing fixed-income securities, generating forward curves, computing inflation premiums, and monitoring business cycles (Bolder, Johnson, and Metzler 2004). Indeed, there is substantial interest in both forecasting and providing adequate descriptions of yield curve behavior (e.g, Diebold and Li 2006; Bolder et al. 2004).

As commonly shown in FDA, it has been well-established that yield curves are best modeled as functions of the maturity, rather than vectors containing the observed maturities (e.g., Jungbacker, Koopman, and Van der Wel 2013; Hays et al. 2012). For example, our data consist of (discrete) yield curve observations at \( m_t^{(c)} \) maturities for economy \( c \) on day \( t \). Multivariate methods would treat these observations as an \( m_t^{(c)} \)-dimensional vector, and then have to account for between-maturity dependence of this vector—in addition to the other dependence structures. Such treatment is usually inadequate, and ignores the fundamental structure implied by the application: yield curves are functions of the maturity, and therefore are interpretable over a continuous set of (potentially unobserved) maturities.

The dependences induced by the time-ordered and contemporaneous yield curve observations are also of interest themselves. Apart from common forecasting considerations, we may, for example, study how a specific segment of the yield curve—or even the shape of the yield curve—reacts to various monetary policies, and estimate how long such reactions persist. We may also compare these monetary policies across economies, or examine any associated lead/lag relationships or spillover effects. Such interesting investigations must both rely on and account for these additional dependence structures on the functional observations.

The paper proceeds as follows. In Section 3 we present the model in its most general form, then highlight an important submodel. In Section 4 we develop our (factor loading) curve estimation technique, beginning with splines and building up to a novel Bayesian approach for fitting smooth curves under constraints. In Section 5 we apply and discuss several variations of our model to multi-economy yield curve data and find that it provides highly competitive forecasts compared to other common estimators. We provide the details
of our Gibbs sampling algorithm and forecasting in the Appendix.

3 The General Model

Suppose we observe functions \( Y_t^{(c)}: \mathcal{T} \to \mathbb{R} \) at times \( t = 1, \ldots, T \) for outcomes \( c = 1, \ldots, C \), where \( \mathcal{T} \) is a closed interval on the real line. We present a model in the hierarchical dynamic linear model (DLM) framework of West and Harrison (1997) and Gamerman and Migon (1993), but allow for functional observations. We refer to the following model as the Functional Dynamic Linear Model (FDLM):

\[
\begin{align*}
Y_t(\tau) &= F(\tau)\beta_t + \epsilon_t(\tau), \quad \epsilon_t(\tau) \overset{\text{iid}}{\sim} N(0, E_t) \\
\beta_t &= A_t\theta_t + \nu_t, \quad \nu_t \overset{\text{iid}}{\sim} N(0, V_t) \\
\theta_t &= G_t\theta_{t-1} + \omega_t, \quad \omega_t \overset{\text{iid}}{\sim} N(0, W_t) \\
\int_{\tau \in \mathcal{T}} F'(\tau)F(\tau) \, d\tau &= I_{KC} \quad \text{for identifiability}
\end{align*}
\]

where \( Y_t(\tau) = (Y_t^{(1)}(\tau), Y_t^{(2)}(\tau), \ldots, Y_t^{(C)}(\tau))' \) is the \( C \)-dimensional vector of multivariate functional observations at time \( t \) evaluated at \( \tau \in \mathcal{T} \); \( F(\tau) \) is the \( C \times KC \) block diagonal matrix of \( K \)-dimensional row vectors of factor curves evaluated at \( \tau \in \mathcal{T} \), with \( K \) denoting the number of factors per outcome; \( \beta_t = (\beta_t^{(1)}, \ldots, \beta_t^{(1)}, \beta_t^{(2)}, \ldots, \beta_t^{(C)})' \) is the \( KC \)-dimensional vector of factors that serve as the time-dependent weights on the factor loading curves; \( A_t \) is the known \( KC \times p \) matrix of covariates at time \( t \), where \( p \) is the total number of covariates in the model; \( \theta_t \) is the \( p \)-dimensional vector of parameters associated with the covariates to adequately model the factors \( \beta_t \); \( G_t \) is the \( p \times p \) evolution matrix of the covariate parameters \( \theta_t \) at time \( t \); and \( \epsilon_t(\tau) \), \( \nu_t \), and \( \omega_t \) are mutually independent error vectors with variance matrices \( E_t \), \( V_t \), and \( W_t \), respectively.

To interpret this framework, first note that the data level of the model combines the functional component \( F(\tau) \) with the multivariate time series component \( \beta_t \). In scalar notation,
we can write the data level as

\[ Y_t^{(c)}(\tau) = \sum_{k=1}^{K} f_k^{(c)}(\tau) \beta_k^{(c),t} + \epsilon_t^{(c)}(\tau) \]  

in which \( \epsilon_t^{(c)}(\tau) \) are the elements of the vector \( \epsilon_t(\tau) \). In our construction, we can always write the data level of (1) as (2); simplifications for the other levels will depend on the choice of submodel. Nonetheless, there are three primary interpretations of the model, which provide insight into useful extensions and submodels.

First, we can view (2) as a basis expansion of the functional observations \( Y_t^{(c)} \), with models for the basis coefficients \( \beta_{k,t}^{(c)} \) to account for the additional dependence structures, including covariates and stochastic volatility. Using (2), our identifiability constraint on \( F(\tau) \) simplifies to \( \int_{\tau \in T} f_k^{(c)}(\tau) f_j^{(c)}(\tau) = 1(k = j) \) for all outcomes \( c \), where \( 1(\cdot) \) is the indicator function and \( k, j \in \{1, \ldots, K\} \). Since this constraint expresses orthonormality with respect to the \( L^2 \) inner product, we can interpret the set of factor loading curves \( \{f_1^{(c)}, \ldots, f_K^{(c)}\} \) as an orthonormal basis for the functional observations \( Y_t^{(c)} \). In contrast to common basis expansion procedures (e.g., B-splines and Fourier series) that assume the basis functions are known and only the coefficients need to be estimated (e.g., Bowsher and Meeks 2008), here we allow our basis functions \( f_k^{(c)} \) to be estimated from the data. As a result, the \( f_k^{(c)} \) will be more closely tailored to the data, which reduces the number of functions \( K \) needed to adequately fit the data.

Then, conditional on the \( f_k^{(c)} \), we can specify the \( \beta_{t}^{(c)} \) and \( \theta_{t}^{(c)} \)-levels of (1) to appropriately model the remaining dependence among the \( Y_t^{(c)} \).

Similarly, we can interpret (1) as a dynamic factor analysis, which is a common approach in yield curve modeling (e.g., Hays et al. 2012; Jungbacker et al. 2013). Under this interpretation, the \( \beta_{k,t}^{(c)} \) are dynamic factors and the \( f_k^{(c)} \) are factor loading curves; we will use this terminology for the remainder of the paper. Compared to a standard factor analysis, (1) has two major modifications: the factors \( \beta_{k,t}^{(c)} \) are dynamic and therefore have an accompanying (multivariate) time series model, and the \( f_k^{(c)} \) are functions rather than vectors. Since the
factor analysis requires an identifiability constraint, we select the constraint based on the functional nature of our factor loading curves.

Finally, (1) has strong connections to a hierarchical DLM. Typically, DLMs assume that the state and evolution matrices $F$ and $G_t$ are known, as well as the error variance matrices $E_t$, $V_t$ and $W_t$. Here, we treat $F$ as unknown, but impose the necessary identifiability constraints; $G_t$ can be handled similarly, depending on the choice of submodel. We also treat the error variance matrices as unknown, but typically there are simplifications available depending on the application and model choice; for example, the lower level error variances $V_t$ and $W_t$ depend on our choices for $\theta_t$ and $A_t$. Then, conditional on $F(\tau)$, $G_t$, and the error variances $E_t$, $V_t$ and $W_t$, (1) is a hierarchical DLM. This result is important, since it allows us to sample from the posterior of $\beta_t$ and $\theta_t$ using standard (hierarchical) DLM algorithms (e.g., Petris, Petrone, and Campagnoli, 2009) within our Gibbs sampler. As a result, efficient sampling of the factors $\beta_t$ and the covariate parameters $\theta_t$ requires minimal implementation effort.

To interpret the $\beta_t$- and $\theta_t$-levels of (1), it is most natural to adopt the hierarchical DLM interpretation. Specifically, we use $\beta_t$ and $\theta_t$ to model both the time series and multivariate structures in the data. The DLM framework is quite general: we can accommodate any traditional autoregressive integrated moving average (ARIMA) or vector autoregression (VAR) model, using an appropriate definition of $\beta_t$ and careful placement of zeros (West and Harrison, 1997). Furthermore, the hierarchical structure allows us to include covariates relevant to each outcome (or all outcomes) and incorporate additional dimension reduction or shrinkage through $\theta_t$. Although Gamerman and Migon (1993) suggest that $\text{dim}(\theta_t) < \text{dim}(\beta_t)$ for strict dimension reduction in the hierarchy, we relax this assumption to allow for covariate information. We present the model in its current form because it explicitly identifies the factors $\beta_t$ as the time-dependent weights on the factor loading curves and allows for a more general hierarchical structure.

Finally, there is the issue of the choice of $K$. In the yield curve application, two natural
choices are $K = 3$ and $K = 4$ for comparison with the common parametric yield curves models: the Nelson-Siegel model \cite{Nelson1987} and the Svensson model \cite{Svensson1994}, both of which can be expressed as submodels of \cite{1}. More formally, we can treat $K$ as a parameter and estimate it using reversible jump MCMC methods \cite{Green1995}, or select $K$ based on relevant information criteria.

\subsection{Common Factor Loading Curves for Multivariate Modeling}

An important submodel of \cite{1} is given by $f^{(c)}_k = f_k$, so that all outcomes share a common set of factor loading curves. In the basis interpretation of the FDLM, this corresponds to the assumption that the functional observations for all outcomes $Y^{(c)}_t$, $c = 1, \ldots, C$, $t = 1, \ldots, T$ share a common basis. We find this approach to be useful and intuitive, since it pools information across outcomes and suggests a more parsimonious model. Equally important, the common factor loading curves approach allows for direct comparison between factors $\beta^{(c)}_{k,t}$ and $\beta^{(c')}_{k,t}$ for outcomes $c$ and $c'$, since these factors serve as weights on the same factor loading curve (or basis function) $f_k$.

An example of this submodel is

\begin{equation}
\begin{cases}
Y^{(c)}_t(\tau) = \sum_{k=1}^{K} f_k(\tau) \beta^{(c)}_{k,t} + \epsilon^{(c)}_t(\tau), \quad \epsilon^{(c)}_t(\tau) \overset{\text{indep}}{\sim} N(0, \sigma^2_{(c)}) \\
\beta_{k,t} = \Psi_k \beta_{k,t-1} + \omega_{k,t}, \quad \omega_{k,t} \overset{\text{indep}}{\sim} N(0, W_{k,t})
\end{cases}
\end{equation}

where $\beta_{k,t} = (\beta^{(1)}_{k,t}, \ldots, \beta^{(C)}_{k,t})'$ is modeled as a $C$-dimensional vector autoregression (VAR) with time-dependent error variance, or volatility, $W_{k,t} = \text{diag}(\sigma^2_{(1),k,t}, \ldots, \sigma^2_{(C),k,t})$, for $k = 1, \ldots, K$. In this specification, we model the dependence between outcomes $c$ and $c'$ across time $t$ for fixed $k$, but assume independence between $\beta_{k,t}$ and $\beta_{j,t}$ to reduce the number of parameters.

In applications of multivariate time series, it is very common—and often necessary for proper inference—to include a model for the volatility \cite[e.g.,][]{Taylor1994, Harvey1994}. In this context, we introduce a model for the volatility (or variance) at each time point $t$ for each outcome $c$.
It is reasonable to suppose that applications of multivariate functional time series may also require volatility modeling. For example, in the yield curve application, a volatility model is indeed necessary to account for the observed volatility clustering. The hierarchical Bayesian approach of the FDLM seamlessly incorporates volatility modeling: conditional on the volatilities, standard DLM algorithms require no additional adjustments for posterior sampling. In our Gibbs sampler, such conditioning is automatic; all we need is a model for the volatility.

Within the Bayesian framework, it is most natural to use a stochastic volatility model (e.g., Kim, Shephard, and Chib, 1998; Chib, Nardari, and Shephard, 2002). Stochastic volatility models are parsimonious, which is important in hierarchical modeling, yet are highly competitive with more heavily parameterized GARCH models (Danielsson, 1998). In particular, for the nonzero elements $\sigma^2_{(c),k,t}$ of $W_{k,t}$, Kim et al. (1998) model the log-volatility $\log(\sigma^2_{(c),k,t})$ as a stationary AR(1) process (for fixed $c$ and $k$) and provide an efficient MCMC sampling algorithm; we provide more details in the Appendix. Using the full conditional distributions from the FDLM Gibbs sampler, we may implement the Kim et al. (1998) algorithm directly, which allows us to sample from the posterior distribution of the log-volatilities.

4 Factor Loading Curve Estimation

For the factor loading curves $f_k^{(c)}$, note that (2) has the form of a standard regression problem, conditional on the factors $\beta_{k,t}^{(c)}$. Naturally, we would like to model the factor loading curves in a smooth, flexible, and computationally appealing manner. Clearly, the latter two attributes are important for broader applicability and larger data sets—including larger $T$, larger $C$, and larger $m_t^{(c)}$. The smoothness requirement is fundamental as well: as documented in Jungbacker et al. (2013) and Jungbacker, Koopman, and van der Wel (2009), smoothness constraints can improve forecasting, despite the small biases imposed by such constraints.
Smooth curves also tend to be more interpretable, since gradual trends are usually easier to explain than sharp changes or discontinuities.

However, there are some additional complications. First, we must incorporate the identifiability constraints, preferably without severely detracting from the smoothness and goodness-of-fit of the factor loading curves. We also have \(K\) curves to estimate for each outcome—or perhaps \(K\) curves common to all outcomes—similar to the varying-coefficients model of [Hastie and Tibshirani (1993)], conditional on the factors \(\beta_{k,t}^{(c)}\). Finally, the observation points \(\tau\) for the functions \(Y_t^{(c)}\) are likely different for each outcome \(c\), and may also vary across time \(t\).

4.1 Splines

To model the factor loading curves, we use splines. Splines provide a broad and flexible framework for nonparametric and semiparametric regression (e.g., [Eubank, 1999; Wahba, 1990], [Wang, 2011], [Ruppert, Wand, and Carroll, 2003]). In the typical nonparametric regression setting, we observe \((x_i, y_i), i = 1, \ldots, n\) and suppose that \(y_i = g(x_i) + \epsilon_i\) with \(E(\epsilon_i) = 0\). The goal is to estimate \(g\) under few assumptions, e.g., \(g\) has two continuous derivatives. A common approach is to express \(g\) as a linear combination of known basis functions, and then estimate the associated coefficients by maximizing a (penalized) likelihood. This classical approach can be adapted to the Bayesian setting, which we discuss in Section 4.2. We will use B-spline basis functions for their numerical properties and easy implementation, but our methods can accommodate other bases as well. For now, we ignore dependence on \(c\) for notational convenience; this corresponds to either the univariate case \((C = 1)\) or \(C > 1\) with the factor loading curves estimated independently for each outcome. The common factor loading curve submodel suggested in Section 3.1 is considered in Section 4.4.

To fully specify the B-spline basis, we must select an order of the spline and a knot sequence. Following [Wand and Ormerod, 2008], we use cubic splines and the knot sequence \(a = \kappa_1 = \ldots = \kappa_4 < \kappa_5 < \ldots < \kappa_{M+4} < \kappa_{M+5} = \ldots = \kappa_{M+8} = b\), with \(\phi = (\phi_1, \ldots, \phi_{M+4})\)
the associated cubic B-spline basis, $M$ the number of interior knots, and $\mathcal{T} = [a, b]$. We will discuss our choice of knots at the end of this subsection. While we could allow each factor loading curve $f_k$, $k = 1, \ldots, K$ to have its own B-spline basis $\phi_k$ and accompanying sequence of knots, there is no obvious reason for doing so. Then explicitly, we write $f_k(\tau) = \phi'(\tau)d_k$ where $d_k$ is the $(M+4)$-dimensional vector of coefficients that needs to be estimated. Therefore, our function estimation problem is reduced to a vector estimation problem.

In classical nonparametric regression, we estimate $d_k$ by maximizing a penalized likelihood, or equivalently solving

$$
\min_{d_k} (-2 \log [Y|d_k] + \lambda_k \mathcal{P}(d_k)) \quad (4)
$$

where $[Y|d_k]$ is the likelihood, $\mathcal{P}$ is a convex penalty function, and $\lambda_k > 0$. We write (4) as a log-likelihood multiplied by $-2$ so that for a Gaussian likelihood, (4) is simply a penalized least squares objective. To penalize roughness, a standard choice for $\mathcal{P}$ is the $L^2$-norm of the second derivative of $f_k$, which can be written in terms of $d_k$:

$$
\mathcal{P}(d_k) = d_k' \Omega_{\phi} d_k = \int_{\tau \in \mathcal{T}} (\dddot{f}_k(\tau))^2 d\tau \quad (5)
$$

where $\dddot{f}_k$ denotes the second derivative of $f_k$ and $\Omega_{\phi} = \int_{\mathcal{T}} \phi'(\tau)\phi''(\tau)d\tau$, which is easily computable for B-splines. With this choice of penalty, (4) balances goodness-of-fit with smoothness, where the trade-off is determined by $\lambda_k$.

Since $\mathcal{P}$ is a quadratic in $d_k$, (4) is straightforward to solve for many likelihoods, in particular a Gaussian likelihood. Letting $\hat{d}_k$ be this solution, we can estimate $f_k(\tau)$ for any $\tau \in \mathcal{T}$ with $\hat{f}_k(\tau) = \hat{\phi}'(\tau)\hat{d}_k$. For a general knot sequence, the resulting estimator $\hat{f}_k$ is an O’Sullivan spline, or $O$-spline, introduced by O’Sullivan (1986) and explored in Wand and Ormerod (2008). In the special case of $M = T$ in which there is a knot at every observation point, $\hat{f}_k$ is a natural cubic smoothing spline (e.g., Green and Silverman, 1993). Alternatively, if we choose a sparser sequence of knots ($M < T$) and take $\lambda_k \to 0$, ...
\( \hat{f}_k \) is a regression spline (e.g., Ramsay and Silverman 2005). Therefore, O-splines have two important types of splines as special cases. In addition, O-splines are numerically stable, possess natural boundary properties, and can be computed efficiently, especially for \( M < T \); see Wand and Ormerod (2008) for a more thorough discussion.

In the multivariate functional time series setting, the knot sequence must be chosen with some care. For example, we may observe \( Y_t^{(c)} \) and \( Y_t^{(c')} \) at different points; i.e., the observation points can vary across outcomes or time, or both. In the yield curve data, the number and location of the observed maturities are different across economies, and for a given economy occasionally vary across time. Therefore, smoothing splines are less appealing: knot placement must be determined by a common set of observation points, which would ignore all of the non-common data points. We prefer O-splines with a quantile-based placement of knots such as the default method described in Ruppert et al. (2003). This approach is responsive to the location of observation points in the data yet is computationally inexpensive.

4.2 Bayesian Splines

Splines also have a convenient Bayesian interpretation (e.g., Wahba 1978, 1983, 1990; Van der Linde 1995; Gu 1992; Berry, Carroll, and Ruppert 2002). Returning to (4), we notably have a likelihood term and a penalty term, where the penalty is a function only of the vector of coefficients \( d_k \) and known quantities. Therefore, conditional on \( \lambda_k \), the term \( \lambda_k \mathcal{P}(d_k) \) provides prior information about \( d_k \), for example that \( f_k = \phi'd_k \) is smooth. Under this general interpretation, (4) combines the prior information with the likelihood to obtain an estimate of \( d_k \). A natural Bayesian approach is therefore to construct a prior for \( d_k \) based on the penalty \( \mathcal{P} \), in particular so that the posterior mode of \( d_k \) is the solution to (4). For the most common settings in which the likelihood is Gaussian and the penalty \( \mathcal{P} \) is (5), the posterior distribution of \( d_k \) will be normal, so the posterior mean will also solve (4).

To construct a prior from \( \mathcal{P} \), we use the partially informative normal distribution of
A random vector \( z \) is partially informative normal, written \( z \sim PIN(\mu, \Sigma) \), if the density of \( z \) is proportional to \( |\Sigma|^{1/2} \exp(-\frac{1}{2}(z - \mu)'\Sigma(z - \mu)) \), where the operator \(|\cdot|_+\) denotes the product of the nonzero eigenvalues of \( \Sigma \), which is restricted to be symmetric and positive semidefinite. If \( \Sigma \) is positive definite, then we return to the multivariate normal case with \( z \sim N(\mu, \Sigma^{-1}) \); if \( \Sigma \) is positive semidefinite but not positive definite, then the distribution is improper. For the penalty (5), we select the prior \( d_k \sim PIN(0, \lambda_k \Omega_{\varphi}) \), for which the resulting posterior mode is clearly the solution to (4).

Note that this prior is improper: constant and linear functions are unpenalized by \( P \), so the prior is flat over this space and thus \( \Omega_{\varphi} \) is not full rank. We could instead place a diffuse but proper distribution over the space of constant and linear functions to obtain a proper prior for \( d_k \), but in practice the effect on the posterior distribution is negligible.

To incorporate Bayesian spline methods into the FDLM, we must specify the likelihood of the model. Since we assume conditional independence between levels of (1), our conditional likelihood for the factor loading curves is simply (2), but we ignore dependence on \( c \) for now:

\[
Y_t(\tau) = \sum_{k=1}^{K} \beta_{k,t} f_k(\tau) + \epsilon_t(\tau) = \sum_{k=1}^{K} \beta_{k,t} \varphi'(\tau) d_k + \epsilon_t(\tau)
\]  

where \( \epsilon_t(\tau) \overset{iid}{\sim} N(0, \sigma^2) \) for simplicity; the results are similar for more sophisticated error variance structures. In particular, (6) describes the distribution of the functional data \( Y_t \) given the factor loading curves \( f_k \) (or \( d_k \)), also conditional on \( \beta_{k,t} \) and \( \sigma^2 \).

Under the likelihood of model (6) and the penalty (5), the solution to (4) conditional on \( d_j, j \neq k \) is given by \( \hat{d}_k = V_k u_k \) where \( V_k^{-1} = \sum_{t=1}^{T} \sum_{\tau \in \mathcal{T}_t} \sigma^{-2} \beta_{k,t} \varphi(\tau) \varphi'(\tau) + \lambda_k \Omega_{\varphi} \), \( u_k = \sum_{t=1}^{T} \sum_{\tau \in \mathcal{T}_t} \sigma^{-2} \beta_{k,t} (Y_t(\tau) - \sum_{j \neq k} \beta_{j,t} f_j(\tau)) \varphi(\tau) \), and \( \mathcal{T}_t \subset \mathcal{T} \) denotes the discrete set of \( |\mathcal{T}_t| = m_t \) observation points for \( Y_t \) at time \( t \), which may vary across \( t \). Note that if \( \mathcal{T}_t = \mathcal{T}_1 \) for \( t = 2, \ldots, T \), then \( V_k \) and \( u_k \) may be rewritten more conveniently in vector notation. Most importantly for our purposes, under the same likelihood induced by (6) and the prior \( d_k \sim PIN(0, \lambda_k \Omega_{\varphi}) \), the posterior distribution of \( d_k \) is multivariate normal with
mean $\hat{d}_k$ and variance $V_k$. For convenient computations, Wand and Ormerod (2008) provide an exact construction of $\Omega_\phi$ and suggest efficient algorithms for $\hat{d}_k$ based on the Cholesky decomposition; we provide more details in the Appendix.

In the Bayesian setting, the smoothing parameter $\lambda_k$ has a natural interpretation: it is the prior precision associated with the penalty $\mathcal{P}$. In our case, $\lambda_k$ is the prior precision that $f_k$ is smooth. Therefore, we can model $\lambda_k$ as a precision, which provides a natural and data-driven method for estimating the smoothing parameter. The associated likelihood is $PIN(0, \lambda_k \Omega_\phi)$; details on the sampling and choice of prior are provided in the Appendix.

### 4.3 Constrained Factor Loading Curves

We extend the Bayesian spline approach to accommodate the necessary identifiability constraints for the FDLM. In particular, we impose the orthonormality constraints $\int_T f_k(\tau)f_j(\tau) = 1(k = j), \quad k, j \in \{1, \ldots, K\}$. The unit-norm constraint preserves identifiability with respect to scaling, i.e., relative to the factors $\beta_{k,t}$. The orthogonality constraints distinguish between pairs of factor loading curves, and in our approach identifies the factor loading curves with distinct posterior distributions.

While other identifiability constraints are available for the factor loading curves, orthonormality is appealing for a number of reasons. As discussed in Section 3, the orthonormality constraint suggests that we can interpret the factor loading curves $\{f_1, \ldots, f_K\}$ as an orthonormal basis for the functional observations $Y_t$. As such, the orthogonality constraint helps eliminate any information overlap between factor loading curves, which keeps the total number of necessary factor loading curves to a minimum. Furthermore, the unit norm constraint allows for easier comparisons among factor loading curves. Of course, the factor loading curves will be weighted by the factors $\beta_{k,t}$, so they can still have varying effects on our estimation of $Y_t$. Finally, using the previously defined B-spline basis functions, we can
write the constraints conveniently in terms of the vectors $d_k$ and $d_j$:

$$\int_{\tau \in T} f_k(\tau) f_j(\tau) \, d\tau = \int_{\tau \in T} \phi'(\tau) d_k \phi'(\tau) \, d\tau = d_k' J_\phi d_j = 1 (k = j)$$

(7)

where $J_\phi = \int_{\tau \in T} \phi(\tau) \phi'(\tau) \, d\tau$, which can be easily computed for B-splines. As with $\Omega$, $J_\phi$ only needs to be computed once, prior to any MCMC sampling.

The addition of an orthogonality constraint to a (penalized) least squares problem has an intuitive regression-based interpretation, which we present in the following theorem:

**Theorem 1.** Let $\hat{d} = Vu$ be the solution to the penalized least squares objective $\sigma^{-2} \sum_{i=1}^{n} (y_i - Z_i' d)^2 + \lambda d' \Omega d$, where $y_i \in \mathbb{R}$, $d$ is an unknown $(M + 4)$-dimensional vector, $Z_i$ is a known $(M + 4)$-dimensional vector, $\Omega$ is a known $(M + 4) \times (M + 4)$ positive-definite matrix, and $\sigma^2, \lambda > 0$ are known scalars. Consider the same objective, but subject to the $J$ linear constraints $d' L = 0$ for $L$ a known $(M + 4) \times J$ matrix. The solution is $\tilde{d} = V \tilde{u}$, where $\tilde{u}$ is the vector of residuals from the generalized least squares regression $u = L \Lambda + \delta$ with $\mathbb{E}(\delta) = 0$ and $\text{Var}(\delta) = V$.

Proof. The Lagrangian is $\mathcal{L}(d, \Lambda) = \sigma^{-2} \sum_{i=1}^{n} (y_i - Z_i' d)^2 + \lambda d' \Omega d + d' L \Lambda$, where $\Lambda$ is the $J$-dimensional vector of Lagrange multipliers associated with the $J$ linear constraints. It is straightforward to minimize $\mathcal{L}(d, \Lambda)$ with respect to $d$ and obtain the solution $\tilde{d} = V \tilde{u} = V(u - L \Lambda)$. Similarly, solving $\nabla \mathcal{L}(\tilde{d}, \Lambda) = 0$ for $\Lambda$ implies that $\Lambda = (L' VL)^{-1} L' V u$, which is the solution to the generalized least squares regression of $u$ on $L$ with error variance $V$. 

The result is interpretable: to incorporate linear constraints into a penalized least squares regression, we find $\tilde{u}$ nearest to $u$ under the inner product induced by $V$ among vectors in the space orthogonal to $\text{Col}(L)$. Hence, in our setting, extending (4) under a Gaussian likelihood to accommodate the (linear) orthogonality constraints $d_k' J_\phi d_j = 0$ may be described via a regression of the unconstrained solution on the orthogonality constraints. Note that the orthogonality constraints are linear, while the unit norm constraint is nonlinear; we address
this at the end of this subsection.

Since our Gibbs sampler allows us to condition on \( d_j \) for \( j \neq k \), we can impose the orthogonality constraints \textit{sequentially}. Specifically, we require that each factor loading curve \( f_k \) is orthogonal to \( f_j \), but \textit{only for} \( j < k \). Of course, when we reach the final factor loading curve \( f_K \), the joint orthonormality constraint will be satisfied across all \( k = 1, \ldots, K \). The advantage is that the number of constraints we impose on each factor loading curve is kept small. Instead of the \( K - 1 \) constraints associated with orthogonality for all \( j \neq k \), we have only \( k - 1 \) constraints for each factor loading curve \( f_k \). For example, in posterior expectation, \( f_1 \) will minimize (4) as a unit-norm O-spline, \( f_2 \) will do the same subject only to \( \int_{\tau \in T} f_1(\tau)f_2(\tau)d\tau = 0 \), and so on. Moreover, the sequential nature of the constraints suggests a natural ordering of the \( f_k \). This feature is extremely important because it helps guard against label-switching problems in the MCMC sampling.

Using the results of Theorem 1, we propose an extension of the unconstrained Bayesian splines of Section 4.2 to incorporate the identifiability constraints. We can rewrite the sequential orthogonality constraints \( d_k'J\phi d_j = 0 \) for \( j < k \) as the linear constraints in Theorem 1 with \( L_1:(k-1) = (J\phi d_1, \ldots, J\phi d_{k-1}) \) and \( J = k - 1 \). It follows from Theorem 1 that the solution to (4) under the likelihood of model (6), the penalty (5), and subject to the linear constraints \( d_k' L_1:(k-1) = 0 \) is given by \( \tilde{d}_k = V_k \tilde{u}_k \), where \( \tilde{u}_k = u_k - L_1:(k-1) \Lambda_1:(k-1) \) and \( \Lambda_1:(k-1) = (L_1':(k-1)V_k L_1:(k-1))^{-1}L_1':(k-1) V_k u_k \) is the vector of Lagrange multipliers associated with the orthogonality constraints. Therefore, akin to Section 4.2, a natural full conditional posterior distribution for \( d_k \) is multivariate normal with mean \( \tilde{d}_k \) and variance \( V_k \).

The posterior has a fascinating interpretation: we can think of the usual posterior distribution for Bayesian splines as being exponentially tilted by the orthogonality constraint from the Lagrangian. Recall that for a density function \( g_0(y) \) and its corresponding (finite) cumulant generating function \( b_{cgf}(\eta) \), \( g(y|\eta) \equiv g_0(y) \exp(y'\eta - b_{cgf}(\eta)) \propto g_0(y) \exp(y'\eta) \) is in the exponential family of distributions. Hence, if we take any density with a finite cumulant generating function and exponentially tilt it by some parameter, we will obtain a density
from the exponential family of distributions. In particular, tilting a normal distribution by \( \eta \) simply results in another normal distribution with the mean shifted by \( \eta \)—similar to the shift in mean by \(-V_k L_{1:(k-1)} \Lambda_{1:(k-1)}\) above to account for the orthogonality constraints.

There remains the issue of the unit norm constraint. Because we omit this constraint, the conditions of Theorem 1 are satisfied and the full conditional posterior distribution of \( d_k \) is normal, both of which are convenient results. Yet the unit norm constraint does not change the shape of the curve; it only affects the scaling of \( f_k \). Therefore, we can sample identifiable factor loading curves by drawing \( d_0 \sim N(V_k \tilde{u}_k, V_k) \) and then normalizing: \( d_k^* = d_0 / \sqrt{d_0' J \phi d_0} \). This normalization step is interpretable, corresponding to a projection of a normal distribution onto the unit sphere. Furthermore, we could absorb the normalizing constant \( \sqrt{d_0' J \phi d_0} \) into \( \beta_{k,t} \) for \( t = 1, \ldots, T \) (and the square into \( \lambda_k \)) without changing the value of the associated Lagrangian—or the shape of the resulting factor loading curve. In practice, this step is unnecessary: after the normalization of \( d_0 \), we immediately sample from the full conditional distribution of \( \beta_t \), which depends on \( d_k^* \) but not on the past MCMC sample values of \( \beta_t \). Specifically, we sample the \( \beta_t \) conditional on \( d_k^* \) rather than \( d_0 \), so the full conditional posterior distribution of \( \beta_t \) automatically accounts for the normalization step. Sampling the \( \lambda_k \) is similar, which we discuss in the Appendix.

4.4 Multiple Outcomes: Common Factor Loading Curves

Now we reintroduce dependence on \( c \) for the factor loading curves \( f_k^{(c)} \). If \( C = 1 \) or we wish to estimate independent factor loading curves for each outcome \( c \), we can sample from the relevant posterior distributions separately for \( c = 1, \ldots, C \) using the methods of Section 4.3. The more interesting case is the common factor loading curves model, i.e., \( f_k^{(c)} = f_k \), as in Section 3.1. In basis notation, this implies \( f_k^{(c)}(\tau) = \phi'_{(c)}(\tau) d_k^{(c)} = f_k(\tau) \). However, since the factor loading curves for each outcome are identical, it is reasonable to assume that they have the same vector of basis functions \( \phi \), so \( f_k^{(c)} = f_k \) is equivalent to \( d_k^{(c)} = d_k \). Moreover, by writing \( f_k^{(c)}(\tau) = \phi'(\tau) d_k \) we can use all of the \( m_t^{(c)} \) observation points for outcome \( c \) at
time \( t \), yet our parameter of interest \( \mathbf{d}_k \) will only be \((M + 4)\)-dimensional, with \( M < m_t^{(c)} \).

Modifying our previous approach, we use the likelihood of model (2) with the simple error distribution \( \epsilon_t^{(c)}(\tau) \sim N(0, \sigma_{(c)}^2) \). The implied full conditional posterior distribution for \( \mathbf{d}_k \) is again \( N(\mathbf{V}_k\tilde{\mathbf{u}}_k, \mathbf{V}_k) \), but now with

\[
\mathbf{V}_k = \sum_{c=1}^C \sum_{t \in T(c)} \sum_{\tau \in T_t^{(c)}} \sigma_{(c)}^{-2} \beta_{k,t}^{(c)} Y_t^{(c)}(\tau) - \sum_{j \neq k} \beta_{j,t}^{(c)} f_j(\tau) \phi(\tau),
\]

For full generality, we allow the (discrete) set of times \( T^{(c)} \) to vary for each outcome \( c \) and the (discrete) set of observation points \( T_t^{(c)} \) to vary with both time \( t \) and outcome \( c \), with \( |T_t^{(c)}| = m_t^{(c)} \). Note that we reuse the same notation to emphasize the similarity of the multivariate results to the univariate (or separate factor loading curve) results; for example, \( \Lambda_{1:(k-1)} \) is defined as before, but with the new values of \( \mathbf{V}_k \) and \( \mathbf{u}_k \).

5 Data Analysis and Results

We jointly analyze weekly yield curves provided by the Federal Reserve (Fed), the Bank of England (BOE), the European Central Bank (ECB), and the Bank of Canada (BOC; Bolder et al. 2004) from late 2004 to early 2014 \( (T = 491 \text{ and } C = 4) \). In each case, the yield curves are estimated differently: the Fed uses quasi-cubic splines, the BOE uses cubic splines with variable smoothing parameters (Waggoner, 1997), the ECB uses Svensson curves, and the BOC uses exponential splines (Li, DeWetering, Lucas, Brenner, and Shapiro, 2001).

Therefore, the functional observations have already been smoothed, although by different procedures. These data are publicly available and published daily on the respective central bank websites—and as such, we treat them as reliable estimates of the yield curves.

Each economy estimates the yield curve for a different set of maturities, with the following values of \( \max_{t=1,\ldots,T^{(c)}}(m_t^{(c)}) \): 11 (Fed), 100 (BOE), 354 (ECB), and 120 (BOC), ranging from 1-3 months up to 300-360 months. Typically, when \( m_t^{(c)} < \max_{t=1,\ldots,T^{(c)}}(m_t^{(c)}) \) for week \( t \) and economy \( c \), the unobserved maturities occur at the very short end (small \( \tau \)) or the very long end (large \( \tau \)) of the yield curve. Regardless, our model handles these cases automatically,
without the need for additional missing data procedures. For an example of the yield curves, see Figure 1. We plot the multi-economy yield curves observed at adjacent times on July 29, 2011 and August 5, 2011. The differences are subtle just one week apart.

![Multi-Economy Yield Curves on 2011-07-29 and 2011-08-05](image)

Figure 1: Multi-economy yield curves, one week apart. The dashed lines indicate the second week.

The literature on yield curve modeling is extensive. Most commonly, yield curve models rely on the Nelson-Siegel parameterization, often within a state space framework (e.g., Diebold and Li 2006, Diebold, Rudebusch, and Aruoba 2006, Diebold, Li, and Yue 2008, Koopman, Mallee, and Van der Wel 2010). Many Bayesian models also use the Nelson-Siegel or Svensson parameterizations (e.g., Laurini and Hotta 2010, Cruz-Marcelo, Ensor, and Rosner 2011). However, the Nelson-Siegel parameterization does not extend to other applications, and often requires solving computationally intensive nonlinear optimization problems. More similar to our approach are the Functional Dynamic Factor Model (FDFM) of Hays et al. (2012) and the Smooth Dynamic Factor Model (SDFM) of Jungbacker et al.
and Jungbacker et al. (2009), both of which feature nonparametric functional components within a state space framework. The FDFM cleverly uses an EM algorithm to jointly estimate the functional and time series components of the model. However, the EM algorithm makes more sophisticated (multivariate) time series models more challenging to implement, and introduces some difficulties with generalized cross-validation (GCV) for estimation of the nonparametric smoothing parameters. The SDFM avoids GCV and instead relies on hypothesis tests to select the number and location of knots—and therefore determine the smoothness of the curves. However, this suggests that the smoothness of the curves depends on the significance levels used for the hypothesis tests, of which there can be a substantial number as \( m^{(c)} \), \( C \), or \( T \) grow large. By comparison, our smoothing parameters naturally depend on the data through the posterior distribution, which notably does not create any difficulties for inference.

5.1 Competing Estimators

We consider two alternative models for comparison: a random walk (RW) model and the Diebold and Li (DL; 2006) model. In the RW model, we use the estimator \( \tilde{Y}^{(c)}(\tau) \equiv Y^{(c)}_{t-1}(\tau) \) for each economy \( c \). Since the yield curves are observed weekly rather than monthly, this is a reasonable baseline: in Figure 1, the solid lines forecast the dashed lines moderately well. The DL model uses a two-stage estimation approach to forecasting in which they estimate the Nelson-Siegel linear parameters at every time point \( t \) and then fit a VAR to the ensuing estimates, all while keeping the nonlinear parameter fixed (at 0.0609). Therefore, the time series structure is not accounted for in estimation; it is only utilized in post hoc forecasting.

5.2 Forecasting and Estimation

We compute one-step yield curve forecasts for the FED, the BOE, the ECB, and the BOC to assess the out-of-sample performance of the FDLM. Specifically, we estimate one-step forecasts for the 52 weeks following each of the start dates 1/1/2008, 6/1/2010, and 1/1/2013.
For example, if $T^*$ corresponds to 1/1/2008, we forecast $Y_{T^*+1}^{(c)}$ given $Y_{1:T^*}^{(c)}$, then forecast $Y_{T^*+2}^{(c)}$ given $Y_{1:(T^*+1)}^{(c)}$ and so on, up to the date $T^* + 52$, corresponding to 1/1/2009. By altering $T^*$, we change the amount of data with which we can estimate the parameters and provide a new set of forecast periods. For these starting dates, the initial yield curve sample sizes $T^*$ are 173, 299, and 434, respectively.

The details of the FDLM forecasting are in the Appendix. Most notably, we use approximate one-step forecasts for computational convenience, which likely inflates the FDLM forecast errors. For the RW and DL models, it is straightforward to recompute exact one-step forecasts, assuming the nonlinear parameter is fixed in the DL model. To evaluate the forecasts, we compute mean squared errors (MSEs) and integrated squared errors (ISEs) for each forecast period. Specifically, we define \[ \text{MSE}(\hat{Y}^{(c)}) \equiv (52)^{-1} \sum_{\ell=1}^{52} \sum_{\tau \in \mathcal{T}} (\hat{Y}_{T^*+\ell}^{(c)}(\tau) - Y_{T^*+\ell}^{(c)}(\tau))^2/m_t^{(c)} \] and \[ \text{ISE}(\hat{Y}^{(c)}) \equiv (52|\mathcal{T}|)^{-1} \sum_{\ell=1}^{52} \int_{\mathcal{T}} (\hat{Y}_{T^*+\ell}^{(c)}(\tau) - Y_{T^*+\ell}^{(c)}(\tau))^2 d\tau, \] where $|\mathcal{T}| = b - a$ and $\hat{Y}_t^{(c)}$ is an estimate of $Y_t^{(c)}$. While the ISE is a more natural evaluator for our functional observations, the MSE may be appropriate if we are particularly interested in the performance of our estimator at the observations points.

In Table 1, we present the ISEs and MSEs for one-step forecasts from the RW model, the DL model, and four variations of the FDLM: $K \in \{3, 4\}$ with separate factor loading curves for each economy and $K \in \{4, 5\}$ with common factor loading curves (see Section 3.1). For each FDLM, we use independent ARIMA(1,1,0) models for the factors with the \cite{Kimetal1998} stochastic volatility model for the error variances. For the VAR in the DL model, we include 2 lags for reasonable comparison with the FDLM ARIMA specification. The $K = 3$ FDLM is most comparable to the DL model based on the number of factor loading curves, and outperforms the DL model in overall ISE. Naturally, we would expect the FDLMs with larger $K$ to perform better, which tends to be true overall. However, the effect of increasing $K$ is not uniform for all economies. For example, when we increase the number of common factor loading curves from $K = 4$ to $K = 5$, the forecast errors for the BOC decrease substantially more than for other economies, suggesting that $f_5$ is
exceptionally important for the BOC. Similarly, certain economies react very favorably to common factor loading curves: the BOE and ECB forecast errors are actually smaller for $K = 4$ common factor loading curves than for $K = 4$ economy-specific factor loading curves. Therefore, the common factor loading curve approach of Section 3.1 is not only useful for parsimony and interpretability, but also may improve accuracy.

Although the RW model is typically not competitive with the FDLMs or the DL model, large ISEs from the RW model can indicate periods of relative yield curve volatility for each economy. For example, the RW ISEs suggest that the Fed yield curves were highly volatile in 2008-2009, during the latter stages of the financial crisis. The FDLMs performed exceptionally well during this period, and tended to outperform the DL model during periods of large RW ISE. Overall, the FDLMs provide highly competitive forecasts.

The FDLM is perhaps even more useful for inference, since we obtain exact (up to MCMC error) posterior distributions for all terms in the model. We fit an FDLM to all $T = 491$ yield curve observations for each economy using submodel (3) with $K = 5$, and the results are encouraging: if we estimate each yield curve with the posterior mean of $\sum_{k=1}^K f_k(\tau)\hat{\beta}_{k,t}^{(c)}$ and compare it to the DL estimate, the FDLM is superior by an order of magnitude in both ISE and MSE for all economies.

We include some interesting plots from this FDLM. In Figure 2, we plot the common factor loading curves $f_k$, along with the analogous Nelson-Siegel curves. We can interpret these $f_k$ as estimates of the time-invariant underlying functional structure of the yield curves shared by the Fed, the BOE, the ECB, and the BOC. Therefore, the functional structure is not tied to a specific economy—or to an economy’s method of yield curve estimation. The factor loading curves are very smooth, with $f_1, f_2, f_3$ somewhat similar to the Nelson-Siegel curves, but with different curvatures. Notably, most of the variability is near the short-term boundary (one-month), which agrees with the FDLM forecasting results: comparing the contributions of individual maturities to the forecast MSEs, we find that the FDLMs perform consistently well for mid- to longer-term maturities, and struggle the most near the...
<table>
<thead>
<tr>
<th>Economy</th>
<th>Model</th>
<th>ISE</th>
<th></th>
<th></th>
<th></th>
<th>MSE</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>'08-'09</td>
<td>'10-'11</td>
<td>'13-'14</td>
<td>Total</td>
<td>'08-'09</td>
<td>'10-'11</td>
<td>'13-'14</td>
<td>Total</td>
</tr>
<tr>
<td>The Fed</td>
<td>Diebold-Li</td>
<td>0.0498</td>
<td>0.0144</td>
<td>0.0187</td>
<td>0.0829</td>
<td>0.0096</td>
<td>0.0166</td>
<td>0.0242</td>
<td>0.0904</td>
</tr>
<tr>
<td></td>
<td>Random Walk</td>
<td>0.0927</td>
<td>0.0114</td>
<td>0.0193</td>
<td>0.1232</td>
<td>0.1124</td>
<td>0.0071</td>
<td>0.0091</td>
<td>0.1286</td>
</tr>
<tr>
<td></td>
<td>K = 3; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0372</td>
<td>0.0165</td>
<td>0.0077</td>
<td>0.0614</td>
<td>0.0602</td>
<td>0.0122</td>
<td>0.0067</td>
<td>0.0792</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0328</td>
<td>0.0132</td>
<td>0.0060</td>
<td>0.0520</td>
<td>0.0532</td>
<td>0.0083</td>
<td>0.0044</td>
<td>0.0660</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0423</td>
<td>0.0116</td>
<td>0.0093</td>
<td>0.0632</td>
<td>0.0573</td>
<td>0.0089</td>
<td>0.0059</td>
<td>0.0721</td>
</tr>
<tr>
<td></td>
<td>K = 5; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0411</td>
<td>0.0129</td>
<td>0.0082</td>
<td>0.0622</td>
<td>0.0584</td>
<td>0.0101</td>
<td>0.0050</td>
<td>0.0735</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0320</td>
<td>0.0098</td>
<td>0.0350</td>
<td>0.0769</td>
<td>0.0196</td>
<td>0.0098</td>
<td>0.0291</td>
<td>0.0585</td>
</tr>
<tr>
<td></td>
<td>Random Walk</td>
<td>0.0349</td>
<td>0.0103</td>
<td>0.0786</td>
<td>0.1238</td>
<td>0.0499</td>
<td>0.0100</td>
<td>0.0601</td>
<td>0.1201</td>
</tr>
<tr>
<td></td>
<td>K = 3; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0462</td>
<td>0.0176</td>
<td>0.0161</td>
<td>0.0763</td>
<td>0.0386</td>
<td>0.0176</td>
<td>0.0161</td>
<td>0.0724</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0328</td>
<td>0.0117</td>
<td>0.0118</td>
<td>0.0563</td>
<td>0.0351</td>
<td>0.0167</td>
<td>0.0142</td>
<td>0.0660</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0270</td>
<td>0.0095</td>
<td>0.0143</td>
<td>0.0508</td>
<td>0.0349</td>
<td>0.0092</td>
<td>0.0155</td>
<td>0.0596</td>
</tr>
<tr>
<td></td>
<td>K = 5; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0280</td>
<td>0.0147</td>
<td>0.0126</td>
<td>0.0553</td>
<td>0.0343</td>
<td>0.0119</td>
<td>0.0144</td>
<td>0.0606</td>
</tr>
<tr>
<td>The BOE</td>
<td>Diebold-Li</td>
<td>0.0140</td>
<td>0.0141</td>
<td>0.0073</td>
<td>0.0353</td>
<td>0.0134</td>
<td>0.0138</td>
<td>0.0059</td>
<td>0.0330</td>
</tr>
<tr>
<td></td>
<td>Random Walk</td>
<td>0.0368</td>
<td>0.0270</td>
<td>0.0153</td>
<td>0.0790</td>
<td>0.0368</td>
<td>0.0271</td>
<td>0.0153</td>
<td>0.0792</td>
</tr>
<tr>
<td></td>
<td>K = 3; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0191</td>
<td>0.0179</td>
<td>0.0079</td>
<td>0.0456</td>
<td>0.0191</td>
<td>0.0179</td>
<td>0.0075</td>
<td>0.0444</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0187</td>
<td>0.0141</td>
<td>0.0076</td>
<td>0.0404</td>
<td>0.0187</td>
<td>0.0138</td>
<td>0.0074</td>
<td>0.0398</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0186</td>
<td>0.0139</td>
<td>0.0067</td>
<td>0.0391</td>
<td>0.0185</td>
<td>0.0137</td>
<td>0.0067</td>
<td>0.0389</td>
</tr>
<tr>
<td></td>
<td>K = 5; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0185</td>
<td>0.0138</td>
<td>0.0063</td>
<td>0.0386</td>
<td>0.0185</td>
<td>0.0135</td>
<td>0.0063</td>
<td>0.0383</td>
</tr>
<tr>
<td>The ECB</td>
<td>Diebold-Li</td>
<td>0.0261</td>
<td>0.0094</td>
<td>0.0082</td>
<td>0.0436</td>
<td>0.0261</td>
<td>0.0095</td>
<td>0.0080</td>
<td>0.0436</td>
</tr>
<tr>
<td></td>
<td>Random Walk</td>
<td>0.0221</td>
<td>0.0079</td>
<td>0.0280</td>
<td>0.0580</td>
<td>0.0219</td>
<td>0.0079</td>
<td>0.0280</td>
<td>0.0578</td>
</tr>
<tr>
<td></td>
<td>K = 3; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0232</td>
<td>0.0087</td>
<td>0.0088</td>
<td>0.0407</td>
<td>0.0229</td>
<td>0.0087</td>
<td>0.0089</td>
<td>0.0406</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0143</td>
<td>0.0079</td>
<td>0.0085</td>
<td>0.0306</td>
<td>0.0141</td>
<td>0.0079</td>
<td>0.0085</td>
<td>0.0305</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0204</td>
<td>0.0074</td>
<td>0.0087</td>
<td>0.0365</td>
<td>0.0201</td>
<td>0.0073</td>
<td>0.0087</td>
<td>0.0361</td>
</tr>
<tr>
<td></td>
<td>K = 5; $f_k^{(c)} = f_k \forall k$</td>
<td>0.0135</td>
<td>0.0067</td>
<td>0.0088</td>
<td>0.0290</td>
<td>0.0134</td>
<td>0.0067</td>
<td>0.0088</td>
<td>0.0289</td>
</tr>
<tr>
<td>The BOC</td>
<td>Diebold-Li</td>
<td>0.1217</td>
<td>0.0477</td>
<td>0.0692</td>
<td>0.2388</td>
<td>0.1087</td>
<td>0.0496</td>
<td>0.0673</td>
<td>0.2255</td>
</tr>
<tr>
<td></td>
<td>Random Walk</td>
<td>0.1863</td>
<td>0.0566</td>
<td>0.1410</td>
<td>0.3840</td>
<td>0.2210</td>
<td>0.0521</td>
<td>0.1125</td>
<td>0.3857</td>
</tr>
<tr>
<td></td>
<td>K = 3; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.1258</td>
<td>0.0577</td>
<td>0.0405</td>
<td>0.2240</td>
<td>0.1408</td>
<td>0.0565</td>
<td>0.0392</td>
<td>0.2365</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} \neq f_k \forall k$</td>
<td>0.0985</td>
<td>0.0469</td>
<td>0.0339</td>
<td>0.1793</td>
<td>0.1211</td>
<td>0.0467</td>
<td>0.0345</td>
<td>0.2023</td>
</tr>
<tr>
<td></td>
<td>K = 4; $f_k^{(c)} = f_k \forall k$</td>
<td>0.1083</td>
<td>0.0424</td>
<td>0.0389</td>
<td>0.1897</td>
<td>0.1308</td>
<td>0.0392</td>
<td>0.0367</td>
<td>0.2067</td>
</tr>
<tr>
<td></td>
<td>K = 5; $f_k^{(c)} = f_k \forall k$</td>
<td>0.1011</td>
<td>0.0481</td>
<td>0.0359</td>
<td>0.1851</td>
<td>0.1246</td>
<td>0.0422</td>
<td>0.0345</td>
<td>0.2014</td>
</tr>
</tbody>
</table>

Table 1: One-step (weekly) forecast ISEs and MSEs over the periods 1/1/2008-1/1/2009, 6/1/2010-6/1/2011, and 1/1/2013-1/1/2014. We compare the DL and RW forecasts to four variations of the FDLM: $K \in \{3, 4\}$ with separate factor loading curves for each economy ($f_k^{(c)} \neq f_k$) and $K \in \{4, 5\}$ with common factor loading curves ($f_k^{(c)} = f_k$). For each FDLM, we use independent ARIMA(1,1,0) models for the $\beta_{k,i}$. The smallest ISEs and MSEs for each economy (summing over the three periods) are denoted by italics; the smallest overall (summing over economies) ISE and MSE are in bold italics.
short-term boundary.

Figure 2: Factor loading curves $f_{k}^{(c)} = f_{k}$ for FDLM with $K = 5$. The gray lines are random samples from the posterior distribution, the black lines are the posterior means, and the dashed red lines are the Nelson-Siegel curves with fixed nonlinear parameter.

It is important that the dominant hump-like features of these curves occur at different maturities—following from the orthonormality constraint—because it allows the model to fit a variety of yield curve shapes. In particular, Figure 3 exhibits this flexibility: even during the abnormal conditions following the Lehman Brothers bankruptcy, our $K = 5$ common factor loading curves are sufficient for estimating the very different yield curve shapes that were observed across economies.

In Figure 4, we plot the factors $\beta_{k,t}^{(c)}$. There is substantial movement for each $\beta_{k,t}^{(c)}$ following Lehman Brothers bankruptcy, and more subtle movement following S&P downgrade of U.S. credit. Most noticeably, the trends are very similar across economies for each $k$, which is
Figure 3: Yield curve estimates for each economy on 9/19/2008, four days after the Lehman Brothers filed for bankruptcy. The black dots are the observed yield curve values, the dashed red line is the DL fit, the solid blue line is the posterior expectation of $\hat{Y}(c)$ from the FDLM (3) with $K = 5$, and the gray lines are samples from the posterior distribution of $\hat{Y}(c)$. Helpful for broader interpretation of the common factor loading curves $f_k$. Interestingly, the $\beta_{k,t}^{(c)}$ may be cointegrated [Granger and Newbold, 1974], in particular for fixed $k$ across economies $c$. This is a promising area for further research, since it may help identify long-run equilibriums among the yield curves for major economies—as well as deviations from these equilibriums during certain periods.
6 Conclusions

The FDLM provides a general framework to model complex dependence structures among functional observations. Because we separate out the functional component through appropriate conditioning, we can model the remaining dependence using familiar scalar and multivariate methods; for the yield curve data, we used a hierarchical DLM for multivariate time series. Then, conditional on the non-functional dependence model, we incorporate Bayesian spline theory, convex optimization, and the theory of exponential families to model the functional component as a set of smooth, nonparametric, and optimal curves subject to constraints. We finally combine these components using a Gibbs sampler, which allows us to sample from the joint posterior distribution and therefore perform inference on any of the model parameters, such as covariate coefficients.

The FDLM’s performance on the multi-economy yield curve data demonstrates the flexibility, accuracy, and interpretability of the model. Our results suggest commonality among the yield curves, despite the different estimation techniques used for each economy’s data. This offers some reconciliation among the competing yield curve estimation techniques and provides insight into between-economy yield curve dynamics. Further study of the factor trends, interactions, and responses to covariates of interest is readily available within the FDLM framework, and may help answer interesting economic and financial questions.
Figure 4: The factors $\beta_{k,t}^{(c)}$ over time $t$ for model $\bar{M}$ with $K = 5$. The magenta line denotes 9/19/2008, four days after the Lehman Brothers filed for bankruptcy; the cyan line denotes 8/5/2011, the date of the S&P downgrade of U.S. credit.
References


7 Appendix

7.1 Gibbs Sampling Algorithm

To sample from the joint posterior distribution of all parameters, we use a Gibbs sampler (Casella and George 1992). Because the Gibbs sampler requires blocks of parameters to be conditioned on all other blocks of parameters, it is a convenient approach for our model. First, dynamic linear model (DLM) algorithms typically require that the factors be the only random components, which we can accommodate by conditioning appropriately. Second, our sequential orthonormality approach for \( f_k \) fits nicely within a Gibbs sampler, and we can use variations of the algorithms described in Wand and Ormerod (2008). And third, the hierarchical structure of our model imposes natural conditional independence assumptions, which allows us to easily divide up the parameters into appropriate blocks.

7.1.1 Initialization

For initial values of the factors \( B_k^{(c)} = (\beta_{k,1}^{(c)}, \ldots, \beta_{k,T}^{(c)})' \), we compute the singular value decomposition (SVD) of the data matrix \( Y^{(c)} = UDV' \) and set \( (B_1^{(c)}, \ldots, B_K^{(c)}) = U_{1:K}D_{1:K} \), where \( U_{1:K} \) are the first \( K \) columns of \( U \) and \( D_{1:K} \) is the upper \( K \times K \) submatrix of \( D \). Note that to obtain a data matrix \( Y^{(c)} \), with rows corresponding to time \( t \) and columns to observations points \( \tau \), we need to estimate \( Y_t^{(c)}(\tau) \) for any unobserved \( \tau \) at each time \( t \), which may be computed quickly using splines. However, these estimated data values are only used for the initialization step. Then, letting \( f_k^{(c)} \) be the vector of factor loadings evaluated at all observation points \( \bigcup_t T_t^{(c)} \) for outcome \( c \), we set \( (f_1^{(c)}, \ldots, f_K^{(c)}) = V_{1:K} \), where \( V_{1:K} \) is the first \( K \) columns of \( V \). The \( f_k^{(c)} \) are orthogonal in the sense that \( f_k^{(c)} f_j^{(c)} = 1(k = j) \), but
they are not smooth. This approach is similar to the initializations in Matteson, McLean, Woodard, and Henderson (2011) and Hays et al. (2012).

Given \( B_k^{(c)} \) and \( f_k^{(c)} \) from the SVD, we can estimate each \( \sigma^2_k^{(c)} \) (or more generally, \( E_t^{(c)} \)) as a conditional MLE, using the likelihood from the data level of model (1). Similarly, we can estimate each \( \lambda_k^{(c)} \) conditional on \( f_k^{(c)} \) by maximizing the partially informative normal likelihood. Then, given \( \lambda_k^{(c)}, \sigma_k^{(c)}, B_k^{(c)}, \) and \( f_k^{(c)} \), we can estimate each \( d_k^{(c)} \) by normalizing the full conditional posterior expectation given in the main paper; i.e., solving the relevant quadratic program and then normalizing the solution. The lower level initializations proceed similarly as conditional MLEs, but depend on the form chosen for \( \theta_t, G_t, V_t, \) and \( W_t \). In our application, this conditional MLE approach produces very reasonable starting values for all variables.

7.1.2 Sampling

Our algorithm proceeds in three main blocks:

1. Sample the basis coefficients \( d_k^{(c)} \) and the smoothing parameters \( \lambda_k^{(c)} \) for the factor loading curves. For \( \lambda_k^{(c)} \), we use a Gamma(\( \gamma_1, \gamma_2 \)) prior distribution, which is conjugate to the partially informative normal likelihood. The full conditional posterior distribution is Gamma(\( \gamma_1 + \text{rank}(\Omega_{\phi})/2, \gamma_2 + d_k^{(c)'} \Omega_{\phi} d_k^{(c)}/2 \)). We use the hyperparameters \( \gamma_1 = \gamma_2 = 0.001 \), although the effect of the hyperparameters is negligible as long as \( \gamma_1 \) and \( \gamma_2 \) are small relative to \( \text{rank}(\Omega_{\phi})/2 \) and \( d_k^{(c)'} \Omega_{\phi} d_k^{(c)}/2 \), respectively. After sampling the \( \lambda_k^{(c)} \), we sample and then normalize the \( d_k^{(c)} \) with a modified version of the efficient Cholesky decomposition approach of Wand and Ormerod (2008):

(a) Compute the (lower triangular) Cholesky decomposition \( V_k^{-1} = \tilde{V}_L \tilde{V}_L' \);

(b) If \( k = 1 \), set \( L_{1:(k-1)}'A_{1:(k-1)} = 0 \);

If \( k > 1 \), use forward substitutions to obtain \( \tilde{x} \) and \( \tilde{y} \) from the equations \( \tilde{V}_L \tilde{x} = L_{1:(k-1)}' \) and \( \tilde{V}_L \tilde{y} = u_k \), and let \( A_{1:(k-1)} \) be the solution to the regression of \( \tilde{y} \) on \( \tilde{x} \);
(c) Use forward substitution to obtain $\bar{u}$ as the solution to $V_L\bar{u} = u_k$, then use backward substitution to obtain $d_0$ as the solution to $L'\bar{u} = \bar{u} + \bar{z}$, where $\bar{z} \sim N(0, I_{M+4})$;

(d) Retain the vector $d^*_k = d_0 / \sqrt{d'^0 J \phi d_0}$.

The definitions of $V_k$ and $u_k$ depend on whether or not we use the common factor loading curve model with $f^{(c)}_k = f_k$; see Section 4. The algorithm efficiently produces $A_{1:(k-1)} = (L'_{1:(k-1)} V_k L_{1:(k-1)})^{-1} L'_{1:(k-1)} V_k u_k$, as desired. Moreover, the extra orthogonality step (b) utilizes the Cholesky decomposition—which we must compute regardless—and adds only the computational cost of a simple linear regression for each $k > 1$. Note that since we retain the unit-norm vectors $d^*_k$ at every MCMC iteration, the $\lambda_k$ are therefore conditional on the normalized $d_k$ in the Gibbs Sampler, and no other unknowns. Hence, adjusting the $\lambda_k$ for the normalization of $d_k$ would have no effect on the full conditional posterior distribution from which we sample the $\lambda_k$.

2. Sample the factors $\beta_t$ (and $\theta_t$, if present) conditional on all other parameters in (I) using the DLM implementation of forward filtering backward sampling (e.g., Petris et al. 2009) or the hierarchical version in Gamerman and Migon (1993). For the prior distributions, we only need to specify the distribution of $\beta_0$ (and $\theta_0$); the remaining distributions are computed recursively using $F(\tau)$, $A_t$, and $G_t$. For simplicity, we let $\beta^{(c)}_{k,0} \sim N(0, 10^6)$, which is a common choice for DLMs. Alternatively, we could use past data not included in our analysis to estimate these initial values. However, the resulting estimates for $t > 1$ in our application are not noticeably different.

3. Sample each of the remaining error variance parameters individually: $E_t$, $V_t$, and $W_t$. As previously mentioned, these distributions depend on our assumptions for the model structure. Typically, we favor conjugate priors and, for the stochastic volatility model, the distributions given in Kim et al. (1998). In particular, letting $\sigma_{k,c,t}^2 = \exp(h^{(c)}_{k,t})$, Kim et al. (1998) propose the model $h^{(c)}_{k,t} = \xi^{(c)}_{k,0} + \xi^{(c)}_{k,1} (h^{(c)}_{k,t-1} - \xi^{(c)}_{k,0}) + \gamma^{(c)}_{k,t}$, where $\gamma^{(c)}_{k,t} \sim$
\( N(0, \sigma^2_{H,k,(c)}) \) for \( t = 2, \ldots, T \) and \( h_{k,1}^{(c)} \sim N(\xi_{k,0}^{(c)}, \sigma^2_{H,k,(c)}/(1 - (\xi_{k,1}^{(c)})^2)) \) with \( |\xi_{k,1}^{(c)}| < 1 \) for stationarity. Kim et al. (1998) also suggest priors for \( \xi_{k,0}^{(c)}, \xi_{k,1}^{(c)} \), and \( \sigma^2_{H,k,(c)} \) and provide an efficient MCMC sampling algorithm. For additional motivation for the stochastic volatility approach over GARCH models, see Danielsson (1998).

### 7.2 Forecasting

Recall that we estimate one-step forecasts for the 52 weeks following each of the start dates 1/1/2008, 6/1/2010, and 1/1/2013. For example, if \( T^* \) corresponds to 1/1/2008, we forecast \( Y_{T^* + 1}^{(c)} \) given \( Y_{1:T^*}^{(c)} \), then forecast \( Y_{T^* + 2}^{(c)} \) given \( Y_{1:(T^* + 1)}^{(c)} \) and so on, up to the date \( T^* + 52 \), corresponding to 1/1/2009.

Forecasting the FDLM is similar to standard DLM forecasting, although since \( \mathbf{F}, \mathbf{G}_t \), and the error variances are unknown, we no longer have a closed form for the posterior predictive distribution \( [Y_{T^* + 1}^{(c)} | Y_{1:T^*}^{(c)}] \). However, we may approximate the posterior predictive distribution by sampling from \( [Y_{T^* + 1}^{(c)} | Y_{1:T^*}, \mathbf{\beta}_{T^* + 1}, \mathbf{F}, \sigma^2_{(c)}] \sim N(\mathbf{F} \mathbf{\beta}_{T^* + 1}, \sigma^2_{(c)}) \) within the Gibbs sampler. Here, we only consider point forecasts of \( Y_{T^* + 1}^{(c)} \), which simply requires computation of the expectation \( \mathbf{F} \mathbf{\beta}_{T^* + 1} \). Therefore, for \( \ell = 1, \ldots, 52 \), we estimate \( Y_{T^* + \ell}^{(c)} \) via \( \mathbf{F} \mathbf{\beta}_{T^* + \ell} | Y_{1:(T^* + \ell - 1)}^{(c)} \), then estimate \( Y_{T^* + \ell + 1}^{(c)} \) via \( \mathbf{F} \mathbf{\beta}_{T^* + \ell + 1} | Y_{1:(T^* + \ell)}^{(c)} \), and so on, for each \( T^* \). Hence, to obtain exact FDLM forecasts, we must update the entire FDLM—\( \mathbf{\beta}_{T^* + \ell}, \mathbf{F}, \mathbf{G}_{T^* + \ell} \), and the error variances, which may be time-dependent—based on the new observation \( Y_{T^* + \ell}^{(c)} \).

In general, there are no convenient updating algorithms for an entire DLM. However, there are recursive updating algorithms for \( \mathbf{\beta}_{T^* + \ell} \) (e.g., Petris et al. 2009). Therefore, we may obtain an approximate point forecast of \( Y_{T^* + \ell}^{(c)} \) by conditioning \( \mathbf{\beta}_{T^* + \ell} \) on all \( Y_{1:(T^* + \ell - 1)}^{(c)} \), but conditioning the other FDLM parameters only on \( Y_{1:T^*}^{(c)} \). Essentially, this assumes that \( \mathbf{F}, \mathbf{G}_t \), and the error variances can be adequately estimated from \( Y_{1:T^*}^{(c)} \), and would not change substantially given \( Y_{T^* + \ell}^{(c)}, \ell = 1, \ldots, 52 \). As a result, our forecast errors are likely overestimated for the FDLM: in particular, we observe that forecast errors for \( Y_{T^* + \ell}^{(c)} \) increase with \( \ell \) for all outcomes (economies). Of course, we could run a Gibbs sampler for each \( T^* + \ell \) to
obtain a forecast estimate (and distribution) of $Y_{T^*+\ell}^{(c)}$, but that would require $52 \times 3 = 156$ Gibbs samplers. Note that in practice, this is typically not an issue: we need a single Gibbs sampler for each starting date, of which there would usually only be one.