UNIVERSITY OF CALIFORNIA,
IRVINE

Essays in Flexible and Nonlinear Time Series Econometrics

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Economics

by

Robert MacDonald

Dissertation Committee:
Associate Professor Ivan Jeliazkov, Chair
Professor Fabio Milani
Professor Eric Swanson

2024

DEDICATION

This work is dedicated to my family, without whom it would not be possible. Danielle was always there to offer love and encouragement, for which I will be eternally grateful. My parents fostered in me a passion for learning and inquiry. All that I have to offer as both a researcher and a person is due to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to thank my advisor, Ivan Jeliazkov for his invaluable support and guidance. His excellent teaching seduced me down the path of econometrics. He shepherded me along the way, always ready to offer insightful feedback and interesting directions for further research. I could not have asked for a better mentor, collaborator, or friend.

I would also like to thank Fabio Milani, Eric Swanson, and other professors from the UC Irvine Economics and Statistics faculty for their outstanding instruction and counsel. This dissertation would not be possible without the knowledge and skills they imparted.

Chapter 3 is a version of a joint project with Alexander Luttmann and Ivan Jeliazkov.

Work that contributed to chapters 1 and 2 was funded by fellowships from the University of California, Irvine.

VITA

Robert MacDonald

EDUCATION

Doctor of Philosophy in Economics **2024**
University of California, Irvine *Irvine, California*

Bachelor of Arts in **2018**
International and Development Economics
University of San Francisco *San Francisco, California*

RESEARCH EXPERIENCE

Graduate Student Researcher **2022**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2018–2023**
University of California, Irvine *Irvine, California*

ABSTRACT OF THE DISSERTATION

Essays in Flexible and Nonlinear Time Series Econometrics

By

Robert MacDonald

Doctor of Philosophy in Economics

University of California, Irvine, 2024

Associate Professor Ivan Jeliazkov, Chair

Time series econometrics is essential to empirical studies in macroeconomics, finance, and many other areas. While the canonical models in this literature are small, parametric, and linear, there is growing interest in models in which the data generating process is either nonlinear or of a flexible, adaptive form. This dissertation proposes several new models and techniques in the area of flexible and nonlinear time series econometrics.

Chapter 1 proposes a novel methodology for determining the specification of factor-augmented vector autoregression (FAVAR) models. Without strong a priori beliefs about the set of possible models, the complexity of the problem renders traditional model selection techniques infeasible. By contrast, my proposed solution only requires the estimation of a single model. This makes the process easy to scale in both the cross-sectional and time series dimensions. An efficient optimization algorithm for model estimation is developed. Monte Carlo studies show the technique to be highly effective in small samples, even in the presence of a low signal-to-noise ratio and missing data. Applications to large datasets of monthly and quarterly U.S. macroeconomic variables identify observed factors not normally considered in the FAVAR literature. The methodology is then used to analyze the asset-pricing model of Fama and French (1993). I find that their constructed factors for firm size and book-to-market equity ratio are likely observed components, but excess market return is not.

Chapter 2 proposes a regime-switching linear model with time-varying transition probabilities, endogenous switching, and a nonparametric error distribution. The last two qualities are achieved by letting the conditional mean of the normalized observation errors be a potentially nonlinear function of the errors in the state equation. We demonstrate that this specification permits a very flexible marginal distribution for the observation error. A Markov Chain Monte Carlo algorithm for sampling from the posterior distribution of parameters is developed. A simulation study demonstrates that existing parametric switching models yield biased parameter estimates when the data is generated by a model with nonlinear endogenous switching. We apply the model to US quarterly output growth. The proposed model is shown to fit the data better than parametric switching models.

Count data models are at the core of a large and diverse empirical literature in the social and natural sciences. A key component in this class of models is the mean function, which defines the relationship between the covariates and the conditional expectation of the count process. Chapter 3 considers a general approach for representing the mean function that is adaptable, tractable, and dispenses with problematic facets of count data models such as explosive covariate effects and restrictive time series properties. The methodology is broadly applicable in cross-sectional, longitudinal, and time-series settings, with likelihood-based, generalized linear, copula and other models. We provide theoretical results that distinguish our methodology from existing work and implement it in two examples that demonstrate its relevance and practical appeal.

Chapter 1

Specification of FAVAR Models

Factor-augmented vector autoregressions (FAVARs) are a popular tool in high-dimensional series analysis. The central assumption of any factor model is that much of the variation in a large panel of dependent variables can be explained by a relatively small number of common components. A standard FAVAR with an intercept is written as

$$X_t = \mu_X + \Lambda^f f_t + \Lambda^y y_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Sigma), \tag{1.1}$$

$$\begin{bmatrix} f_t \\ y_t - \mu_y \end{bmatrix} = \Phi(L) \begin{bmatrix} f_{t-1} \\ y_{t-1} - \mu_y \end{bmatrix} + \eta_t, \quad \eta_t \sim N(0, \Omega). \tag{1.2}$$

X_t is an $N \times 1$ vector of dependent variables with unit variance, f_t is an $r_f \times 1$ vector of latent factors, y_t is an $r_y \times 1$ vector of observed factors, Λ^f and Λ^y are matrices of loading parameters, μ_X is the intercepts, and ε_t is an $N \times 1$ vector of error terms. The variables are observed in each time period $t = 1, \ldots, T$. The common factors f_t and y_t are assumed to explain all of the covariance in X_t. The idiosyncratic errors ε_t are thus assumed to have diagonal covariance matrix $\Sigma = diag(\sigma_1^2, \ldots, \sigma_N^2)$. The FAVAR reduces to a multivariate regression when $r_f = 0$ and a dynamic factor model (DFM) when $r_y = 0$. This specification

allows for parsimonious modeling of high-dimensional data when $r = r_f + r_y << N$, thus offering an alternative to highly-parameterized vector autoregressions (VARs).

The FAVAR was originally developed for structural macroeconomic analysis by Bernanke, Boivin, and Eliasz (2005).[1] The authors assumed the Federal Funds Rate was the only observed factor and did not perform any model comparison with alternative observed factors. Models with observed factors, though not typically cast in terms of an FAVAR, are also common in the asset-pricing literature[2]. The number of possible observed factors to consider has grown quite large. Choosing the best subset from the available "factor zoo" (Cochrane, 2011) is of interest to researchers and investors alike.

There is no existing feasible method for comparing all of the possible FAVAR specifications. Model selection requires knowing the observed factors, the number of latent factors, and the lag order. Exhaustive model comparison would require estimating millions of models, even with modestly sized datasets. I propose a solution that only requires the estimation of a single model. The procedure exploits the fact that any FAVAR has an equivalent representation as a DFM. I first determine the total number of factors r and the lag order p using existing information criterion methods. The main contribution of this chapter is the identification of observed factors. If an observed factor is modeled as a dependent variable in a DFM, then the associated idiosyncratic error term ε_{it} will have true variance $\sigma_i^2 = 0$. I estimate a DFM with all observed variables and place a Bayesian variable selection prior on each σ_i^2. The prior is designed to shrink small variances towards 0 while exercising little influence on larger variances.

Model selection is achieved through maximum a posteriori (MAP) estimation. To facilitate fast model selection, I develop several extensions to the Expectation-Maximization (EM) algorithm for DFMs. The proposed algorithm leverages the rotation and scale invariance of

1. See Belviso and Milani (2006); Boivin, Giannoni, and Stevanović (2013); Fernald, Spiegel, and Swanson (2014); Paccagnini (2017) among others for further detail as well as interesting extensions and applications.
2. Among many others, see Chen, Roll, and Ross, 1986; Fama and French, 1993; Fama and French, 2015.

the likelihood to obtain a solution significantly faster than the basic EM algorithm.

The identification of observed factors has been addressed solely by the frequentist literature until this point. The two papers most closely related to this project are Bai and Ng (2006) and Parker and Sul (2016). Bai and Ng (2006) observe that if we can consistently estimate the factor space, then any observed factors will be linear combinations of the estimated factors. Their procedure relies on statistical tests in which some candidate variable is an observed factor under the null hypothesis. This approach is reasonable when the set of possible observed factors is small, but will encounter problems when the set is large. To systematically find the correct observed factors in a dataset with many variables, this requires running dozens or hundreds of independent tests and then performing a correction for multiple hypothesis testing. This method is unlikely to select the true model and may produce incoherent results, such as concluding there are more observed factors than total possible factors. Parker and Sul (2016) build upon the work of Bai and Ng (2006) to develop a criterion for finding a set of candidate observed factors. When combined with a clustering algorithm, the criterion is effective at finding the set of all possible observed factors. However, this approach is agnostic about choosing between highly correlated candidates. If X_i is an observed factor and $X_j = X_i + \varepsilon_j$, where ε_j has a small variance, the criterion may conclude that either variable could be an observed factor. Both papers assume a balanced panel dataset and do not address the case of missing data. My model selection process does not encounter the same problems.

I apply the procedure to large datasets of monthly and quarterly macroeconomic data, as well as the asset-pricing data of Fama and French (1993). The analysis of quarterly macroeconomic data yields surprising results. Rather than selecting the Federal Funds Rate as an observed factor, the default assumption in the monetary FAVAR literature, the procedure selects the Total Capacity Utilization index. A model of monthly macroeconomic data selects the spread between the 10-Year Treasury Rate and the Federal Funds Rate as the only

observed factor. Models restricted to the period following the 2007 Financial Crisis find that employment measures are more likely to be observed factors. I also estimate a model with the same variables as Fama and French (1993). Variables that measure the excess returns attributable to firm size and book-to-market equity ratio are classified as observed factors, while the excess return from a market portfolio is not.

The remainder of the chapter proceeds as follows. Section 1.1 recasts the model selection process as an optimization problem. Section 1.2 develops an efficient EM algorithm for MAP estimation. Section 1.3 investigates the performance of the proposed procedure through Monte Carlo studies. Section 1.4 applies the new approach to large macroeconomic and financial datasets, and section 1.5 concludes. Mathematical proofs and technical details can be found in the appendix.

1.1 The Model Selection Procedure

1.1.1 Rotation Invariant Likelihood

A perennial issue in factor analysis is that the likelihood is invariant under rotations of the factor basis. Consider the case of a DFM with likelihood $f(X|\theta)$. For any square invertible matrix A, $\Lambda f_t = \Lambda A^{-1} A f_t$. Let $F = (f_1, \ldots, f_T)$, $\Lambda^* = \Lambda A^{-1}$, $F^* = AF$, $\Phi_l^* = A\Phi_l A^{-1}$, and $\Omega^* = A\Omega A'$. Assuming the first p instances of the factors come from the stationary distribution, where p is the VAR lag order in (2), we obtain the equality

$$\begin{aligned} f(X|\theta) &= \int f(X|F, \Lambda, \Sigma)\pi(F|\Phi, \Omega)dF \\ &= \int f(X|F^*, \Lambda^*, \Sigma)\pi(F^*|\Phi^*, \Omega^*)dF^* \\ &= f(X|\theta^*). \end{aligned} \tag{1.3}$$

This means that parameter restrictions are required to identify the likelihood. Unfortunately, a priori restrictions can lead to model misspecification. Identification is usually achieved through restrictions on an $r \times r$ submatrix of Λ. However, the restrictions are only valid if the true submatrix is invertible.

1.1.2 Rewriting the FAVAR as a DFM

If we stack X_t and y_t in a single vector, we can then rewrite the FAVAR as a special case of a DFM:

$$\begin{bmatrix} X_t \\ y_t \end{bmatrix} = \begin{bmatrix} \mu_X + \Lambda^y \mu_y \\ \mu_y \end{bmatrix} + \begin{bmatrix} \Lambda_f & \Lambda_y \\ 0_{r_y \times r_f} & I_{r_y} \end{bmatrix} \begin{bmatrix} f_t \\ y_t - \mu_y \end{bmatrix} + \varepsilon_t^\dagger, \qquad \varepsilon_t^\dagger \sim N\left(0, \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\right). \tag{1.4}$$

Now consider rotating the factors by an arbitrary invertible matrix A: $f_t^* = A[f_t' \;\; y_t' - \mu_y']'$. The FAVAR can then be written as:

$$X_t^* = \mu_{X^*} + \Lambda^* f_t^* + \varepsilon_t^*, \qquad \varepsilon_t^* \sim N\left(0, \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}\right), \tag{1.5}$$

$$f_t^* = \Phi^*(L) f_{t-1}^* + \eta_t^*, \quad \eta_t^* \sim N(0, \Omega^*), \tag{1.6}$$

where $X_t^* = [X_t' \;\; y_t']'$, μ_{X^*} is the intercept from (1.4), and the remaining parameters with asterisks are defined as in section 2.1. We can thus conclude that any DFM in which some idiosyncratic variances are 0 is equivalent to an FAVAR where the corresponding variables are observed factors.

Rather than comparing estimates from different FAVAR specifications. I will estimate a DFM that nests all possible FAVARs with total number of factors r and lag order p. Identification is not an issue if your only aim is to determine the observed factors, r, and p. The elements of Σ do not change when the factor basis is rotated. Since the procedure I propose makes

use of a MAP estimate from a Gaussian state-space model, the posterior can be optimized using an EM algorithm, which does not require an identified likelihood to converge to a maximum. This helps us avoid any model misspecification problems that can arise from a priori restrictions. The only normalization I assume is $\Omega = I_r$. This helps to scale identify the factors and facilitates jumping between points of equal probability to accelerate the EM algorithm. Once MAP estimates are obtained, the researcher is free to choose his or her preferred identifying restrictions and rotate the factor basis accordingly.

1.1.3 Determining the Observed Factors

Let us consider the problem of identifying the observed factors when r and p are known. The theoretically ideal method would be exhaustive Bayes factor comparisons. However, the combinatorial complexity of such a procedure requires prohibitively large computing resources even when r is small. One would have to estimate marginal likelihoods for $\sum_{r_y=0}^{r} \binom{N}{r_y}$ models. This amounts to over 4 million marginal likelihood estimations in the modest case of $N = 100$ and $r = 4$. Even if one used the Bayesian information criterion (BIC) to approximate the marginal likelihood, the model selection process would take months on a standard personal computer.

Since any variables with idiosyncratic variances of 0 must be observed factors, a closely related approach would be to place spike-and-slab priors on the variances. This would take the form

$$\pi(\sigma_i^2) = (1 - \rho_i)\delta_0(\sigma_i^2) + \rho_i \psi_1(\sigma_i^2). \tag{1.7}$$

While the spike-and-slab prior recasts the problem in the context of a single model, it does not alleviate the problem of combinatorial complexity. To produce a posterior that is easier

to traverse, let us consider a continuous approximation of the point-mass mixture prior. After adding a hierarchical prior on the mixing weight and a latent indicator for the components of the mixture, the prior for σ_i^2 is expressed as

$$\pi(\sigma_i^2|\gamma_i) = \psi_0(\sigma_i^2)^{1-\gamma_i}\psi_1(\sigma_i^2)^{\gamma_i}, \tag{1.8}$$

$$\psi_q(\sigma_i^2) = \alpha_q e^{-\alpha_q\sigma_i^2}, \tag{1.9}$$

$$\gamma_i \sim \text{Bernoulli}(\rho_i), \tag{1.10}$$

$$\rho_i \sim \mathcal{B}(a, a). \tag{1.11}$$

The spike-and-slab densities are exponential distributions. By setting $\alpha_0 >> \alpha_1$ and α_1 close to 0, we can place virtually all of the probability mass of the spike distribution (ψ_0) near 0 while placing nearly all of the mass of the slab distribution (ψ_1) away from 0. Figure 1.1 shows the mixture density for increasingly large α_0's and $\rho = 0.5$. This is equivalent to the marginal prior after γ_i and ρ_i have been integrated out. We can see that this continuous prior approaches the point-mass mixture prior as $\alpha_0 \to \infty$. I employ parameter expansion by augmenting the prior with the latent indicator variable γ_i. The latent variable formulation is amenable to closed-form updates of variance estimates within an EM algorithm.

Parameter estimates are obtained by solving the optimization problem

$$\begin{aligned}\hat{\theta}_{MAP} &= argmax_\theta \pi(\theta|X) \\ &= argmax_\theta f(X|\theta)\pi(\theta) \\ &= argmax_\theta \ln f(X|\theta) + \ln\pi(\theta).\end{aligned} \tag{1.12}$$

Care must be taken when optimizing a model for which some $\hat{\sigma}^2_{i,MAP} = 0$. The EM algorithm requires the output from a Kalman smoother, which gives imprecise estimates when idiosyncratic variances are sufficiently small. I avoid this problem by first estimating a constrained model in which $\sigma_i^2 \geq 10^{-15}$. We can maintain numerical stability with variances of this size by using a square root Kalman smoother that leverages the QR decomposition

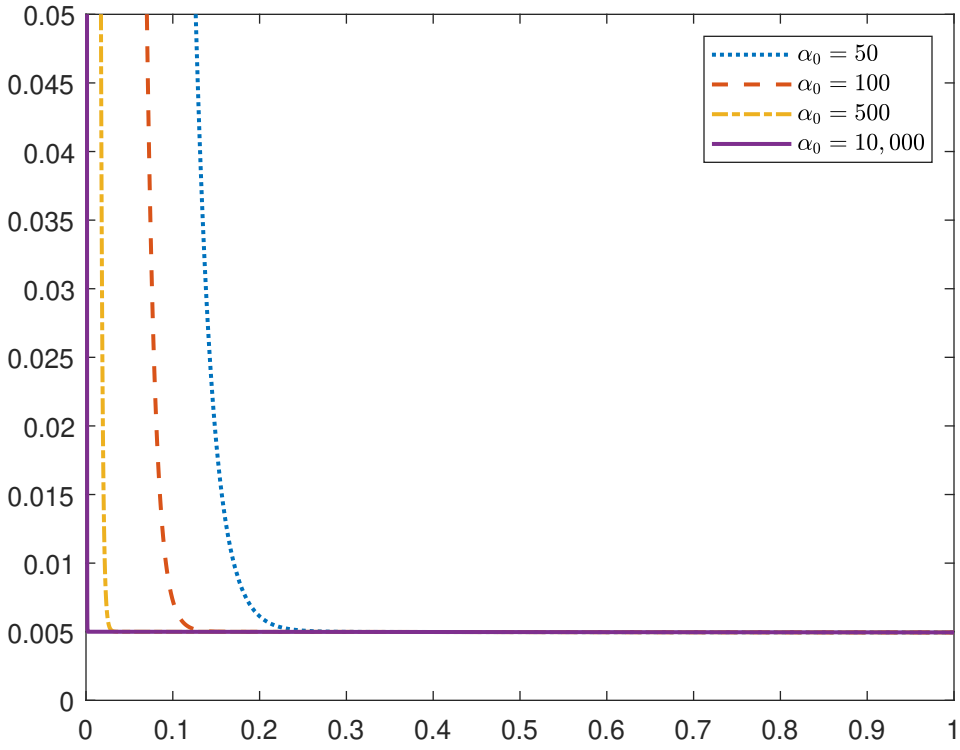


Figure 1.1: Spike-and-Slab Priors for σ_i^2, $\alpha_1 = 0.01$, $\rho = 0.5$

(Tracy, 2022). After termination of the EM algorithm, I check to see if the posterior density can be increased further by setting any $\hat{\sigma}^2_{i,MAP} < 10^{-8}$ to 0. The model with exacts 0's was preferred in all estimations.

1.1.4 Selecting the Number of Factors

Before applying the model selection process, we must first know the number of factors. There has been a great deal of work done with regard to estimating r. Many approaches in the frequentist literature develop information criteria (Bai and Ng, 2002; Hallin and Liška, 2007; Ahn and Horenstein, 2013). Another approach to estimating r is using an overidentified model, with more factors than is likely true, and then forcing factor loadings towards 0. Frequentist methods accomplish this by applying regularization techniques like the LASSO to

factors estimated using Principle Components Analysis (PCA) (Zou, Hastie and Tibshirani, 2006; Witten, Tibshirani and Hastie, 2009). Bayesian solutions typically place hierarchical shrinkage priors on the loading parameters (Carvalho et al., 2008; Frühwirth-Schnatter and Lopes, 2009; Knowles and Ghahramani, 2011; Ročková and George, 2016). These methods pertain to models in which the factors are assumed to be uncorrelated across time. There has been some recent work that extends the approach to restricted DFMs (McAlinn, Ročková, and Saha 2018; Luo and Yu, 2021).

Many Bayesian approaches for selecting r are unfortunately ill-suited to the problem at hand. Methods based on variable selection priors are less effective when the idiosyncratic variances are very small, which will occur for any observed factors as well as any other variables that are particularly well-explained by the common components. When continuous shrinkage priors are used, such as in Ročková and George (2016), very small variances reduce the variable selection penalty to effectively 0. Another issue arises when the factors are highly correlated, a situation that is not precluded by the model under consideration. In fact, highly correlated factors are likely to result when the model is overidentified. Overidentification does not create the same issue in static factor models because the factors are independent a priori. Point mass-density priors may be less susceptible to the problem of small variances, but they are still likely to encounter difficulties with highly correlated factors. Likelihood-based criteria such as Bayes factors and the Deviance Information Criterion unfortunately have a tendency to overfit the number of factors (Beyeler and Kaufmann, 2021). While BIC performed well in simulation studies, it exhibited the same overfitting property in empirical applications, continuing to decrease as the number of factors increased. It is for this reason I instead use the IC_{p2} criterion of Bai and Ng (2002). It can be computed quickly, has very good finite sample properties and tends to give reasonable factor estimates for models calibrated to macroeconomic applications (Stock and Watson, 2016).

1.1.5 Lag Length Selection

Conditional on the number of factors, I use BIC to select p. Calculation of the BIC is not obvious in this case because I have chosen to maximize the unidentified likelihood, which attains a maximum not at a single point but at a ridge. The derivation of the BIC requires the maximum to be unique for the Laplace approximation to be valid, meaning BIC can only be calculated for an identified model. However, the maximum likelihood value found for the unidentified likelihood will correspond to the maximum likelihood of any model with correct identifying restrictions (assuming there is at least one nonsingular $r \times r$ submatrix in $\hat{\Lambda}_{MLE}$). We can thus use the maximum likelihood from the unidentified model and the penalty from the identified model to compute

$$\text{BIC} = -2\ln f(X|\hat{\theta}_{MLE}) + \ln T(N(r+2) + pr^2 - r(r-1)/2). \tag{1.13}$$

1.2 Estimation Algorithm

Our goal is to maximize $\pi(\theta|X) \propto \int f(X|F,\theta)\pi(F|\theta)\pi(\theta)dF$. Maximization of the posterior is achieved using a variant of the EM algorithm. While the posterior can be maximized with a conventional EM algorithm, convergence is considerably slower. I instead use a combination of two EM variants with faster convergence properties: the Expectation/Conditional Maximization (ECME) algorithm (Liu and Rubin, 1994) and the Parameter-Expanded Expectation-Maximization (PX-EM) algorithm (Liu, Rubin, and Wu, 1998). The basic EM algorithm is an iterative process in which one can find parameter updates that monotonically increase the value of an integrated density, such as $\pi(\theta|X)$, by maximizing the posterior expectation of the full data density (Dempster, Laird, and Rubin, 1977). The parameter updates take the form

$$\theta_n = argmax_\theta \mathbb{E}[\ln f(X|F,\theta) + \ln\pi(F|\theta)|X,\theta_{n-1}] + \ln\pi(\theta) = argmax_\theta Q(\theta|\theta_{n-1}). \quad (1.14)$$

1.2.1 The PX-EM Algorithm for a DFM

The PX-EM algorithm exploits the rotational invariance of the likelihood. The proposed algorithm proceeds by first maximizing $Q(\theta|\theta_{n-1})$ with respect to $\theta^* = \{\Lambda^*, \Phi^*, \Omega^*, \Sigma\}$. We then take A_n^L to be the lower triangular Cholesky factor of Ω_n^* and set $\Lambda_n = \Lambda_n^* A_n^L$ and $\Phi_{l,n} = A_n^{L^{-1}} \Phi_{l,n}^* A_n^L$ for each lag l. By adopting improper priors for Λ and Φ, we obtain a posterior that is also rotation invariant. The priors for Λ and Φ are $\pi(\Lambda) \propto 1$, and $\pi(\Phi) \propto \mathbb{1}\{\Phi \in \mathcal{A}\}$, where $\mathcal{A}$ is the region of the parameter space for which the roots of the VAR polynomial lie outside the unit circle. The main advantage of improper priors is that they are rotation invariant, which will make the posterior easier to traverse. Improper priors can create convergence issues in Markov Chain Monte Carlo (MCMC) estimation, but they are not a problem in MAP estimation. The sequence of density ordinates generated by the EM updates will still converge to a stationary point (Wu, 1983). Any solution found by optimization is also a solution under an appropriately diffuse, Uniform prior. Diffuse proper priors are unlikely to impact posterior inference for Σ, but they do create difficulties for optimization. One may be inclined to choose diffuse semiconjugate priors such as $\pi(\Lambda_i) = f_N(\Lambda_i|0, \nu_\Lambda I_r)$ and $\pi(\Phi) \propto \mathbb{1}\{\Phi \in \mathcal{A}\} \prod_{i,l} f_N(\Phi_{il}|0, \nu_\Phi I_r)$. Such priors do little to identify the posterior because they are invariant under orthonormal rotations as well as sign and order permutations. However, they are not invariant under oblique rotations such as A^L, so we can no longer use the PX-EM algorithm. Proper priors merely restrict the modes of the posterior to a smaller ridge while making the parameter space more difficult to explore.

The original EM algorithm for maximum likelihood estimation of a DFM can be found in

Watson and Engle (1983). I adapt their algorithm to account for the variable selection prior on Σ and the estimation of Ω^*. Details of the PX-EM algorithm are given in Algorithm 1. Define $\hat{f}_t \equiv \mathbb{E}[f_t|X,\theta_n]$, $\hat{F} \equiv (\hat{f}_1,\ldots,\hat{f}_T)'$, $\hat{P} \equiv \sum_t \mathbb{E}[(f_t-\hat{f}_t)(f_t-\hat{f}_t)'|X,\theta_n]$, and $\hat{\gamma}_i \equiv \Pr(\gamma_i = 1|X,\theta_n)$. All conditional moments related to the factors are available directly from the output of a Kalman smoother in which the state vector has been augmented to include an additional lag of f_t. The calculation of $\hat{\gamma}_i$ follows from a straightforward application of Bayes' formula.

Algorithm 1 PX-EM Algorithm

while $\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}))$ > tolerance level **do**

E Step:

Run a Kalman smoother to obtain $\hat{F}$, $\hat{G}$ $\hat{P}$, $\hat{C}$, and $\hat{P}_g$.

for $1 \leq i \leq N$ **do**

$\hat{\gamma}_i = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}}\frac{\alpha_1}{\alpha_0}\exp((\alpha_0-\alpha_1)\sigma^2_{i,n-1}))^{-1}$.

end for

M Step:

$\Lambda^*_n = X'\hat{F}(\hat{F}'\hat{F}+\hat{P})^{-1}$

for $1 \leq i \leq N$ **do**

$SS_i = \sum_t (X_{it} - \Lambda^*_{i,n}\hat{f}_t)^2 + \Lambda^*_{i,n}\hat{P}\Lambda^{*\prime}_{i,n}$

$\alpha^*_i = (1-\hat{\gamma}_i)\alpha_0 + \hat{\gamma}_i\alpha_1$

$\sigma^2_{i,n} = \frac{SS_i}{\frac{1}{2}(T+\sqrt{T^2+2\alpha^*_i SS_i})}$

$\rho_{i,n} = \frac{\hat{\gamma}_i+1-a}{2a-1}$

end for

$\{\Phi^*_n, \Omega^*_n\} = argmax_{\Phi^*_n,\Omega^*_n} Q(\theta|\theta_{n-1})$

Rotation Step:

Calculate A^L_n, the lower Cholesky factor of Ω^*_n.

$\Lambda_n = \Lambda^*_n A^L_n$

for $1 \leq l \leq p$ **do**

$\Phi_{l,n} = A^{L^{-1}}_n \Phi^*_{l,n} A^L_n$

end for

end while

1.2.2 The PX-ECME Algorithm for a DFM

The algorithm described in the previous section has faster convergence properties than the standard EM algorithm, but still encounters some difficulties maximizing the posterior, especially with respect to Σ, the parameters of greatest interest. To overcome this issue, I occasionally supplement the iterations of the PX-EM algorithm with an iteration from an ECME algorithm. The ECME algorithm works by iteratively maximizing functions of parameter blocks that are conditioned on the values of the remaining parameters from the last iteration. Better convergence properties are obtained by allowing the functions to be either conditional Q functions, such as $\mathbb{E}[\ln\pi(\theta_1|X, F, \theta_{2,n-1})|X, \theta_{n-1}]$ or conditional log integrated densities, such as $\ln\pi(\theta_1|X, \theta_{2,n-1})$. For the problem at hand, I choose to update ρ using a conditional Q function and update the remaining parameters with conditional posterior densities. All of the conditional maximizations must be unique in order for the sequence of posterior ordinates generated by the ECME algorithm to converge (Liu and Rubin, 1994). The posterior distribution obviously does not have a unique maximum, but the conditional distributions do when the parameters are grouped by observation equation (1) and state equation (2). One option for an ECME iteration would be to first maximize $\ln\pi(\Lambda, \Sigma|X, \Phi_{n-1}, \Omega = I_r)$ with respect to Λ_n and Σ_n, then maximize $\ln\pi(\Phi|X, \Lambda_n, \Sigma_n, \Omega = I_r)$ with respect to Φ_n. I instead modify this step with parameter expansion by first maximizing $\ln\pi(\Lambda^*, \Sigma|X, \Phi_{n-1}, \Omega = I_r)$ with respect to Λ_n^* and Σ_n, then maximizing $\ln\pi(\Phi^*, \Omega^*|X, \Lambda_n^*, \Sigma_n)$ with respect to Φ_n^* and Ω_n^*. The solutions are then rotated back to the scale identified model, as in the PX-EM algorithm. Details are given in Algorithm 2. Except for ρ, all maximizations are done numerically using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. Computation time is decreased by using the closed form solution for the gradient that results from the identity $\nabla\ln\pi(\theta_n|X) = \nabla Q(\theta_n|\theta_n)$ (Ruud, 1991).

Algorithm 2 PX-ECME Algorithm

while $\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1})) >$ tolerance level **do**

E Step:

for $1 \leq i \leq N$ **do**

$\hat{\gamma}_i = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}} \frac{\alpha_1}{\alpha_0} \exp((\alpha_0 - \alpha_1)\sigma^2_{i,n-1}))^{-1}$.

end for

M Step:

for $1 \leq i \leq N$ **do**

$\rho_{i,n} = \frac{\hat{\gamma}_i + 1 - a}{2a-1}$

end for

$\{\Lambda^*_n, \Sigma\} = argmax_{\Lambda^*,\Sigma} \ln\pi(\Lambda, \Sigma|X, \rho_n, \Phi_{n-1}, \Omega = I_r)$

$\{\Phi^*_n, \Omega^*_n\} = argmax_{\Phi^*_n,\Omega^*_n} \ln\pi(\Phi^*, \Omega^*|X, \rho_n, \Lambda^*_n, \Sigma_n)$

Rotation Step:

Calculate A^L_n, the lower Cholesky factor of Ω^*_n.

$\Lambda_n = \Lambda^*_n A^L_n$

for $1 \leq l \leq p$ **do**

$\Phi_{l,n} = A^{L^{-1}}_n \Phi^*_{l,n} A^L_n$

end for

end while

1.2.3 Approximations to the Stationary Likelihood

Working with the stationary likelihood is theoretically ideal, but presents several challenges. Maximization of $Q(\theta|\theta_n)$ with respect to the VAR parameters must be done numerically, as no closed form solutions exist. Let us consider the state equation when it is rewritten from a $VAR(p)$ equation to a $VAR(1)$ equation. Let $g_t = (f'_t, f'_{t-1}, \ldots, f'_{t-p+1})'$.

$$g_t = Bg_{t-1} + \begin{bmatrix} \eta_t \\ 0_{r(p-1)\times 1} \end{bmatrix} \tag{1.15}$$

Rather than work with the stationary variance of the factor process, I will instead approximate the stationary likelihood by assuming the first p presample instances of the factors follow the distribution

$$g_0 \sim N(0, \nu_{g_0} I_{pr}). \tag{1.16}$$

Integration over these presample instances of f_t yields a distribution for the first p instances of f_t of the form

$$g_p \sim N(0, \Omega_g + \nu_{g_0} B^p B^{p\prime} + \sum_{j=1}^{p-1} B^j \Omega_g B^{j\prime}), \;\; \Omega_g = \begin{bmatrix} \Omega & \\ & 0_{r(p-1)} \end{bmatrix}. \tag{1.17}$$

This functions as an approximation to the stationary distribution. The approximation could be made arbitrarily accurate by making the number of presample factors τ sufficiently large. As $\tau \to \infty$, the marginal covariance matrix of g_p will converge to the stationary covariance matrix P_0. However, such an approach will drastically reduce the efficiency of the EM algorithm. As τ increases, the curvature of $Q(\theta|\theta_n)$ increases, leading to smaller steps being taken in each parameter update. There is likely a more optimal choice of τ. A researcher may run a preliminary algorithm with a small number of iterations, then select his or her preferred τ by choosing a number such that $||P_{0,\tau} - P_0|| < m$, where m is a positive tuning parameter, $||\cdot||$ is a matrix norm, and $P_{0,\tau}$ is the covariance matrix of g_p for a given τ. Alternatively, one could increase τ until the convergence time of the algorithm begins to suffer. While these approximations greatly increase the efficiency of the algorithm, the likelihood is no longer rotation invariant. The parameter-expanded algorithms I have developed are thus no longer guaranteed to produce updates that monotonically increase the likelihood. One way to make the algorithm monotonic is to only perform rotations if they increase the posterior density, and just perform regular EM updates otherwise. Another option is to only do parameter-expanded steps for a set number of iterations, then switch to a basic EM algorithm. Despite the loss of monotonicity, any fixed points of the parameter-expanded algorithms will also be fixed points of the basic EM algorithm. Non-monotonic updates were not an issue in

applications to simulated or real data, while convergence was markedly faster. The PX-EM algorithm with the likelihood approximation is given in Algorithm 3. In addition to the posterior moments defined previously, let $\hat{g}_t \equiv \mathbb{E}[g_t|X,\theta_n]$, $\hat{G} \equiv (\hat{g}_0,\ldots,\hat{g}_{T-1})'$, $\hat{P}_g \equiv \sum_t \mathbb{E}[(g_{t-1}-\hat{g}_{t-1})(g_{t-1}-\hat{g}_{t-1})'|X,\theta_n]$, and $\hat{C} \equiv \sum_t \mathbb{E}[(f_t-\hat{f}_t)(g_{t-1}-\hat{g}_{t-1})'|X,\theta_n]$. The reader will note that all parameters updates are now available in closed-form.

Another option, should one wish to work with the exact stationary likelihood, is to only use ECME steps for updating the parameters of the state equation. Gradient-based methods for this problem require care. Calculation of the numerical gradient requires many runs of either a Kalman filter or a precision-based method for obtaining the integrated likelihood, as well as many high-dimensional matrix inversions to calculate the stationary variance. A precise approximation of the gradient can be computed in significantly less time by augmenting the state vector with many presample factors and using the fact that $\nabla \ln\pi(\theta_n|X) = \nabla Q(\theta_n|\theta_n)$ (Ruud, 1991). Justification for this approach is given by Proposition 1.1.

Proposition 1.1

Let $Q_{\tau-p}(\theta|\theta_n) \equiv \mathbb{E}[\ln\pi(X,F,f_0,f_{-1},\ldots,f_{-\tau+p+1},\theta|f_{-\tau+p},\ldots,f_{-\tau+1})|X,\theta_n]$ and assume θ_n is an interior point of the parameter space.

$$\lim_{\tau\to\infty} \nabla Q_{\tau-p}(\theta_n|\theta_n) = \nabla \ln\pi(\theta_n|X).$$

A proof of this proposition can be found in Appendix A.

The benefit of this approach is that it only requires one matrix inversion and one Kalman smoother run, as opposed to many matrix inversions and Kalman filter runs. One only has to work with the conditional elements of the likelihood, so the gradient is available in closed

Algorithm 3 PX-EM Algorithm with Approximate Likelihood

while $\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1})) >$ tolerance level **do**

E Step:

Run a Kalman smoother to obtain $\hat{F}$, $\hat{G}$ $\hat{P}$, $\hat{C}$, and $\hat{P}_g$.

for $1 \leq i \leq N$ **do**

$\hat{\gamma_i} = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}} \frac{\alpha_1}{\alpha_0} \exp((\alpha_0 - \alpha_1)\sigma^2_{i,n-1}))^{-1}$.

end for

M Step:

$\Lambda^*_n = X'\hat{F}(\hat{F}'\hat{F} + \hat{P})^{-1}$

for $1 \leq i \leq N$ **do**

$SS_i = \sum_t (X_{it} - \Lambda^*_{i,n}\hat{f}_t)^2 + \Lambda^*_{i,n}\hat{P}\Lambda^{*\prime}_{i,n}$

$\alpha^*_i = (1 - \hat{\gamma_i})\alpha_0 + \hat{\gamma_i}\alpha_1$

$\sigma^2_{i,n} = \frac{SS_i}{\frac{1}{2}(T + \sqrt{T^2 + 2\alpha^*_i SS_i})}$

$\rho_{i,n} = \frac{\hat{\gamma_i}+1-a}{2a-1}$

end for

$\Phi^*_n = (\Phi^*_{1,n}, \ldots, \Phi^*_{p,n}) = (\hat{F}'\hat{G} + \hat{C})(\hat{G}'\hat{G} + \hat{P}_g)^{-1}$

$\Omega^*_n = T^{-1}(\sum_t (\hat{f}_t - \Phi^*\hat{g}_t)(\hat{f}_t - \Phi^*\hat{g}_t)' + \Phi^*\hat{P}_g\Phi^{*\prime} - \Phi^*\hat{C}' - \hat{C}\Phi^{*\prime})$

Rotation Step:

Calculate A^L_n, the lower Cholesky factor of Ω^*_n.

$\Lambda_n = \Lambda^*_n A^L_n$

for $1 \leq l \leq p$ **do**

$\Phi_{l,n} = A^{L^{-1}}_n \Phi^*_{l,n} A^L_n$

end for

end while

form. This result is also applicable to stationary VARs and vector autoregressive moving average (VARMA) models. It could be used for maximizing the likelihoods of these models or for efficient simulation from the posterior distribution using Hamiltonian Monte Carlo, which is an area of active research (Heaps, 2023; Binks et al., 2023).

1.2.4 Specification of α_q

α_1 should be set so as to have minimal influence on variance estimates. I adopt the convention of $\alpha_1 = 0.01$ in all estimations. Optimal specification of α_0 is not obvious. Values that are too small will not impose sufficient shrinkage on small variances. However, setting α_0 too high means that the EM algorithm is unlikely to assign significant weight to the spike component of the prior, and the resulting estimates will be close to the maximum likelihood estimates. Rather than try to find a single optimal α_0, I adopt the dynamic posterior exploration approach developed by Ročková and George (2016). The authors, drawing on concepts from deterministic annealing, estimate a series of models with increasingly pronounced spike distributions. This is done by using a ladder of increasing spike parameters $\alpha_0 \in I = \{\alpha_0^1, \alpha_0^2, \ldots, \alpha_0^L\}$. α_1 is held constant. Small values of α_0 produce a flatter posterior density that is easier to traverse. As α_0 increases, the posterior becomes spikier. Each estimation is initialized with the MAP estimates from the previous estimation. This "warm start" approach makes the global mode easier to find. The intersection point of the spike and slab densities is given by $\delta(\alpha_1, \alpha_0, \rho) = \frac{1}{\alpha_0 - \alpha_1} \ln\left(\frac{\alpha_0}{\alpha_1}\frac{1-\rho}{\rho}\right)$. The sequence I is defined implicitly by the sequence $\delta(\alpha_1, \alpha_0, \rho = .5) \in I_\delta = \{\delta^1, \delta^2, \ldots, \delta^L\}$. I use the sequence $I_\delta = \{.5, .25, .1, .05, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ for all estimations.

1.2.5 Missing data

It is rare for a researcher to be blessed with a balanced panel of data. Very often variables are not available for the entire sample period. It may be that they were only recorded after a certain date or were eventually discontinued. It could also be the case that certain entries are intentionally trimmed by the researcher to control for outliers. Missing data represent a major issue in the frequentist literature when factors are estimated with PCA. Missing values must be imputed with a consistent estimator (Stock and Watson, 2002; Jin, Miao, and Su, 2021). Likelihood-based method do not have the same problem. One only has to restrict the vector of dependent variables in each period to those with non-missing values.

I adjust the variance selection prior to account for differing sample sizes in the presence of missing data. Using the same α_0 and α_1 for every time series would disproportionately penalize the variances of variables with many missing values. Let T_i be the number of time periods for which X_{it} is observed. The variable-specific hyperparameters are then defined as $\alpha_{iq} \equiv \frac{T_i}{T}\alpha_q$. The correction term $\frac{T_i}{T}$ ensures that, conditional on γ_i, the prior has the same influence on each σ_i^2.

Optimization can proceed with only minor modifications to the algorithms. Posterior moments are obtained using a Kalman smoother adjusted for missing data. Let the set of time periods O_i be defined as $O_i \equiv \{t : X_{it} \text{ is observed}\}$. Let τ_{im} be the m^{th} entry in O_i. We now define $\hat{F}_i \equiv (\hat{f}_{\tau_{i1}}, \ldots, \hat{f}_{\tau_{iT_i}})'$, $\hat{P}_i \equiv \sum_{t \in O_i} \mathbb{E}[(f_t - \hat{f}_t)(f_t - \hat{f}_t)' | X, \theta_n]$. Algorithm 4 gives the details for a modified PX-EM algorithm.

Algorithm 4 PX-EM Algorithm with Approximate Likelihood and Missing Data

while $\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1}) - (\ln f(X|\theta_{n-1}) + \ln\pi(\theta_{n-1})) >$ tolerance level **do**

E Step:

Run a Kalman smoother to obtain $\{\hat{F}_i\}$, $\hat{G}$, $\{\hat{P}_i\}$, $\hat{C}$, and $\hat{P}_g$.

for $1 \leq i \leq N$ **do**

$\hat{\gamma}_i = (1 + \frac{\rho_{i,n-1}}{1-\rho_{i,n-1}} \frac{\alpha_{i1}}{\alpha_{i0}} \exp((\alpha_{i0} - \alpha_{i1})\sigma^2_{i,n-1}))^{-1}$.

end for

M Step:

for $1 \leq i \leq N$ **do**

$\Lambda^*_{i,n} = X_i' \hat{F}_i (\hat{F}_i' \hat{F}_i + \hat{P}_i)^{-1}$

$SS_i = \sum_{t \in O_i} (X_{it} - \Lambda^*_{i,n} \hat{f}_t)^2 + \Lambda^*_{i,n} \hat{P}_i \Lambda^{*\prime}_{i,n}$

$\alpha^*_i = (1 - \hat{\gamma}_i)\alpha_{i0} + \hat{\gamma}_i \alpha_{i1}$

$\sigma^2_{i,n} = \frac{SS_i}{\frac{1}{2}(T_i + \sqrt{T_i^2 + 2\alpha^*_i SS_i})}$

$\rho_{i,n} = \frac{\hat{\gamma}_i + 1 - a}{2a - 1}$

end for

$\Phi^*_n = (\Phi^*_{1,n}, \ldots, \Phi^*_{p,n}) = (\hat{F}'\hat{G} + \hat{C})(\hat{G}'\hat{G} + \hat{P}_g)^{-1}$

$\Omega^*_n = T^{-1}(\sum_t (\hat{f}_t - \Phi^* \hat{g}_t)(\hat{f}_t - \Phi^* \hat{g}_t)' + \Phi^* \hat{P}_g \Phi^{*\prime} - \Phi^* \hat{C}' - \hat{C} \Phi^{*\prime})$

Rotation Step:

Calculate A^L_n, the lower Cholesky factor of Ω^*_n.

$\Lambda_n = \Lambda^*_n A^L_n$

for $1 \leq l \leq p$ **do**

$\Phi_{l,n} = {A^L_n}^{-1} \Phi^*_{l,n} A^L_n$

end for

end while

1.3 Monte Carlo Studies

This section presents the results of various Monte Carlo studies. In each simulation, the loadings are randomly generated according to $\Lambda_{ij} \sim N(0, 1)$. I consider the case of $p = 1$ lags. Φ is randomly generated using its eigendecomposition $\Phi = VDV^{-1}$. The elements of the eigenvector matrix are distributed $V_{jj'} \sim \mathcal{U}(-1, 1)$ and the eigenvalues are distributed $D_{jj} \sim \mathcal{U}(.4, .6)$. $\Omega = \omega I_r$. ω is chosen such that $\text{Var}(\Lambda_i f_t) = r$.[3]I first examine datasets with every possible combination of N, T, and r for which $N \in \{40, 60, 100, 200\}$, $T \in$

3. This is done by first calculating the stationary covariance matrix P_0 of a process with transition parameters Φ and covariance matrix $\Omega = I_r$. Let C be the lower Cholesky factor of P_0 such that $P_0 = CC'$. The covariance matrix of innovations is then rescaled to $\Omega = \frac{r}{||\text{vech}(C)||_2^2} I_r$.

$\{50, 100, 150, 200, 250\}$, and $r \in \{2, 4, 6\}$. In each case, the number of observed and latent factors are equal: $r_f = r_y = r/2$. The factors are simulated by drawing the first p instances from the stationary distribution and then iterating the data generating process forward through time. Each study analyzes 100 simulated datasets.

The first simulation study assumes a balanced panel and $\sigma_i^2 = r$. The results can be seen in Figure 1.2. The plots give the proportion of datasets for which the procedure correctly identified the true observed factors. The proposed approach is quite good at identifying the observed factors in most cases. The one noticeable limitation is that results suffer when N and T are small and r is large. This is hardly surprising, as we are asking a lot of the model and not providing sufficient data. Thankfully, the success rate is quite high for combinations of N and T that we are likely to encounter in practice.

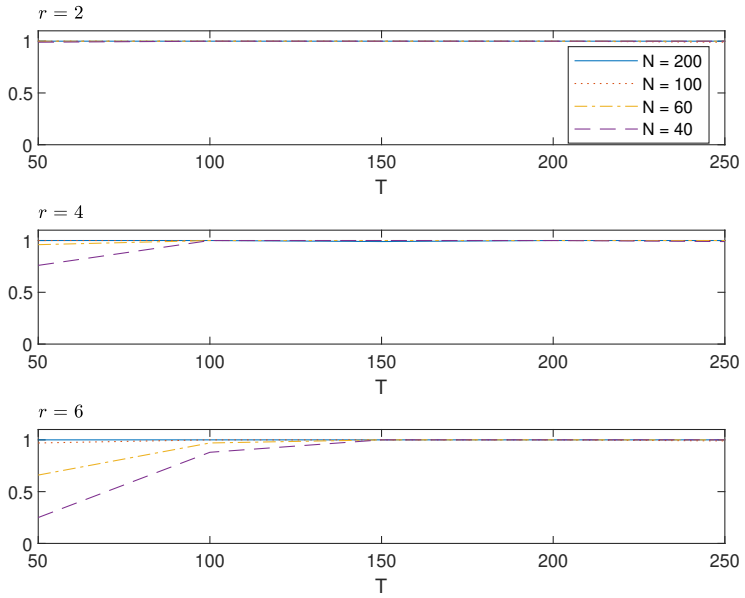


Figure 1.2: Proportion of Models Correctly Identified, $\sigma_i^2 = r$

Figure 1.3 gives the results of a simulation study that uses the exact same parameters and factors as Table Figure 1.2, only the idiosyncratic variance is now set to $\sigma_i^2 = 2r$. We observe a slight decrease in accuracy for small values of N and T. This is to be expected because

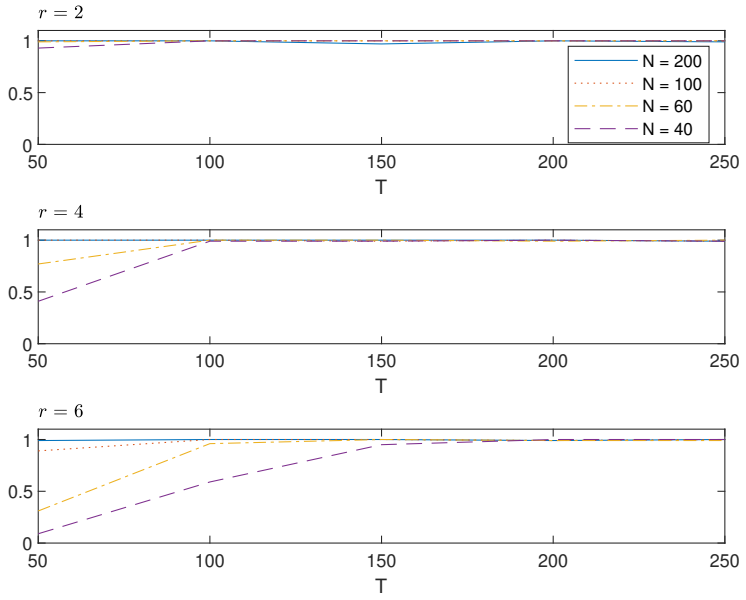


Figure 1.3: Proportion of Models Correctly Identified, $\sigma_i^2 = 2r$

the signal-to-noise ratio has decreased and the factors will not be estimated as precisely. However, we see no noticeable drop in accuracy for $N \geq 60$ and $T \geq 100$.

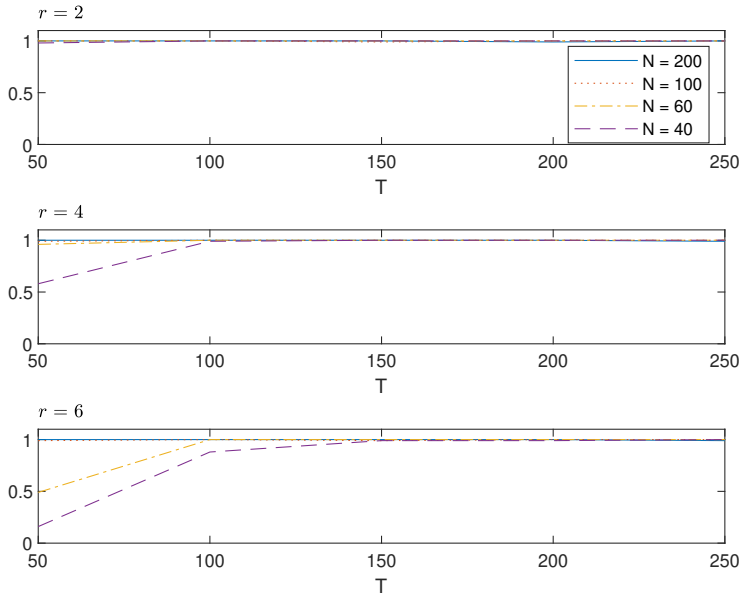


Figure 1.4: Proportion of Models Correctly Identified, $\sigma_i^2 = r$, $p_{miss} = 0.05$

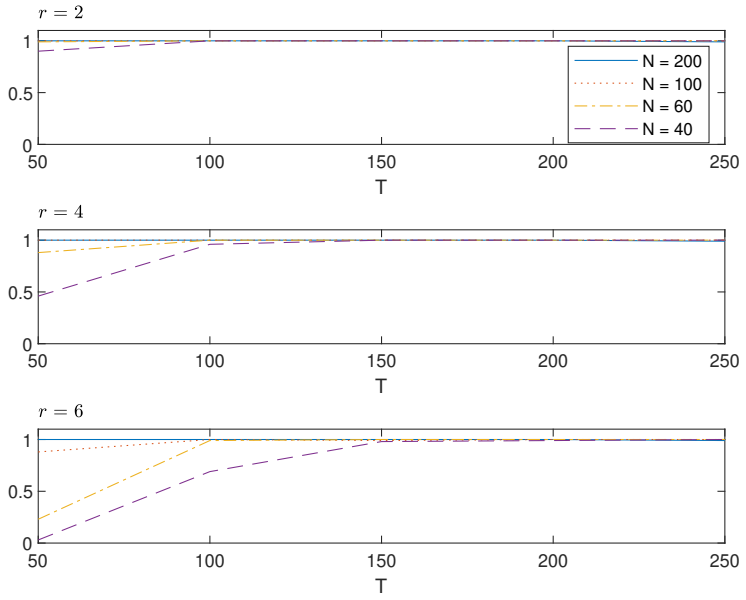


Figure 1.5: Proportion of Models Correctly Identified, $\sigma_i^2 = r$, $p_{miss} = 0.1$

Figures 1.4 and 1.5 give the results of Monte Carlo studies in which $\sigma_i^2 = r$ and a proportion p_{miss} of the data is missing. The values considered are $p_{miss} = 0.05, 0.1$. There appears to be no substantive difference between the results with missing data and the results with a balanced panel for $N \geq 60$ and $T \geq 100$.

1.4 Applications

1.4.1 Quarterly Macroeconomic Data

This section applies the model selection procedure developed above to the FRED-QD dataset (McCracken and Ng, 2020). For the first application, the dataset consists of $N = 246$ macroeconomic variables observed at quarterly intervals. Observations begin in 1959:Q3 and end in 2023:Q1. 38 of the variables were not recorded until midway through the sample period. Variables were transformed to be approximately stationary using the recommended trans-

formation codes of the authors, and then standardized to have unit variance. I performed outlier detection using the same criterion as McCracken and Ng (2020). Any observations that deviated from the sample median by more than ten interquartile ranges were classified as outliers and treated as missing. Initial factor estimates were obtained by replacing missing values with 0 and then using PCA. Jin, Miao, and Su (2021) show that this is a consistent estimator of the true factor space. I analyzed the full sample period as well as the subsamples 1959:Q3 - 2007:Q3 and 2007:Q4 - 2023:Q1. The sample was partitioned to examine any structural changes that may have occurred after the 2007 financial crisis. The outlier classification criterion detected 90 outliers in the full sample, 5 outliers in the pre-financial crisis subsample, and 109 outliers in the post-financial crisis subsample. The IC_{p2} criterion selected 8 factors for the full sample period and 6 factors for each of the subsamples.

As can be seen from Table 1.1, Capacity Utilization: Total Industry (TCU) is selected as an observed factor for both the full sample and pre-2007 estimations. TCU is an index that measures the percentage of potential feasible output that is being produced. This is a surprising but not unreasonable finding. Capacity utilization has long been recognized as a leading indicator for inflation and business cycles (Corrado and Mattey, 1997). That TCU was selected demonstrates the necessity of being able to incorporate missing data. TCU was not recorded until 1967:Q1. The existing frequentist methods require a balanced panel dataset, and thus would not have been able to detect this relationship over the periods considered. Another advantage of the Bayesian approach is that unobserved values of observed factors can be imputed naturally using the output from the Kalman smoother.

Figure 1.6 shows the precise estimates that are obtained for the period 1959:Q3 - 1968:Q4 using this method. TCU is not selected in the post-2007 estimation. With this in mind, the behavior of TCU does seem to be different before and after the financial crisis. Capacity utilization tends to peak in the middle of expansions prior to 2007. The index is already declining prior to the onset of recessions during this period. The inter-recession shape of

Period	N	T	r	y
1959:Q3-2023:Q1	246	255	8	Capacity Utilization: Total Industry
1959:Q3-2007:Q3	246	193	6	Capacity Utilization: Total Industry
2007:Q4-2023:Q1	246	62	6	Business Sector: Real Output All Employees: Service-Providing Industries All Employees: Goods-Producing Industries

Table 1.1: Likely Observed Factors in the U.S. Economy, Quarterly Data

TCU appears different after 2007. It is approximately level during 2007 and only starts to decline after the 2008 recession has already begun. Estimates of all 8 factors from the full sample estimation are plotted in Figure 1.7. One can see that estimates of factor 5 are nearly identical to a one period lag of TCU. This suggests that the true Ω might be of reduced rank (Bai and Ng, 2007). It also indicates that TCU is not only an important driver of the economy, but its impact is also persistent.

1.4.2 Monthly Macroeconomic Data

This sections analyzes the FRED-MD dataset (McCracken and Ng, 2016). It consists of $N = 127$ monthly macroeconomic variables over the period 1959:3-2023:6. The variables are transformed using the authors' recommended transformations and standardized to have unit variance. Outliers are identified and removed using the criterion previously discussed. As with the quarterly data, I analyze the full sample, as well as pre- and post-financial crisis subsamples. Results are given in Table 1.2.

The only variable identified as an observed factor in the full sample estimation is 10-Year Treasury Constant Maturity Minus Federal Funds Rate (T10YFFM). This is very similar to measures of the slope of the yield curve, which has been studied for its relationship to business cycles. Figure 1.8 plots T10YFFM along with NBER recession dates. We can see that T10YFFM often turns negative near the peak of an expansion and then sharply increases

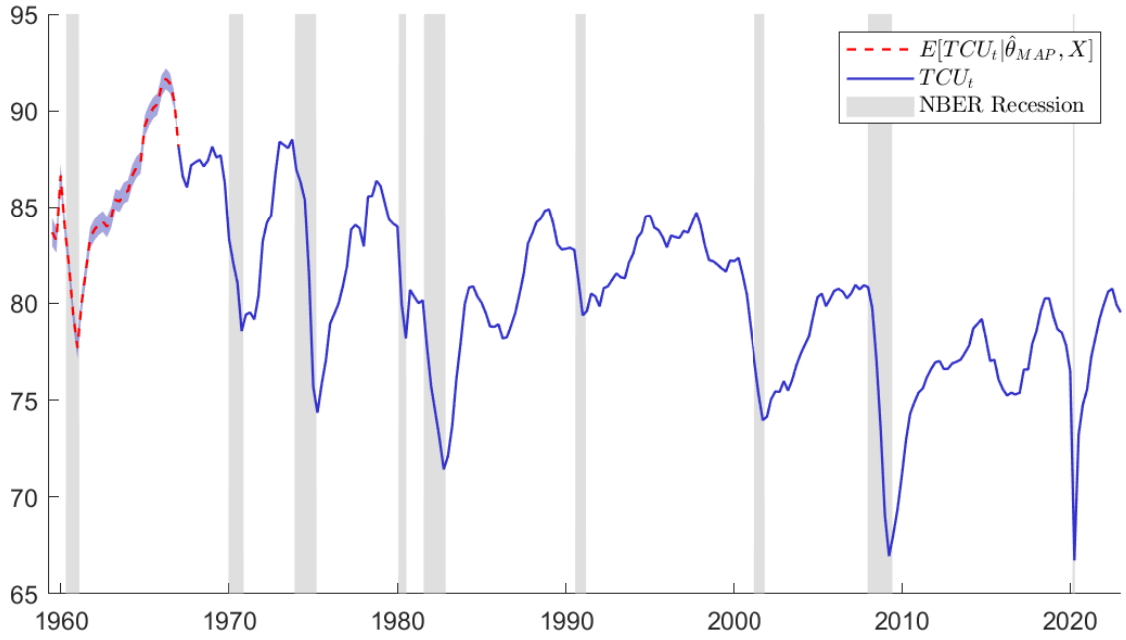


Figure 1.6: Total Capacity Utilization (Observed and Imputed)

Notes: The shaded region around the imputed values of TCU_t is a 95% credible interval. The variance of TCU_t is available directly from the Kalman smoother.

during recessions. The pre-financial crisis estimation selects a related variable: Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate (BAAFFM). The correlation between T10YFFM and BAAFFM during this period is 0.94, so the relationship between BAAFFM and business cycles is nearly identical. The correlation between the two variables drops to 0.89 in the post-2007 subsample. One possible reason that T10YFFM is selected in the full sample estimation is that the Federal Reserve began targeting the yield curve directly after the financial crisis.

While estimations using monthly and quarterly data produce differing results in the pre-financial crisis subsamples, there is some overlap in the selected observed factors for the post-financial crisis subsamples. They both select All Employees: Service-Providing Industries, along with one other employment measure. The importance of service sector employment may stem from its correlation with the impacts of the COVID-19 pandemic.

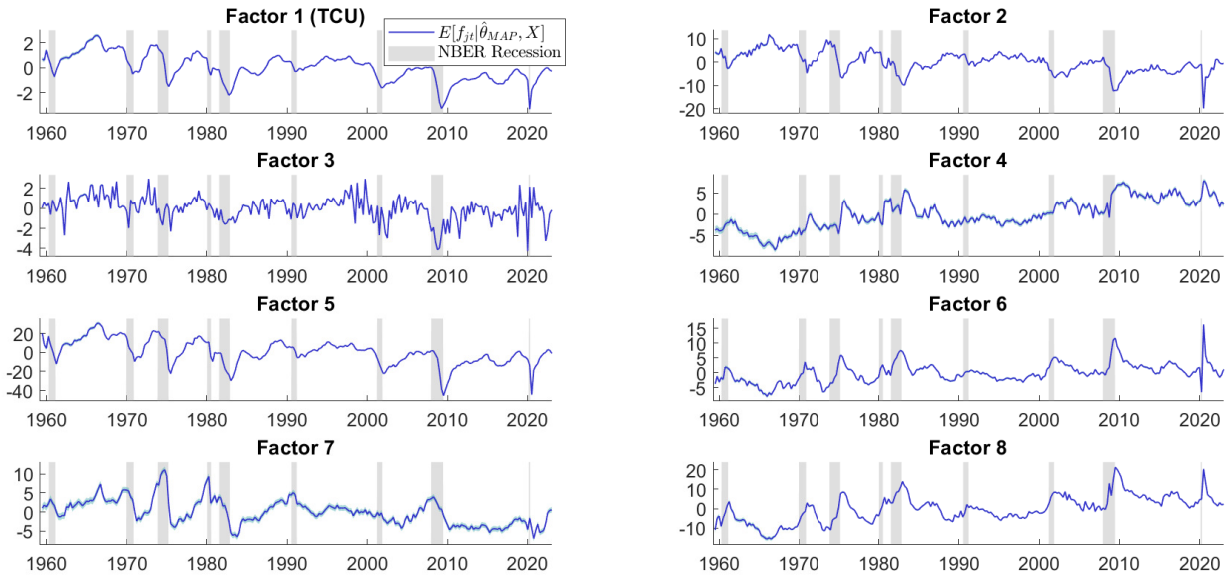


Figure 1.7: Factor Estimates from FRED-QD

Notes: The shaded region around the estimated values of f_{jt} is a 95% credible interval. The variance of f_{jt} is available directly from the Kalman smoother.

1.4.3 Fama-French Portfolio Data

I will now use the model selection process to investigate the asset-pricing model of Fama and French (1993). The authors extend the capital asset pricing model to include factors that measure the excess returns attributable to firm size and book-to-market equity ratio (BE/ME). The Fama-French three-factor model is given by

$$X_{it} = R_{it} - R_t^f = \beta_0 + \beta_{1i}(R_t^m - R_t^f) + \beta_{2i}SMB_t + \beta_{3i}HML_t + \varepsilon_{it}, \tag{1.18}$$

where R_{it} is the return on portfolio i, R_t^m is the return on a market portfolio, R_t^f is the risk-free return, SMB_t is the firm size factor, and HML_t is the BE/ME factor. I estimate models for a dataset that includes the three Fama-French factors, their measure of the risk-free rate, and the excess returns from 100 portfolios. The portfolios are the intersection of 10 portfolios organized by deciles of firm size and 10 portfolios organized by deciles of BE/ME.

Period	N	T	r	y
1959:3-2023:6	127	772	7	10-Year Treasury Constant Maturity Minus Federal Funds Rate
1959:3-2007:9	127	583	7	Moody's Seasoned Baa Corporate Bond Minus Federal Funds Rate
2007:10-2023:3	127	189	7	All Employees, Total Nonfarm All Employees: Service-Providing Industries Consumer Price Index for All Urban Consumers: All Items in U.S. City Average S&P 500

Table 1.2: Likely Observed Factors in the U.S. Economy, Monthly Data

The data was collected from Kenneth French's website [4]. Estimating such a model allows us to test whether the 3 factor specification is supported by the data. Factor observations and incomplete portfolio data are available for the period 1926:7-2023:6. I estimated a model for the full sample period as well as a number of subsamples. The subsamples include the time periods considered by Bai and Ng (2006) as well several others. The time periods not previously examined are the interwar period of 1926:7-1945:8, the Bretton Woods period of 1945:9-1972:12, the pre-financial crisis period of 1997:1-2007:9, and the post-financial crisis period of 2007:10-2023:6. Previous studies only examined the validity of the three-factor model after 1960 and did not include all 100 portfolios. Researchers had to delete several portfolios as well as many time periods because their methods required a balanced panel. Results are given in Table 1.3.

The importance of BE/ME is quite stable. It is selected as an observed factor in the full sample estimation as well as every subsample estimation except for 1973:1-1987:12. Firm size is selected in the full sample estimation, but not in the subsamples for 1973:1-1987:12, 1988:1-1996:12, 1997:1-2007:9, and 1960:1-1996:12. A surprising result is the selection of the portfolio of firms in the tenth deciles of size and BE/ME in 2 subsamples. However, this

4. See `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` for further information.

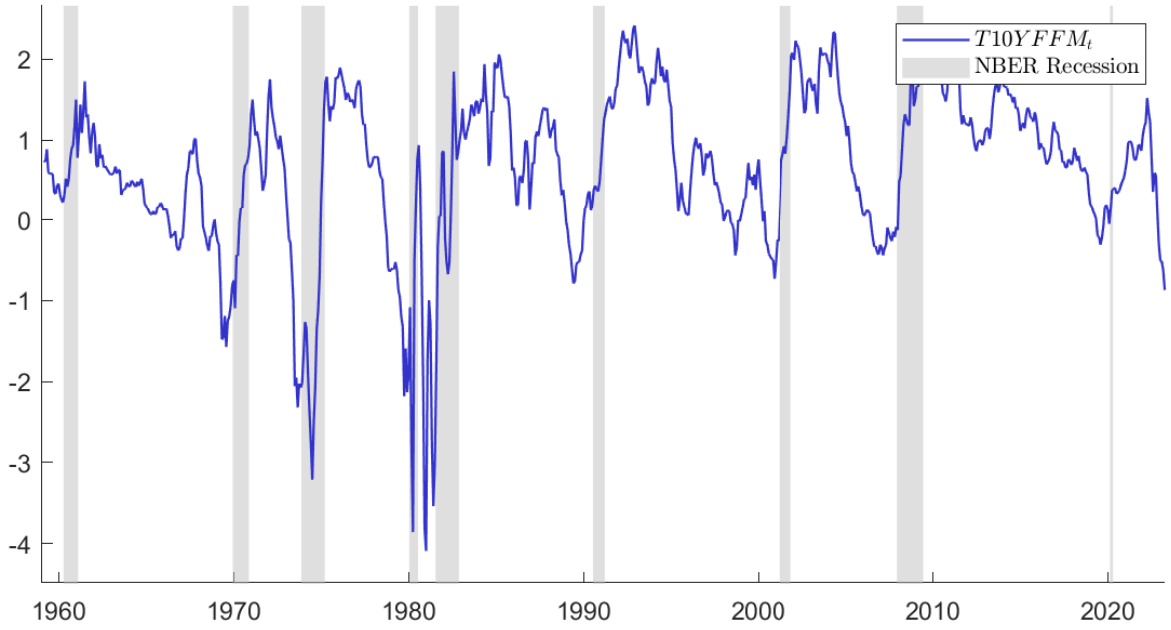


Figure 1.8: 10-Year Treasury Constant Maturity Minus Federal Funds Rate

result should be treated skeptically because there are very few observations of the variable in these subperiods, so there is a good chance of overfitting. The most glaring result is that market excess return is not selected in any estimation. Although the market variable is not selected, we should not interpret this as evidence that market return plays no role in portfolio returns. The estimated variance of the idiosyncratic error for the market variable is less than 0.01 in all but two of the estimations. This suggests that excess market return or some closely related variable is a fundamental factor, but it is not perfectly observed. Figure 1.9 plots the market variable over the entire sample period along with its fitted values. We can see that there is very little difference between the two. A more surprising result is the occasional selection of the risk-free return as an observed factor. This suggests that excess returns depend on R_t^f in a way that is not simply a function of their dependence on excess market return. It is important to note that the two time periods in which R_t^f is selected include the period in the 1970s and early 1980s when interests rates were extremely volatile.

Period	N	T	r	y
1926:7-2023:6	104	1,164	4	Firm Size Book-to-Market Equity Ratio
1926:7-1945:8	104	230	4	Firm Size Book-to-Market Equity Ratio Portfolio of firms in the tenth deciles of size and BE/ME
1945:9-1972:12	104	328	3	Firm Size Book-to-Market Equity Ratio
1973:1-1987:12	104	170	4	Risk-Free Return
1988:1-1996:12	104	108	3	Book-to-Market Equity Ratio
1997:1-2007:9	104	109	5	Book-to-Market Equity Ratio Portfolio of firms in the tenth deciles of size and BE/ME
2007:10-2023:6	104	189	4	Firm Size Book-to-Market Equity Ratio
1960:1-1996:12	104	444	4	Book-to-Market Equity Ratio
1960:1-1982:12	104	276	4	Firm Size Book-to-Market Equity Ratio Risk-Free Return
1982:1-1996:12	104	168	3	Firm Size Book-to-Market Equity Ratio

Table 1.3: Likely Observed Factors in Monthly Fama-French Portfolios

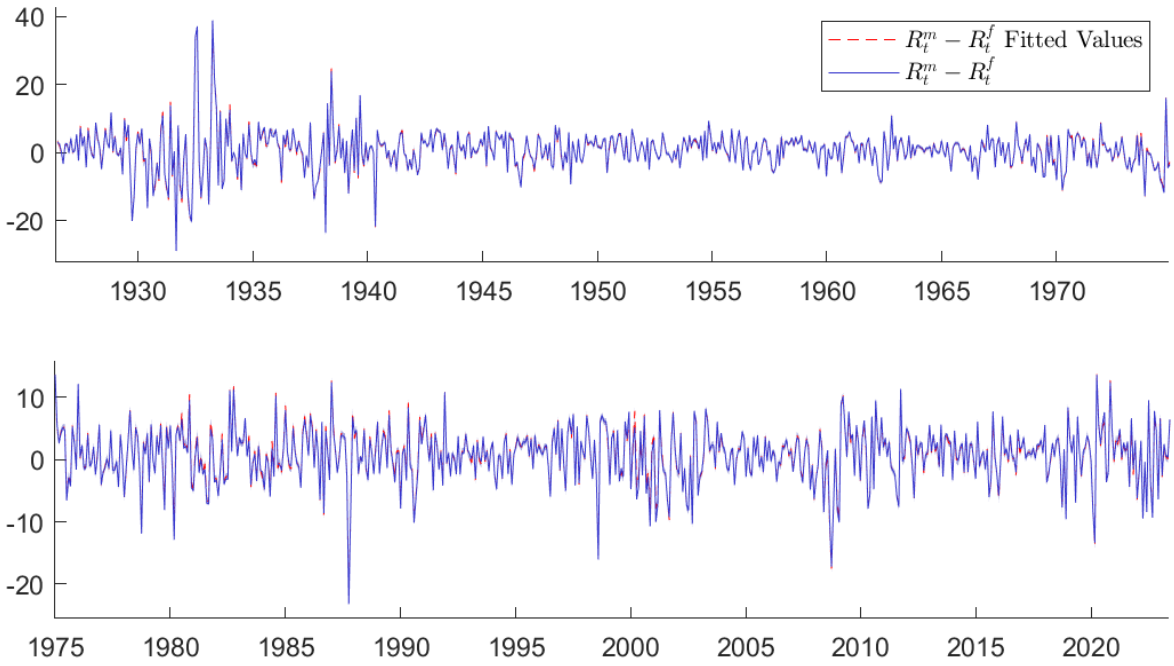


Figure 1.9: Actual and Fitted Values of $R_t^m - R_t^f$

Notes: The shaded region around the fitted values of $R_t^m - R_t^f$ is a 95% credible interval. The variances of the fitted values are available directly from the Kalman smoother.

1.5 Conclusion

I proposed a model selection procedure for FAVARs. Estimation of the total number of factors and the lag length is done using existing methods, although the use of BIC for lag length selection is modified to avoid model misspecification problems. The selection of observed factors is achieved using a Bayesian shrinkage prior. The prior allows us to recast a high-dimensional model selection process as an optimization problem. This enables researchers to differentiate between millions of potential models by estimating just a single model. The procedure has very good small sample properties. Model selection accuracy was virtually 100% in simulated datasets of realistic size.

Several extensions to the EM algorithm for estimating DFMs were proposed. The resulting PX-ECME algorithm exhibited faster convergence properties than the basic EM algorithm. I also developed an efficient and precise method for calculating the gradient of the log-likelihood of stationary VARMA processes, of which the FAVAR is a special case.

The model selection procedure yielded interesting results when applied to macroeconomic and financial data. The Total Capacity Utilization index was the only observed factor detected in a large dataset of quarterly U.S. macroeconomic variables. The spread between the 10-Year Treasury Constant Maturity Rate and the Federal Funds Rate was the only observed factor detected for monthly data. A specification in which the Federal Funds Rate is the only observed factor, the default assumption in the FAVAR literature, was never selected. Finally, I used the model selection procedure to test the assumptions of the Fama and French (1993) asset-pricing model. The variables constructed for firm size and book-to-market equity ratio were often selected as observed factors, but excess market return was not. That excess market return was not selected is more likely the result of mismeasurement of the variable rather than lack of importance.

There are many avenues for further research. While the approach of this chapter seeks to

find the most likely model, it may be the case that there are multiple competing models with significant posterior probabilities. MCMC would be the appropriate means of estimation for this end. This model assumes homoskedastic Normal errors, which is unlikely to be realistic in macroeconomic and financial data. One could incorporate errors with stochastic volatility into the state equation as well as the observation equations. Allowing for stochastic volatility in the observation equations permits the possibility of the observed factors changing over time, which is a perfectly reasonable hypothesis.

Chapter 2

A Nonparametric Endogenous Switching Model with an Application to Macroeconomics

Until the past decade, time series econometrics has focused primarily on parametric models. This was true of both linear vector autoregressions (VARs) (Sims, 1980; Litterman, 1986; Primiceri, 2005; Koop and Korobilis, 2013) and mixture models (Beaudry and Koop, 1993; Sims and Zha, 2006; Uribe and Lopes, 2020). Early work in nonparametric time series models focused on approximating nonlinear conditional mean functions in either univariate or small multivariate processes (Auestad and Tjøstheim, 1990; Härdle, Tsybakov, and Yang, 1998; Hamilton, 2001). A good deal of the more recent work has focused on Dirichlet process mixture models (DPMs). DPMs have been used to model the error distribution of asset returns in stochastic volatility models (Jensen and Maheu, 2010; Delatola and Griffin, 2011). Babak (2009) used the DPM to identify regimes in U.S. real GDP growth, allowing the number of regimes to be selected by the model. Kalli and Griffin (2018) use a DPM to flexibly model VAR processes. Nonparametric VARs are an active area of research. Jeliazkov (2013)

modeled the conditional mean of each dependent variable as a sum of nonlinear univariate functions of explanatory variables. Huber and Rossini (2021) model the conditional mean of a VAR process using Bayesian additive regression trees. A consistent finding in the nonparametric VAR chapters is that relaxing the assumption of linearity leads to better forecasting performance.

One of the primary methodologies for introducing nonlinearity to time series econometrics has been the class of models known as Markov-switching models (MSMs). MSMs are an extension of the hidden Markov model (Baum and Petrie, 1966) to the case of a continuously-distributed dependent variable. They allow model parameters to switch between different regimes. MSMs were introduced by Goldfeld and Quandt (1973) and popularized by Hamilton (1989), who used a model of US Gross National Product growth as an alternative method for dating business cycle turning points. Since then, models with Markov-switching have been applied extensively to business cycles (Albert and Chib, 1993; Boldin, 1996; Ghysels, McCulloch, and Tsay, 1998; Chauvet and Hamilton, 2006) as well as financial data (Vigfusson, 1997; Haas, Mittnik, and Paolella. 2004; Guidolin and Timmermann, 2005). MSMs have expanded to include models with time-varying transition probabilities (TVTP) (Diebold, Lee, and Weinbach. 1993; Filardo, 1994; Filardo and Gordon, 1998) as well as state space models (Kim, 1994; Chauvet, 1998; Kim and Nelson, 1999). A more recent class of models allows for endogenous switching (Chib and Dueker, 2004; Kim, Piger, and Startz, 2008; Hwu, Kim, and Piger, 2021; Kim and Kang, 2022). These models allows the innovations in the equations governing regime transitions to be correlated with innovations in the observation equation. Flexible error distributions are almost entirely missing in the MSM literature. A Normal error distribution for the observation equation is assumed in virtually all models. Two notable exceptions are Dueker (1997) and Hwu and Kim (2023). The former modeled stock returns using a student's t-distribution where the degrees of freedom switch between different regimes. Hwu and Kim (2023) is the only MSM we have found where the observation errors have a nonparametric distribution. He develops a switching mean model where

the error distribution is generated by a Dirichlet process. Hwu and Kim (2023) assumes that switching is exogenous and transition probabilities are constant.

This chapter proposes a regime-switching linear model with TVTP, endogenous switching, and a nonparametric error distribution. Both of these qualities are achieved by letting the conditional mean of the normalized observation errors be a potentially nonlinear function of the errors in the state equation. Our model differs from Hwu and Kim (2023) both in the formulation of the error distribution and robustness to endogeneity and TVTP.

The rest of the chapter is organized as follows. Section 2.1 outlines the proposed model. Section 2.2 describes how samples from the posterior distribution of model parameters are simulated using Markov Chain Monte Carlo (MCMC) methods. Section 2.3 reports the results of a simulation study. Section 2.4 describes model comparison using Bayes factors. Section 2.5 applies the model to US output growth data. Section 2.6 concludes.

2.1 The Proposed Model

2.1.1 Model Setup

Consider the model

$$y_t = x_t'\beta_{s_t} + \sigma_{s_t}\varepsilon_t, \tag{2.1}$$

$$s_t = 1\{s_t^* > 0\}, \tag{2.2}$$

$$s_t^* = z_t'\delta_{s_{t-1}} + \eta_t, \tag{2.3}$$

$$\varepsilon_t \sim N(g(\eta_t), 1), \;\; \eta_t \sim N(0, 1). \tag{2.4}$$

When $g(\eta_t)$ is linear, $g(\eta_t) = \rho\eta_t$, this model is observationally equivalent to the endogenous

switching model of Kim, Piger, and Startz (2008). This can be seen by rewriting the model as

$$y_t = x_t'\beta_{s_t} + \tilde{\sigma}_{s_t}\tilde{\varepsilon}_t, \tag{2.5}$$

$$\tilde{\sigma}_{s_t} = \sigma_{s_t}\sqrt{1+\rho^2}, \tag{2.6}$$

$$\tilde{\varepsilon}_t = \frac{\varepsilon_t}{\sqrt{1+\rho^2}}, \tag{2.7}$$

$$\begin{bmatrix}\tilde{\varepsilon}_t \\ \eta_t\end{bmatrix} \sim N(0_2, \Omega), \tag{2.8}$$

$$\Omega = \begin{bmatrix}1 & \tilde{\rho} \\ \tilde{\rho} & 1\end{bmatrix}, \tag{2.9}$$

$$\tilde{\rho} = \frac{\rho}{\sqrt{1+\rho^2}}. \tag{2.10}$$

The formula for $\tilde{\rho}$ guarantees that Ω is positive definite.

To estimate the model, $g(\eta_t)$ is approximated nonparametrically. Let $g(\eta_t) \approx \hat{g}(\eta_t) = \sum_{n=1}^{p} \rho_n b_n(\eta_t) = \rho' b_t$, where $\{b_n(\eta_t)\}$ are basis functions. The basis functions are normalized to equal 0 at the origin. Without this normalization, the intercept would not be jointly identified with $\hat{g}(\eta_t)$. This was a natural choice of normalization because it means there is no impact from endogeneity when $\eta_t = 0$, just as in a parametric enodogenous switching model. The two types of approximations that we considered were polynomial series and regression splines. We only report results for regression splines because they consistently performed comparably to or better than series regression. The polynomial series specification worked well when $g(\eta_t)$ was also polynomial, as would be expected, but less so for other types of functions. Note that when a linear spline is used to approximate $g(\eta_t)$, the joint distribution of ε_t and η_t becomes a mixture of disjoint truncated Normal distributions. For any other approximation, the joint distribution is nonstandard.

2.1.2 The Implications of $g(\eta_t)$ for the Marginal Distribution of ε_t

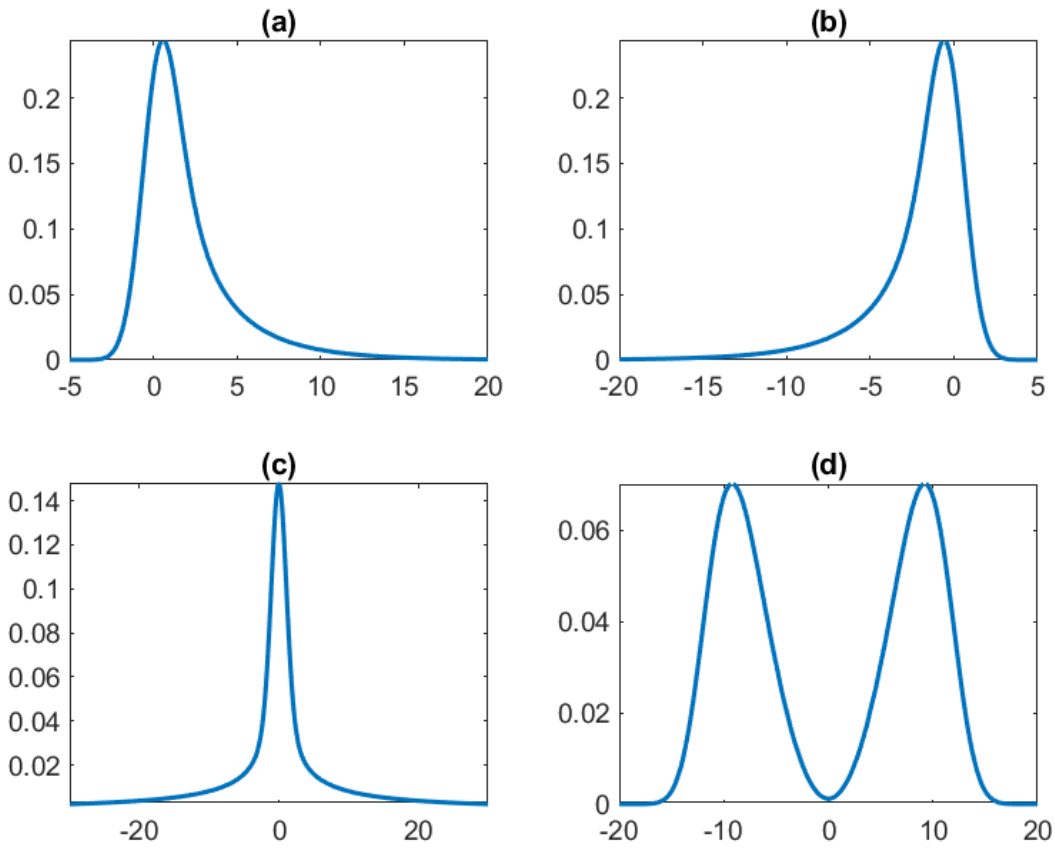
The general form of $g(\eta_t)$ allows for great flexibility in the marginal distribution of ε_t. Figure 2.1 contains plots of $f(\varepsilon_t)$ under various conditional mean functions. It demonstrates that we can induce skewness (1.a and 1.b), excess kurtosis (1.c), and bimodality (1.d) using simple functional forms for $g(\eta_t)$. The reader will observe that neither the unconditional mean nor the unconditional variance are constant with respect to $g(\eta_t)$. This is in contrast to the various types of parametric endogenous switching models (Chib and Dueker, 2004;Kim, Piger, and Startz, 2008; Hwu, Kim, and Piger, 2021). The existing literature models (ε_t, η_t) as a multivariate Normal random variable. This gives the marginal distribution of ε_t the same mean and variance regardless of the correlation structure. In our specification, a greater correlation between ε_t and η_t implies a greater marginal variance of ε_t. We considered marginal moment restrictions on $f(\varepsilon_t)$, namely $E[\varepsilon_t] = 0$ and $Var[\varepsilon_t] = 1$. However, accommodating these restrictions is difficult when using any approximation other than a local polynomial; numerical integration is required to find the mean and variance of $\hat{g}(\eta_t)$. In addition, direct sampling from the full conditional posterior distribution of ρ would no longer be possible. Let $\Theta \equiv \{\beta = \{\beta_j\}, \sigma = \{\sigma_j\}, \delta = \{\delta_j\}, \rho\}$. The dependence between the degree of endogeneity and the marginal variance does not restrict the model overall because $Var[y_t|x_t, s_t, \Theta] = \sigma^2_{s_t} Var[\varepsilon_t]$. $g(\eta_t)$ determines the degree and nature of the endogeneity, while σ controls the variance of the error term. A possible way to weaken this relationship between endogeneity and unconditional variance is to use the alternative model

$$y_t = x_t'\beta_{s_t} + \varepsilon_t, \tag{2.11}$$

$$\varepsilon_t \sim N(g(\eta_t), \sigma^2_{s_t}), \;\; \eta_t \sim N(0,1). \tag{2.12}$$

However, this model offers less flexibility with regard to within-regime variance. Under the alternative model, $Var[y_t|x_t, s_t, \Theta] = Var[g(\eta_t)] + \sigma^2_{s_t}$, as opposed to $\sigma^2_{s_t}(Var[g(\eta_t)] + 1)$ in

the proposed model.



(a) $g(\eta_t) = \eta_t^2$, (b) $g(\eta_t) = -\eta_t^2$, (c) $g(\eta_t) = 10\eta_t^3$, (d) $g(\eta_t) = 10\eta_t^{1/3}$ (the real root).

Figure 2.1: Marginal Distributions of ε_t

2.2 Posterior Sampling

2.2.1 Sampling S_T and S_T^*

Let $Y_t \equiv (y_1, \ldots, y_t)'$, $S_t \equiv (s_1, \ldots, s_t)'$, $S_t^* \equiv (s_1^*, \ldots, s_t^*)'$. Regardless of the functional form of $g(\eta_t)$, the filtered regime probability can be calculated using

$$P(y_t, s_t|s_{t-1}) = \int_{\mathcal{B}_{s_t|s_{t-1}}} f(y_t|s_t, \eta_t) f(\eta_t) d\eta_t, \tag{2.13}$$

$$P(s_t|Y_t) \propto \sum_{s_{t-1}} P(y_t, s_t|s_{t-1}) P(s_{t-1}|Y_{t-1}). \tag{2.14}$$

$B_{s_t|s_{t-1}}$ is the region of integration where the values of η_t are consistent with s_t and s_{t-1}. The constant of proportionality can be obtained by summation over s_t. This enables a straightforward implementation of the algorithm of Chib (1996) for sampling the entire history of regimes as a single block. Numerical integration is performed using the trapezoid method and a fine grid of 500 points. Gauss Legendre quadrature would typically be a superior choice to the trapezoid method as it allows for exact integration of finite order polynomials and requires fewer function evaluations. The trapezoid method was chosen because function evaluations can be saved and reused in sampling S_T^*.

S_T and S_T^* are sampled as a single block by first sampling S_T marginally of S_T^* and then drawing from $\pi(S_T^*|S_T, \Theta, Y_T)$. Conditional on s_t, s_t^* can be drawn independently from the posterior $\pi(S_t^*|S_t, \Theta, Y_T)$ using a Metropolis Hastings step. We obtain near iid samples from the full conditional posterior using a Griddy Gibbs proposal density (Tierney, 1994). The posterior is first discretized by evaluating $f(y_t, S_t^*|s_t, \Theta)$ over an evenly-spaced grid. The discrete probability measure is calculated as

$$P(x_i) = \frac{f(y_t, x_i|s_t, \Theta)}{\sum_k f(y_t, x_k|s_t, \Theta)}. \tag{2.15}$$

A candidate $s_t^{*\prime}$ is obtained by drawing x_i from the discrete distribution and then adding a continuous random variable:

$$s_t^{*\prime} = x_i + u, \quad u \sim N(0, \sigma_u^2). \tag{2.16}$$

The proposal density for $s_t^{*\prime}$ is then obtained by summation of $(s_t^{*\prime}, x_i)$ over the discrete component:

$$q(s_t^{*\prime}) = \sum_i P(x_i) f_N(s_t^{*\prime} - x_i, 0, \sigma_u^2). \tag{2.17}$$

Proposed draws are then accepted with the usual MH acceptance probability.

We follow the common practice in Bayesian MSMs of rejecting samples where s_t is constant over all periods. Accepting such draws can cause the sampler to get stuck in a particular region of the parameter space and mix very slowly. Chib (1996) pointed out that this restriction is not necessary if all priors are proper. Another quirk of Bayesian MSMs of which we must be mindful is label switching (Fruhwirth-Schnatter, 2001). This problem arises in parameter simulation because an unconstrained model with N regimes produces a likelihood with $N!$ modes. Failing to account for label switching can lead to nonsensical parameter estimates if one simply uses the sample mean. One solution is to use identifying restrictions, such as order restrictions on the intercepts or variances.

As η_1 depends on s_0, it must either be sampled or the dependence of η_1 on s_0 must be integrated out. We elect to sample s_0. Since there is no corresponding y_0 for s_0, it can be sampled analytically from its full conditional distribution. Let $\eta(s_0) \equiv s_1^* - z_1' \delta_{s_0}$. The full conditional distribution of s_0 can then be written as

$$P(s_0|Y_T, S_{-0}) \propto f(y_1|s_1, \eta(s_0)) f(s_1^*|s_0) \pi(s_0) \tag{2.18}$$

The constant of proportionality is obtained by summing over all values of s_0. $\pi(s_0)$, the unconditional probability of s_0, can be estimated in several ways. One common approach is to use the stationary distribution of the Markov chain (Albert and Chib, 1993; Chib,

1996). This becomes more complicated when transition probabilities are non-constant. If the variables in z_t are stationary, stationary transition probabilities can approximated by plugging the sample mean of z_t into the equation for s_t^* (Hwu, Kim, and Piger, 2021). However, the approximation is invalid when nonstationary variables like time trends are included. Another solution is to let $\pi(s_0 = 1)$ be a parameter with prior distribution $\pi(s_0 = 1) \sim \mathcal{B}(p_1, p_2)$. One can then sample from the full conditional distribution

$$\pi(s_0 = 1)|Y_T, S_T \sim \mathcal{B}(p_1 + 1 - s_0, p_2 + s_0). \tag{2.19}$$

We use this specification in all estimations that follow.

2.2.2 Sampling β and ρ

Once we condition on S_T, S_T^*, and δ, the model for Y_T becomes linear. We assume the conjugate priors

$$\beta \sim N(b_0, B_0), \tag{2.20}$$
$$\rho \sim N(r_0, R_0). \tag{2.21}$$

We use the hierarchical prior

$$R_0 = \tau_\rho^2 diag(\nu_1, \ldots, \nu_p), \tag{2.22}$$
$$\tau_\rho^2 \sim IG(\alpha_\rho/2, \gamma_\rho/2). \tag{2.23}$$

τ_ρ thus acts as a global smoothness parameter. It can be sampled from the full conditional posterior

$$\tau_\rho^2 \sim IG\left(\frac{\alpha_\rho + p}{2}, \frac{\gamma_\rho + \rho'(diag(\nu_1, \ldots, \nu_p))^{-1}\rho}{2}\right). \tag{2.24}$$

We set $\nu_1, \nu_2, \nu_p = 1$. For other entries, we set $\nu_i = k_i - k_{i-1}$. k_i is the i^{th} knot. Knots are set such that they are evenly spaced across standard Normal quantiles for linear splines and multiples of quantiles for higher order splines.

An equivalent way of writing 2.1 is

$$y_t = \begin{bmatrix} x_t' & s_t x_t' & \sigma_{s_t} b_t' \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 - \beta_0 \\ \rho \end{bmatrix} + \sigma_{s_t}\varepsilon_t^\dagger = x_{s_t}'\beta^* + \sigma_{s_t}\varepsilon_t^\dagger, \tag{2.25}$$

$$\varepsilon_t^\dagger \sim N(0,1). \tag{2.26}$$

Let $X_{S_T} \equiv (x_{s_1}, \ldots, x_{s_T})'$ and $\Sigma_{S_T} \equiv \text{diag}(\sigma_{s_1}^2, \ldots, \sigma_{s_T}^2)$. The likelihood can then be written as

$$f(Y_T|\Theta, S_T, S_T^*) = f_N(Y_T|X_{S_T}\beta^*, \Sigma_{S_T}). \tag{2.27}$$

We then arrive at the full conditional posterior for a classical linear model with heteroskedasticity.

$$\beta^*|Y_T, \Theta_{-\beta^*}, S_T, S_T^* \sim N(\hat{b}^*, \hat{B}^*), \tag{2.28}$$

$$\hat{B}^* = (B_0^{*-1} + X_{S_T}'\Sigma_{S_T}^{-1}X_{S_T})^{-1}, \tag{2.29}$$

$$\hat{b}^* = \hat{B}^*(B_0^{*-1}b_0^* + X_{S_T}'\Sigma_{S_T}^{-1}y), \tag{2.30}$$

$$b_0^* \equiv (b_0', r_0')', \tag{2.31}$$

$$B_0^* \equiv \begin{bmatrix} B_0 & 0_{2k\times p} \\ 0_{p\times 2k} & R_0 \end{bmatrix}. \tag{2.32}$$

When β is unrestricted, the entire vector β^* can be sampled at once. When an identifying restriction is placed on the intercepts, we will use the normalization $\beta_{11} > \beta_{01}$, where β_{j1} is the intercept for $s_t = j$. This leads to the full conditional posterior

$$\beta^*|Y_T, \Theta_{-\beta^*}, S_T, S_T^* \sim TN_{\beta_{k+1>0}^*}(\hat{b}^*, \hat{B}^*). \tag{2.33}$$

Since only one dimension of β^* is truncated, the marginal distribution of β_{k+1}^* is

$$\beta_{k+1}^*|Y_T, \Theta_{-\beta^*}, S_T, S_T^* \sim TN_{(0,\infty)}(\beta_{k+1}^*|\hat{b}_{k+1}^*, \hat{B}_{k+1,k+1}^*). \tag{2.34}$$

A well known result is that is the conditional distributions from a multivariate truncated Normal distribution are also truncated Normal distributions. This fact, combined with the lack of truncation for β_{-k+1}^*, tells us that $f(\beta_{-k+1}^*|\beta_{k+1}^*, Y_T, \Theta_{-\beta^*}, S_T, S_T^*)$ is a multivariate Normal density. We can then sample from $f(\beta^*|Y_T, \Theta_{-\beta^*}, S_T, S_T^*)$ by first sampling from $f(\beta_{k+1}^*|Y_T, \Theta_{-\beta^*}, S_T, S_T^*)$ and then from $f(\beta_{-k+1}^*|\beta_{k+1}^*, Y_T, \Theta_{-\beta^*}, S_T, S_T^*)$.

2.2.3 Sampling σ

Sampling σ_j is complicated by the nonstandard manner in which it enters the likelihood. The full conditional distribution takes the form

$$f(\sigma_j|Y_T, \Theta_{-\sigma_j}, S_T, S_T^*) \propto \pi(\sigma_j) \prod_{s_t=j} f_N(y_t|x_t'\beta_j + \sigma_j\rho' b_t, \sigma_j^2). \tag{2.35}$$

If just σ_j entered the conditional mean parameter of the likelihood, we could use a Normal

prior for σ_j and sample it from a Normal full conditional posterior distribution. If just σ_j^2 entered the conditional variance parameter of the likelihood, we could use an Inverse-Gamma prior for σ_j^2 and sample it from an Inverse-Gamma full conditional posterior distribution. However, the appearance of both σ_j in the conditional mean and σ_j^2 in the conditional variance makes iid sampling from the full conditional posterior infeasible. Luckily, MH sampling with a tailored proposal density is a simple task. The mode of the full conditional distribution of σ_j has an analytical solution for several choices of prior distribution, including Gamma, Inverse-Gamma, and Generalized inverse Gaussian distributions. One first samples from $q(\sigma_j') = f_{t_\nu}(\sigma_j'|\hat{\sigma}_j, c\hat{V}_j)$. $\hat{\sigma}_j$ is the mode of the full conditional posterior. $\hat{V}_j$ is the negative inverse of the second derivative of the log posterior distribution evaluated at $\hat{\sigma}_j$. ν and c are positive tuning parameters. σ_j' is then accepted with probability

$$\alpha = min\left\{1, \frac{q(\sigma_j)\pi(\sigma_j')\prod_{s_t=j} f_N(y_t|x_t'\beta_j + \sigma_j'\rho' b_t, \sigma_j'^2)}{q(\sigma_j')\pi(\sigma_j)\prod_{s_t=j} f_N(y_t|x_t'\beta_j + \sigma_j\rho' b_t, \sigma_j^2)}\right\}. \tag{2.36}$$

2.2.4 Sampling δ

δ_j enters the likelihood in a highly nonlinear fashion via the vector of basis functions. This removes the option of analytical sampling that is present in exogenous and parametric endogenous models. As well, there is no general closed form solution for the mode of the full conditional posterior density. This leaves one with MH sampling and either a tailored proposal distribution that is found numerically or a random walk proposal distribution. A random walk proposal distribution is used in all estimations that follow. Let the prior distribution for $\pi(\delta_j) = f_N(d_{0j}, D_{0j})$. A candidate δ_j' is draw from $q(\delta_j'|\delta_j) = f_N(\delta_j'|\delta_j, \tau_{\delta_j}^2)$. Let $\tilde{\eta}_t \equiv s_t^* - z_t'\delta_j'$. δ_j' is then accepted with probability

$$\alpha = min\left\{1, \frac{\pi(\delta_j')\prod_{s_{t-1}=j} f(y_t|s_t, \tilde{\eta}_t)f(\tilde{\eta}_t)}{\pi(\delta_j)\prod_{s_{t-1}=j} f(y_t|s_t, \eta_t)f(\eta_t)}\right\}. \tag{2.37}$$

Bayesian MSMs sometimes require a strong prior for the transition probabilities for the model to be well-identified. In a model with 2 regimes and fixed transition probabilities, we can select priors to match our expectations about the average duration of a regime (Chib, 1996; Filardo and Gordon, 1998).

2.3 Simulation Study

This section presents the results of a simulation study. We generated 500 datasets with $T = 500$ observations and $k = 3$ variables: an intercept and two variables each drawn from $N(0_T, I_T)$ distributions. A sample of 13,000 draws from the posterior distribution of parameters was obtained for each dataset. Given the earlier discussion of mixing, this may seem like an insufficiently small sample size. However, we observed much faster mixing of the posterior distribution for the simulated datasets than with the output growth dataset used later. We used the identifying restriction $\beta_{01} < \beta_{11}$. In each instance, the first 3,000 draws were discarded as burn-in. For comparison, we also estimated models with parametric endogenous switching and exogenous switching. All datasets were simulated using the function $g(\eta_t) = \eta_t^2$. Table 2.1 reports the parameter estimation errors for all three models. The estimation error of a parameter is taken to be the parameter estimate minus the true parameter value. The results demonstrate that the existing models can produce biased estimates when $g(\eta_t)$ is nonlinear. A surprising result is that the exogenous model outperforms the parametric endogenous model. The large estimation errors of the parametric endogenous model are partly due to bimodality in the empirical error distribution. A natural cubic spline with

	Nonparametric Endogenous Model	Parametric Endogenous Model	Exogenous Model
β_{01}	0.0316	-1.3478	0.2404
	(0.0512)	(1.4932)	(0.0267)
β_{11}	0.1071	2.0731	0.5028
	(0.3088)	(1.5123)	(0.0634)
β_{02}	0.0033	1.3325	-0.0035
	(0.0869)	(1.2015)	(0.0281)
β_{12}	-0.0035	-1.3373	-0.0185
	(0.0975)	(1.1851)	(0.0570)
β_{03}	-0.0076	-1.3320	-0.0007
	(0.0867)	(1.1978)	(0.0271)
β_{13}	0.0066	1.3262	0.0225
	(0.0969)	(1.1751)	(0.0552)
σ_0	0.0194	0.7062	0.3222
	(0.0373)	(0.5127)	(0.0673)
σ_1	0.0089	0.9562	0.1888
	(0.0359)	(0.5127)	(0.0397)
δ_{01}	0.0142	0.7404	0.0211
	(0.1075)	(0.6413)	(0.1239)
δ_{11}	-0.0094	-0.6912	-0.0128
	(0.1074)	(0.6643)	(0.1252)

The values presented are the estimation errors (parameter estimate - true value) for all parameters in θ. Standard deviations are listed below estimation errors in parentheses.

Table 2.1: Average Estimation Errors

9 knots was used to approximate $g(\eta_t)$. Figure 2.2 shows that $g(\eta_t)$ is well-approximated by posterior estimates.

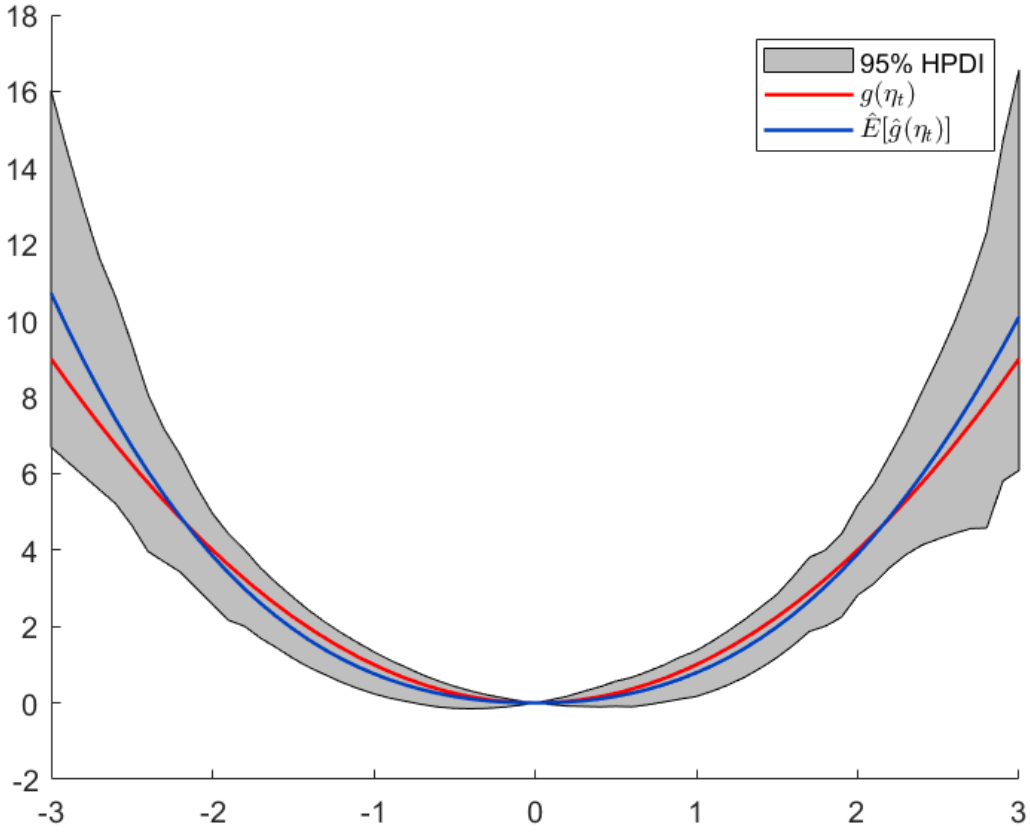


Figure 2.2: The Distribution of $\hat{g}(\eta_t)$, Simulation Study

2.4 Model Comparison

The different models considered in this chapter are compared using Bayes factors (Kass and Raftery, 1995). Since the sampler uses a mix of Gibbs and MH steps, marginal likelihood calculation is done using methods form Chib (1995), Chib (1998), and Chib and Jeliazkov (2001). Bayes factors have also been employed to select the number of regimes in both classical MSMs (Koop and Potter, 1999) and in models with endogenous switching (Kang, 2014). As in Chib (1995), the formula for the marginal likelihood is obtained from a simple application of Bayes' Formula:

$$f(Y_T|\mathcal{M}_i) = \frac{f(Y_T|\Theta^*, \mathcal{M}_i)\pi(\Theta^*|\mathcal{M}_i)}{f(\Theta^*|Y_T, \mathcal{M}_i)}. \tag{2.38}$$

Θ^* is taken to be the posterior mean of Θ. The likelihood $f(Y_T|\Theta^*, \mathcal{M}_i)$ is calculated using a modified version of the forward filtering algorithm of Hamilton (1989). $\pi(s_0 = 1)$ can be integrated out of the likelihood by replacing it with its prior mean. $f(\Theta^*|Y_T, \mathcal{M}_i)$ is rewritten as $f(\beta^*|\sigma^*, \delta^*, Y_T, \mathcal{M}_i)f(\sigma^*|\delta^*, Y_T, \mathcal{M}_i)f(\delta^*|Y_T, \mathcal{M}_i)$. All ordinates are estimated via simulation as in Chib and Jeliazkov (2001).

2.5 Application to GDP Data

We applied the model of (2.1) - (2.4) to data on quarterly real GDP growth. The dataset runs from 1947:Q2 to 2019:Q4. The most recent recession was omitted because the magnitudes of the shifts are much greater than in the rest of the sample. Estimations that included this period did not perform well at identifying previous recessions. They tended to classify every period prior to 2020:Q1 as an expansion. Let the dependent variable be defined as $y_t \equiv \ln(GDP_t) - \ln(GDP_{t-1})$. We estimated a 2-state switching means model with constant scaling factor σ:

$$y_t = \beta_{s_t} + \sigma\varepsilon_t. \tag{2.39}$$

We experimented with different autoregressive specifications, allowing for up to 4 lags of y_t and switching scaling factors. However, the simple switching means model with constant scaling factor performed the best in identifying latent states that correspond to business cycles. $g(\eta_t)$ is approximated using a natural cubic spline with 9 knots. Identification is achieved through the restriction $\beta_0 < \beta_1$. We also estimated a parametric endogenous model

and an exogenous model. We used the prior $\pi(\sigma) = f_{TN_{(0,\infty)}}(\sigma|0,1)$ and the hyperpriors $b_0 = (-.1815, .4196, 0_p')'$, $B_0 = .25I_2$, $\alpha_\rho = 1$, $\gamma_\rho = .1$, $d_0 = (-.6, 1.66)'$, $D_0 = I_2$, $p_1, p_2 = 0$. We set $p = 1$ and $p = 0$ for the parametric endogenous and exogenous models, respectively. b_0 was chosen to match the average growth rates during recessions and expansions as classified by the National Bureau of Economic Research (NBER). d_0 was chosen to match the average durations of recessions and expansions. Each sample of parameters consisted of 300,000 draws after burn-in samples were discarded.

As can be seen in Figure 2.3, the posterior estimate of $\hat{g}(\eta_t)$ is rather nonlinear. To better understand how nonlinearity in the conditional mean of ε_t affects its marginal distribution, we estimated the densities $f(\varepsilon_\tau|Y_T)$ and $\{f(\varepsilon_\tau|Y_T, s_\tau, s_{\tau-1})\}$. The subscript τ is used instead of t to stress that these distributions are not conditioned on any time period in the sample. We would ideally remove the dependence of ε_τ on ρ and δ through direct integration:

$$f(\varepsilon_\tau|Y_T) = \int f(\varepsilon_\tau|\eta_\tau, Y_T, \rho) f(\eta_\tau) f(\rho|Y_T) d\eta_\tau d\rho, \tag{2.40}$$

$$f(\varepsilon_\tau|Y_T, s_\tau, s_{\tau-1}) = \int f(\varepsilon_\tau|\eta_\tau, Y_T, \rho) f(\eta_\tau|Y_T, s_\tau, s_{\tau-1}, \delta) f(\rho, \delta|Y_T) d\eta_\tau d\rho d\delta. \tag{2.41}$$

The intractability of these integrals forces us to instead use a mixture of numerical and monte carlo integration. At each iteration m of the MCMC sampler, we evaluate the integrals $\int f(\varepsilon_\tau|\eta_\tau, Y_T, \rho^{(m)}) f(\eta_\tau) d\eta_\tau$ and $\int f(\varepsilon_\tau|\eta_\tau, Y_T, \rho^{(m)}) f(\eta_\tau|Y_T, s_\tau, s_{\tau-1}, \delta^{(m)}) d\eta_\tau$ using numerical methods. The reader should note that $f(\eta_\tau) = f_N(\eta_\tau|0,1)$ and $f(\eta_\tau|Y_T, s_\tau, s_{\tau-1}, \delta^{(m)})$ is a truncated standard Normal density with region of truncation $\mathcal{B}^{(m)}_{s_t|s_{t-1}}$. The rest of the integration is done by averaging over MCMC draws. We use the approximations

$$f(\varepsilon_\tau|Y_T) \approx M^{-1} \sum_{m=1}^{M} \int f(\varepsilon_\tau|\eta_\tau, Y_T, \rho^{(m)}) f(\eta_\tau) d\eta_\tau, \tag{2.42}$$

$$f(\varepsilon_\tau|Y_T, s_\tau, s_{\tau-1}) \approx M^{-1} \sum_{m=1}^{M} \int f(\varepsilon_\tau|\eta_\tau, Y_T, \rho^{(m)}) f(\eta_\tau|Y_T, s_\tau, s_{\tau-1}, \delta^{(m)}) d\eta_\tau. \tag{2.43}$$

M is the number of remaining MCMC draws after burn-in samples are discarded. Approximations $\hat{f}(\varepsilon_\tau|Y_T)$ and $\hat{f}(\varepsilon_\tau|Y_T, s_\tau, s_{\tau-1})$ are plotted in Figures 2.4 and 2.5, respectively. $\hat{f}(\varepsilon_\tau|Y_T)$ is skewed to the right, making extreme positive values more likely than in a Gaussian distribution. We observe interesting deviations from $\hat{f}(\varepsilon_\tau|Y_T)$ when we condition on past and current regimes. $\hat{f}(\varepsilon_\tau|Y_T, s_\tau = 0, s_{\tau-1} = 0)$ is positively skewed and centered around a positive number, meaning we are more likely to see positive deviations from the average growth rate during a recession. The distribution of errors in $\hat{f}(\varepsilon_\tau|Y_T, s_\tau = 0, s_{\tau-1} = 1)$ is more Gaussian, but there is more mass in the positive region of ε_τ. This implies that average growth is higher in the first period of a recession. We can interpret this as a transitional period between high and low growth. We also see a large amount of probability mass in the positive region of ε_τ for $\hat{f}(\varepsilon_\tau|Y_T, s_\tau = 0, s_{\tau-1} = 1)$. This corresponds to a high growth recovery in which average growth is higher in the first quarter following a recession. $\hat{f}(\varepsilon_\tau|Y_T, s_\tau = 1, s_{\tau-1} = 1)$ is the density that most closely resembles a Gaussian distribution centered at 0. This results from the quasilinear shape of $\hat{E}[\hat{g}(\eta_t)|Y_T]$ in the region [-1,.5].

Parameter estimates are displayed in Table 2.2. Estimates for β are lower for the nonparametric model than the other two. This is likely caused by the positive skew in $f(\varepsilon_\tau|Y_T)$. The posterior estimate for σ is also lowest in the nonparametric model, indicating that there is less residual variation in the data when we allow ε_t to have a nonlinear conditional mean function. The estimate for δ_0 is highest in the nonparametric model, corresponding to less persistent recessions. There does not appear to be a large variation in the persistence of expansions predicted by the three models.

Figure 2.6 shows smoothed recession probabilities from the nonparametric model along with NBER recession dates. We observe a spike in recession probabilities during every recession.

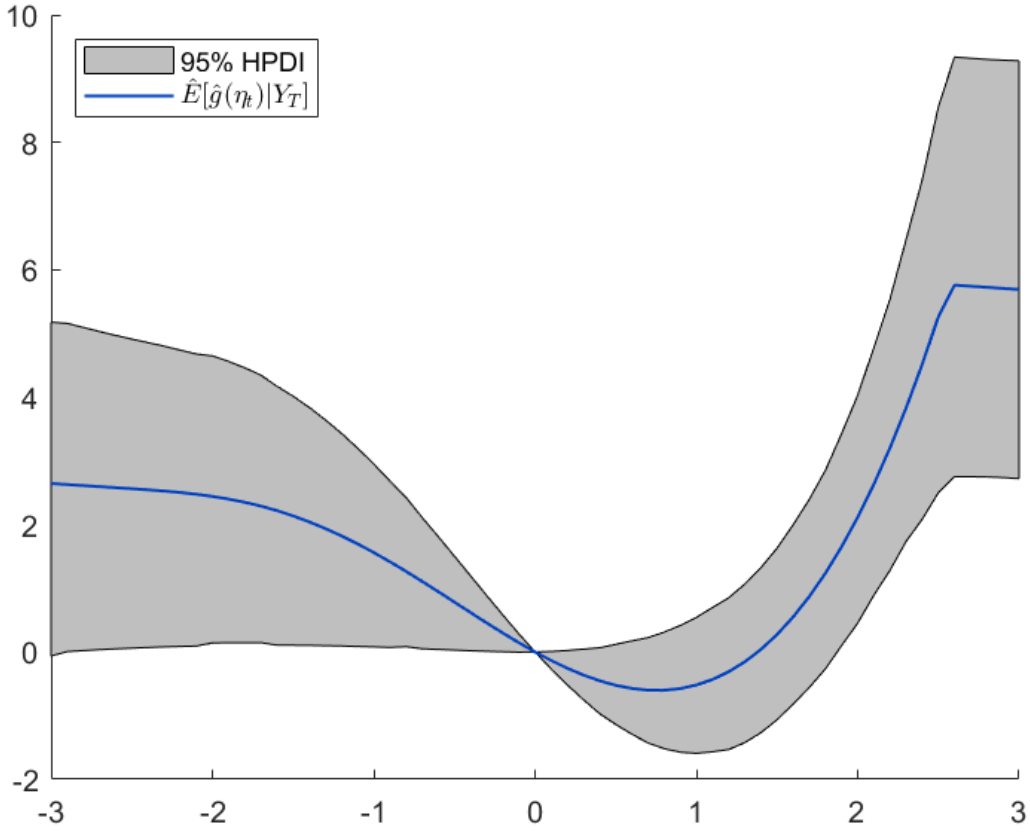


Figure 2.3: The Distribution of $\hat{g}(\eta_t)$, GDP Model

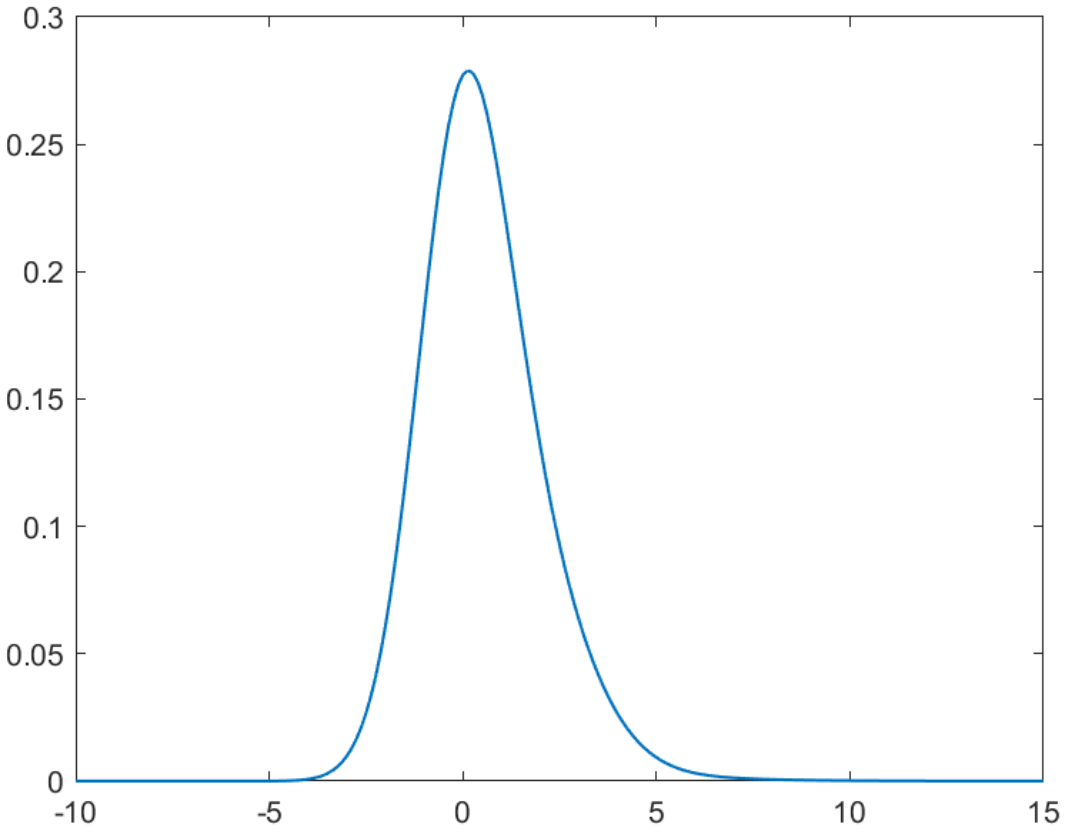


Figure 2.4: $\hat{f}(\varepsilon_\tau|Y_T)$

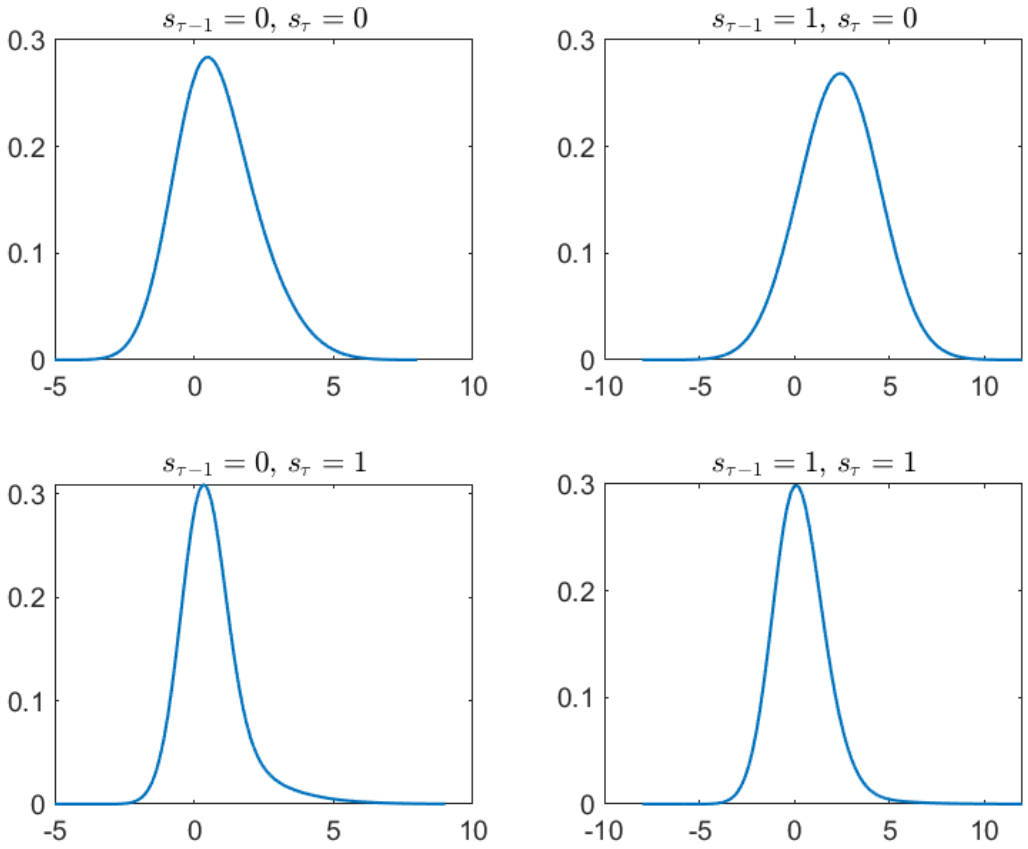


Figure 2.5: $\hat{f}(\varepsilon_\tau|Y_T, s_\tau, s_{\tau-1})$

	Nonparametric Endogenous Model	Parametric Endogenous Model	Exogenous Model
β_0	-0.4089	-0.2368	-0.0049
	(0.1298)	(0.2593)	(0.2080)
β_1	0.3702	0.4500	0.5357
	(0.0642)	(0.1317)	(0.1624)
σ	0.2430	0.3475	0.3497
	(0.0362)	(0.0186)	(0.0178)
δ_0	-0.3880	-0.5434	-0.90901
	(0.2429)	(0.5533)	(0.5449)
δ_1	1.3523	1.6845	1.2814
	(0.2625)	(0.4062)	(0.4811)
$\ln(f(Y_T\|\Theta^*, \mathcal{M}_i))$	-123.4621	-137.8244	-140.4360
$\ln(f(Y_T\|\mathcal{M}_i))$	-139.3520	-143.3834	-144.9914

Standard deviations are listed below parameter estimates in parentheses.

Table 2.2: Posterior Estimates for Output data

The one false positive occurs in the first quarter of the sample. This is a reasonable error for the model to produce because real output growth was negative in this period.

The natural logarithm of the Bayes factors for choosing the Nonparametric model over the exogenous model and the parametric endogenous model are 5.64 and 4.03, respectively. Using an uninformative uniform prior for model probabilities, this makes the posterior probability of the nonparametric model 281.46 times that of the exogenous model and 56.26 times that of the parametric endogenous model.

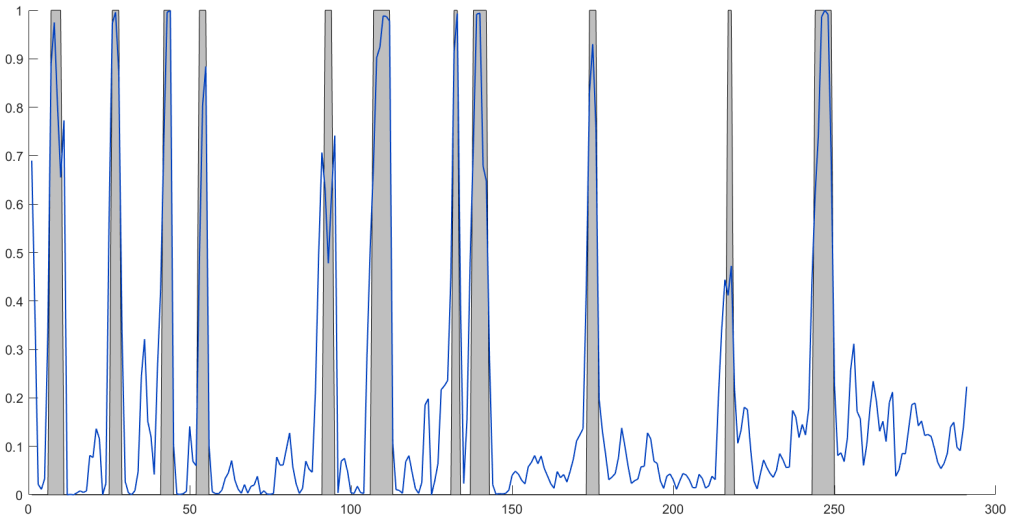


Figure 2.6: Smoothed Recession Probabilities

2.6 Conclusion

We developed a MSM with nonparametric endogenous switching. The nonparametric model offers substantial flexibility with regard to the marginal distribution of observation errors. A simulation study demonstrated that existing parametric models can produce biased results when the true data generating process entails nonlinear endogenous switching. The model was applied to data on US real GDP growth. The estimated model had a significantly higher Bayes factor than estimates for parametric endogenous and exogenous models. The nonparametric model is also able to identify all recessions, as calculated by the NBER. Estimated marginal error distributions indicated the innovations in the observation equation are non-Gaussian and generally skewed to the right.

There are many areas in which this chapter could be extended. One obvious application is to financial data. A flexible error distribution is called for in a landscape where fat tails and skewness are expected. Other directions for further research are extensions to multivariate data and more than 2 regimes.

Chapter 3

A Flexible Conditional Mean Function for Count Data Analysis

This paper proposes a new flexible conditional mean function for use in count data modelling. Count data, primarily modelled with Poisson regression or compound distributions like the Negative Binomial and Poisson-Log Normal, arise in many areas of economic analysis (Blundell, Griffith, and Reenen, 1995; Schoenmaker, 1996; Cameron et al., 1988) as well as other disciplines such as epidemiology (Zeger, 1988; Davis, Dunsmuir, and Wang, 2000; Böhning et al., 1999). Count data models with exogenous covariates have relied almost exclusively on the exponential conditional mean function. Despite several glaring issues with the exponential function - explosive covariate effects, sensitivity to model misspecification, and inability to incorporate lagged dependent variables - it has remained the conditional mean function of choice primarily due to its simplicity. The proposed conditional mean function attempts to solve these problems.

The other strand of the literature to which this paper contributes is models of autocorrelated count data. An early model for time series count data was the Integer-Valued ARMA

(INARMA) model (Al-Osh and Alzaid, 1987; Brännäs and Hall, 2001). In the most basic INAR(1) model, the dependent variable is the sum of a binomial thinning operator applied to last period's dependent variable and a Poisson-distributed error term. The main advantage of using the proposed conditional mean function in a time series model is that it allows all covariates, be they exogenous, or lagged dependent variables and conditional means, to impact the conditional mean in the same manner. Other models tend to specify one process for the autoregressive component of the conditional mean and then multiply that term by an exponential function of exogenous covariates. Zeger (1988) modeled autocorrelation in count data with a stochastic autoregressive mean (SAM), μ_t in a Poisson regression model. He incorporated exogenous covariates by letting $\mathbb{E}[y_t|x_t] = e^{x_t'\beta}\mu_t$. While this model is intuitive, estimation requires either inversion of a high-dimensional covariance matrix or calculation of high-dimensional integrals (Jung, Kukuk, and Liesenfeld, 2006). Heinen (2003) introduced the Autoregressive Conditional Poisson (ACP) model. This model lets μ_t be a linear function of lagged dependent variables and conditional means. Exogenous covariates are introduced in the same way as Zeger (1988). While this model provides a simple method for forecasting, it does not permit negative autocorrelation. The GLARMA model of Davis, Dunsmuir, and Streett (2003) offers another observational approach. They model $\ln \mu_t \equiv \omega_t$ as a linear function of past $\omega_t s$ and error terms that are standardized by $\mu_t^\rho, \rho \in (0, 1]$. Cameron and Trivedi (1998) propose a model in which the observed outcome is the result of an autoregressive conditional ordered probit model (ACOP). The latent variable that determines the outcome is a linear function of lagged dependent variables, lagged values of another latent variable, exogenous covariates, and an error term that follows a standard Normal distribution. This model offers greater flexibility in some respects, but it cannot accommodate a wide range of outcomes.

The rest of this paper is organized as follows. Section 3.1 explains the problems with the exponential conditional mean function. It then proposes a new conditional mean function that solves these problems. Section 3.2 proposes a new model for autocorrelated count data

that incorporates the new conditional mean function. Section 3.3 presents our first empirical application of the new conditional mean function to data examining how daylight savings time impacts the number of fatal car accidents. Section 3.4 presents our second empirical application that evaluates if regulatory policy limiting the number of flights was effective at reducing flight delays at Newark International Airport. Finally, Section 3.5 concludes.

3.1 A New Conditional Mean Function

3.1.1 The Exponential Conditional Mean Function

The traditional conditional mean function used for cross-sectional count data models is the exponential. The conditional mean is thus given by

$$\mathbb{E}[y_i|x_i'\beta] = e^{x_i'\beta} \tag{3.1}$$

The main advantage of the exponential conditional mean function is the simplicity with which it maps $\mathbb{R} \to \mathbb{R}^+$. This mapping is a necessary quality for a conditional mean function, as any count model requires that the conditional mean be nonnegative. However, there are several disadvantages. The first drawback of the exponential conditional mean function is the covariate effects it produces. This is seen by taking the partial derivative with respect to x_{ij}.

$$\frac{\partial}{\partial x_{ij}} e^{x_i'\beta} = e^{x_i'\beta}\beta_j \tag{3.2}$$

This leads to covariate effects that become larger and larger in magnitude as the conditional mean increases, creating unrealistic predictions in applications to economic data. Many foundational economic models assume diminishing marginal effects. On the contrary, the exponential conditional mean function produces increasing marginal effects. Demand for aggregate consumption can at most be an increasing linear function of wealth while still satisfying a consumer's budget constraint. The increasing covariate effects in standard count models thus lead to infeasible predictions.

Another drawback of the exponential conditional mean function is its lack of robustness to model misspecification. Under this assumption, the likelihood of the data given the model will be very steep. This leads to precise estimates when the specification is correct, but makes the model ill-suited to analyze data where the assumption is not satisfied. This will be illustrated in the simulation study below.

The final disadvantage of the exponential conditional mean function that we will discuss is its inability to accept lagged values of the dependent variable as covariates. This problem is apparent when one looks at the expected growth rate of the conditional mean.

$$\mathbb{E}[y_t|x_t'\beta] = \mu_t = e^{x_t'\beta+\phi y_{t-1}}, \quad x_t \overset{iid}{\sim} F(x). \tag{3.3}$$

$$\begin{aligned} \ln\mu_t - \ln\mu_{t-1} &= x_t'\beta + \phi y_{t-1} - x_{t-1}'\beta - \phi y_{t-2}, \\ \mathbb{E}[\ln\mu_t - \ln\mu_{t-1}] &= \mathbb{E}[x_t'\beta - x_{t-1}'\beta] + \mathbb{E}[\phi y_{t-1} - \phi y_{t-2}], \\ \mathbb{E}[\ln\mu_t - \ln\mu_{t-1}] &= \phi\mathbb{E}[y_{t-1} - y_{t-2}]. \end{aligned} \tag{3.4}$$

Since the expected growth rate of the conditional mean is constant, the conditional mean is either explosive or implosive.

3.1.2 The Proposed Conditional Mean Function

The new conditional mean function proposed here attempts to remedy all of the problems outlined above. Our function is obtained through a two-step mapping, $\lambda(x_i'\beta) = g(f(x_i'\beta))$, where $f(\cdot)$ is a logistic cumulative density function (CDF) and $g(\cdot)$ is the inverse CDF of the Weibull distribution with scale parameter c and shape parameter k. This mapping yields the function

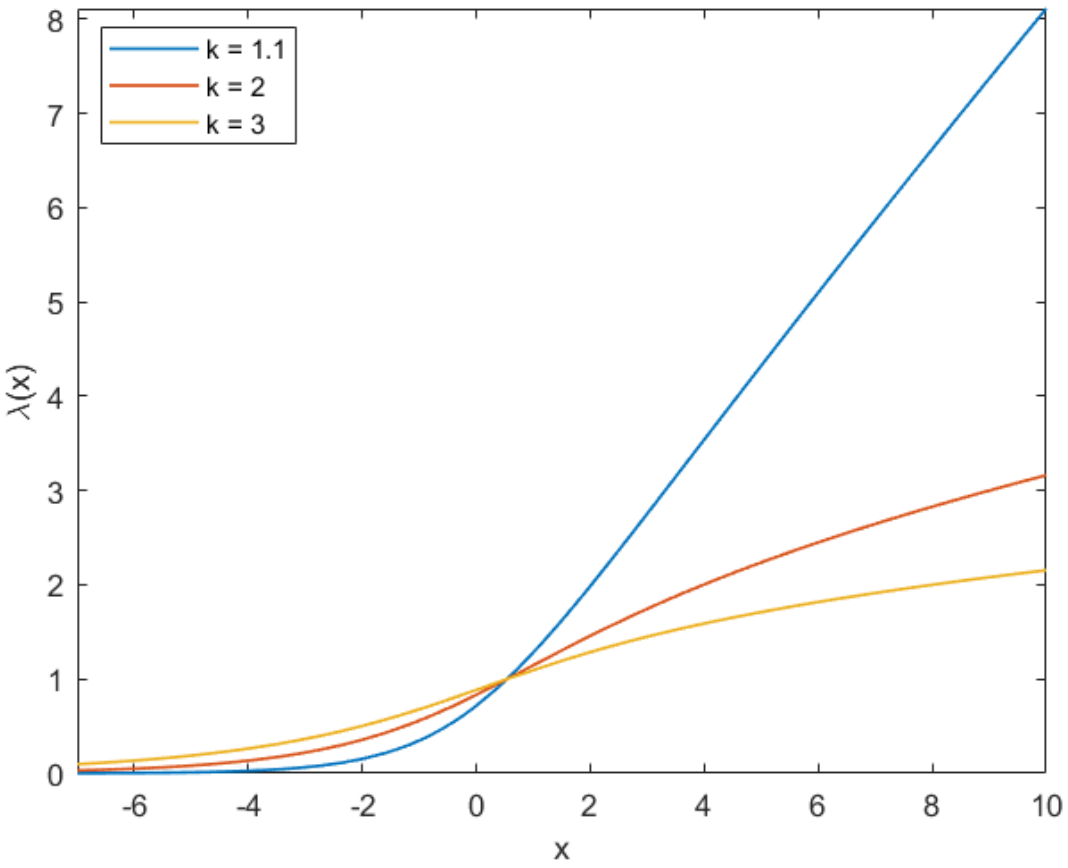
$$\lambda(x_i'\beta) = c\ln(1 + e^{x_i'\beta})^{\frac{1}{k}}, \quad c > 0, k > 1. \tag{3.5}$$

We refer to this function as the Logistic-Weibull (LW) function. If $k = 1$, we obtain a scaled version of the *softplus* activation function used in neural networks (Dugas et al., 2001). However, we restrict k to be greater than 1 to ensure that the function is eventually concave. Figure 3.2 demonstrates the flexibility of the LW function. k controls how quickly the function moves from the quasilinear region to the concave region. c controls the scale of the function as well as the range of $\lambda(\cdot)$ for which the function is approximately exponential. The LW function converges to $ce^{\frac{x_i'\beta}{k}}$ when $x_i'\beta$ is small and $c(x_i'\beta)^{\frac{1}{k}}$ when $x_i'\beta$ is large. To demonstrate convergence to $ce^{x_i'\beta}$, recall that $\ln(1+x) \approx x$ for x sufficiently close to 0. This means that when $x_i'\beta$ is sufficiently negative, making $e^{x_i'\beta}$ sufficiently close to 0,

$$c\ln(1 + e^{x_i'\beta})^{\frac{1}{k}} \approx c(e^{x_i'\beta})^{\frac{1}{k}} = ce^{\frac{x_i'\beta}{k}}.$$

Convergence to $c(x_i'\beta)^{\frac{1}{k}}$ for $x_i'\beta$ sufficiently large can be seen from the fact that $\ln(1+x) \sim \ln(x)$ as $x \to \infty$. This means that when $x_i'\beta$ is sufficiently large,

$$c\ln(1 + e^{x_i'\beta})^{\frac{1}{k}} \approx c\ln(e^{x_i'\beta})^{\frac{1}{k}} = c(x_i'\beta)^{\frac{1}{k}}.$$


k = 1.1
k = 2
k = 3
λ(x)
x
8
7
6
5
4
3
2
1
0
-6
-4
-2
0
2
4
6
8
10

While the arguments made here are asymptotic, Figure 3.1 illustrates that this convergence occurs relatively quickly.

Figure 3.1: Convergence of the LW Function

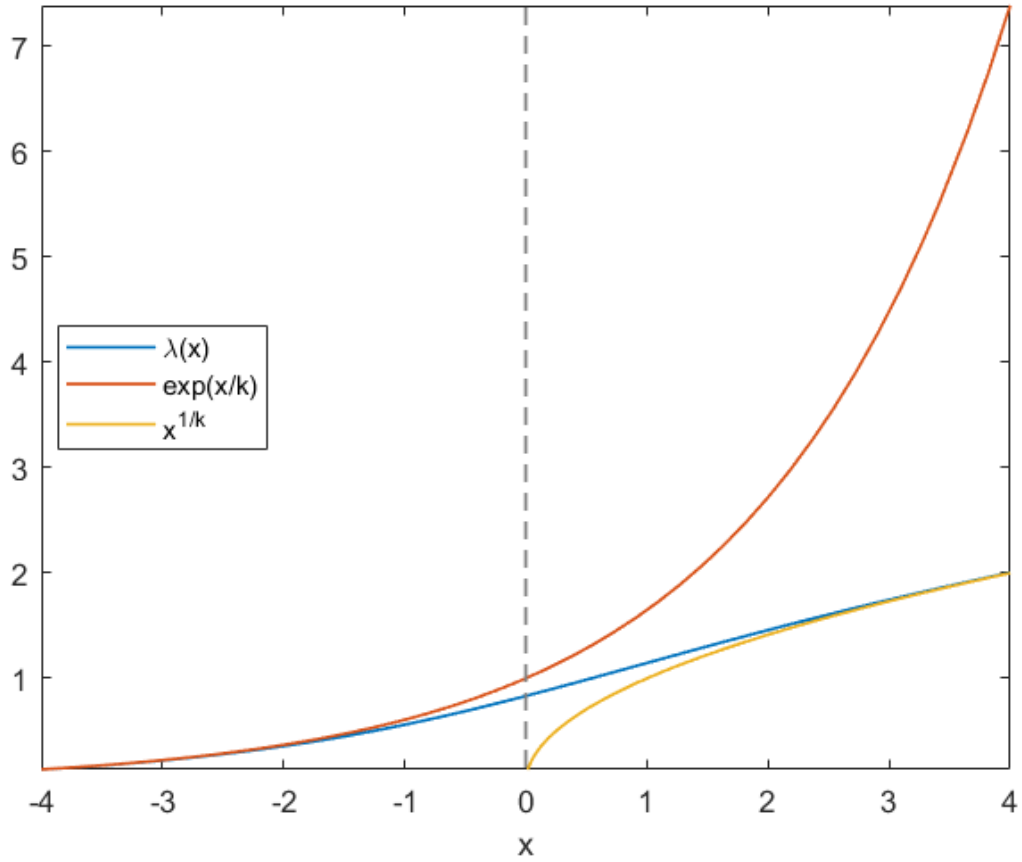


Figure 3.2: The LW Function Under Different Values of k

Covariate Effects

Under this conditional mean function, the marginal effect of a unit change in the covariate x_{ij} is

$$\frac{\partial}{\partial x_{ij}}\lambda(x_i'\beta) = \frac{c}{k}\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\lambda(x_i'\beta)^{\frac{1-k}{k}}\beta_j. \tag{3.6}$$

While this formulation is not immediately intuitive, the term that multiplies β_j has a bell shape that initially increases with $x_i'\beta$ but eventually dies off to 0. In practice, researchers can choose tolerance levels for convergence to the exponential and k^{th} root functions and then identify critical values of $x_i'\beta$ beyond which the conditional mean function can be treated as if it is one of those two. In the k^{th} root region this leads to an approximate covariate effect of $\frac{c}{k}(x_i'\beta)^{\frac{1-k}{k}}\beta_j$. The marginal effect for the exponential region would be $\frac{c}{k}e^{\frac{x_i'\beta}{k}}\beta_j$. As illustrated in Figure 3.1, the nonconvergent region is roughly linear. If we let α be the inflection point for $\lambda(x)$ ($\lambda''(\alpha) = 0$), then we can let the covariate effect be $\lambda'(\alpha)\beta_j$ for values of $x_i'\beta$ around α. It is likely only when $x_i'\beta$ is close to one of the critical values but still in the nonconvergent zone that it would be prudent to use the exact marginal effect.

Model Misspecification

We will now investigate how robust the proposed conditional mean function is to model misspecification. Model misspecification will be assessed through the lens of information theory. We also explore the precision of maximum likelihood estimates when the function is used in Poisson and Negative Binomial regression models.

The units by which we measure the information contained in a single event are given by the negative of the logarithm, with respect to a specific base number, of the probability of the event. Two common choices for the base number of the logarithm are 2 and e. When the base-2 logarithm is used, the units of information are called bits. The quantity of bits associated with a specific event tells us how many times the volume of the probability space of a random variable is divided by 2 when we know the variable takes a specific value. The lower the probability of an event, the greater the information gained by knowing it occurs. For our analysis, we will use the natural logarithm, making the units of information nats.

A common method for evaluating the difference between two probability distributions is

the expected information gain. When data is generated by one of the two distributions, the information gain is the difference in the information an observation gives when we assume the true model and the information from assuming the incorrect model. For a correct distribution A and an incorrect distribution B, it is calculated as

$$I = \sum_i p(y_i|A)(\ln p(y_i|A) - \ln p(y_i|B)). \tag{3.7}$$

This is also known as the Kullback-Leibler Divergence. Computing information gain is complicated in our case because we are not conditioning on a single distribution, but rather the functional form of the distribution. In principle, we could calculate the expected information gain from choosing the true conditional mean function over the false conditional mean function by specifying prior distributions for all model parameters. Let T and F denote the true conditional mean function and the false conditional mean function respectively. We can then compute

$$I = \int_X \int_\theta \sum_i p(y_i|\theta, T, X)[\ln p(y_i|\theta, T, X)p(\theta|T) - \ln p(y_i|\theta, F, X)p(\theta|F)]p(X)d\theta dX. \tag{3.8}$$

We chose not to pursue this approach because it would be computationally infeasible. In particular, the general lack of identification of the scale and shape parameters in the LW function makes it difficult to define appropriate priors. We instead estimated the expected information gain from a given conditional mean function after conditioning on the maximum likelihood estimates. This quantity was estimated via simulation.

We considered four cases for the true Data Generating Process (DGP): a Conditional Poisson Model with an exponential conditional mean function, a Conditional Poisson Model with an LW conditional mean function, a Conditional Negative Binomial Model with an exponential conditional mean function, and a Conditional Negative Binomial Model with an LW

conditional mean function. $c = 10$, $k = 5$ were chosen when the true model used an LW conditional mean function. $B = 1,000$ datasets of size $n = 300$ were simulated from each DGP. We fit models with LW and exponential conditional mean functions for each simulated dataset and saved the log-likelihoods evaluated at the maximum likelihood estimates. Holding beliefs about the DGP for X constant, the simulation estimate of I can be expressed as

$$\hat{I} = \frac{1}{B}\sum_{b=1}^{B} \ln p(y^{(b)}|X^{(b)}, \hat{\theta}_{MLE}, T) - \ln p(y^{(b)}|X^{(b)}, \hat{\theta}_{MLE}, F). \tag{3.9}$$

Table 3.1 reports the estimates of I that arise from the two modelling assumptions. It is interesting to note that the expected information gain is large when the true conditional mean function is LW and slightly negative when the true conditional mean function is exponential. The exponential function is thus ill-suited to problems with concavity in the conditional mean function. On the other hand, the flexibility of the LW function allows it to fit the data generated by an exponential conditional mean function as well as or better than the true model.

True Conditional Mean	Poisson	Negative Binomial
LW	292.280	315.434
Exponential	-0.202	-0.251

Table 3.1: Expected Information Gain

Table 3.2 gives the results of a Monte Carlo study. The values are the sample means and standard deviations (displayed below means in parentheses) of maximum likelihood estimates across $B = 1,000$ datasets of size $n = 300$. Both the DGPs and the estimators hold c and k constant at 10 and 3, respectively. γ is the dispersion parameter in the Negative Binomial model. From simulation results, we see that the maximum likelihood estimator is fairly precise for the Poisson model. The precision of all parameters decreases a bit in the Negative

Binomial model, as would be expected when the conditional distribution of the data has a higher variance.

	True Parameters	Poisson	Negative Binomial
γ	12		12.633
			(2.496)
β_0	1	1.000	1.029
		(0.200)	(0.403)
β_1	2	2.005	2.021
		(0.097)	(0.165)
β_2	3	3.011	3.036
		(0.127)	(0.227)
β_3	4	4.014	4.041
		(0.155)	(0.297)

Table 3.2: Estimation Results for Simulated Data

3.2 A New Time Series Process for Count Data

Let y_t have nonnegative integer values with conditional expectation

$$\mathbb{E}[y_t|\mathcal{F}_{t-1}] = \lambda_t = \lambda(z_t'\delta) = \lambda(x_t'\beta + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \psi_j \lambda_{t-j}), \tag{3.10}$$

where $\mathcal{F}_{t-1}$ is the information set up to time $t-1$, which includes $x_t'\beta$. The following sections will consider processes for which $y_t|\mathcal{F}_{t-1} \sim Poisson(\lambda_t)$ and $y_t|\mathcal{F}_{t-1} \sim NB(\lambda_t, \gamma)$. We will refer to these processes as Poisson ARMA(p,q) and Negative Binomial ARMA(p,q).

3.2.1 Dynamics

Due to the relatively more complicated formulation of $\lambda(x)$ than e^x, direct evaluation of the expected growth rate of the conditional mean is not feasible. However, we can get an idea

of the dynamics in the AR(1) case by writing

$$\mathbb{E}[y_t - y_{t-1}|y_{t-1}] = \lambda(k + \phi y_{t-1}) - y_{t-1}.$$

Since $\lambda(k + \phi y_{t-1}) > 0$ and is either bounded (in the case of $\phi < 0$) or eventually concave, there exists two points, y_l^* and y_h^* (both > 0), such that $\lambda(\beta_0 + \phi y_{t-1}) > y_{t-1}$ for $y_{t-1} < y_l^*$ and $\lambda(k + \phi y_{t-1}) < y_{t-1}$ for $y_{t-1} > y_h^*$. We can then conclude $\mathbb{E}[y_t - y_{t-1}|y_{t-1}]$ is

$$\begin{cases} < 0 & \text{if } y_{t-1} > y_h^* \\ > 0 & \text{if } y_{t-1} < y_l^* \end{cases}.$$

For the general case of an ARMA(p,q) specification with exogenous covariates, we can write

$$\mathbb{E}[y_t - y_{t-1}|\mathcal{F}_{t-1}] = \lambda(z_t'\delta) - y_{t-1}.$$

$\lambda(z_t'\delta) > 0$ and is either bounded (in the case of $\phi_1 < 0$) or eventually concave with respect to y_{t-1} (conditional on all other variables in the information set). Therefore, when all other covariates are held constant, there exists two points, $y_{t-1,l}^*$ and $y_{t-1,h}^*$ (both > 0), such that $\lambda(z_t'\delta) > y_{t-1}$ for $y_{t-1} < y_{t-1,l}^*$ and $\lambda(z_t'\delta) < y_{t-1}$ for $y_{t-1} > y_{t-1,h}^*$. We can then conclude $\mathbb{E}[y_t - y_{t-1}|\mathcal{F}_{t-1}]$ is

$$\begin{cases} < 0 & \text{if } y_{t-1} > y_{t-1,h}^* \\ > 0 & \text{if } y_{t-1} < y_{t-1,l}^* \end{cases}.$$

The AR(1) and ARMA(p,q) processes are thus neither explosive nor implosive. For many combinations of parameters, $y_l^* = y_h^*$ and $y_{t-1,l}^* = y_{t-1,h}^*$. This suggests the existence of mean-reverting dynamics. To conclude that this is indeed mean reversion, we must first demonstrate that an autoregressive process of this sort has an unconditional mean. We

demonstrate this in the next section.

3.2.2 Stationarity Results

This section presents results for the strict stationarity of y_t. While the exact values for the unconditional moments are elusive, we demonstrate their existence by proving that they are bounded. It is common in the literature for time series count models to prove covariance stationarity without respect to the distribution of exogenous covariates (Heinen, 2003; Davis, Dunsmuir, and Streett, 2003). As such, we first present stationarity results for $x_t = 1$. We then show that these results extend easily to moments conditional only on x_t. Finally, we will present unconditional moments, assuming $x_t \overset{iid}{\sim} F(x_t),\ \mathbb{E}[|x_t'\beta|^n] < \infty,\ n \in \mathbb{N}$.

The Conditional Poisson Distribution

Proposition 3.1 For $x_t = 1$:

$$0 < \mathbb{E}[y_t] < \frac{b_1 + m_1|\beta_0|}{1 - m_1(\sum_{i=1}^{p}|\phi_i| + \sum_{j=1}^{q}|\psi_j|)}. \tag{3.11}$$

For $x_t \in \mathbb{R}^{d_x}$:

$$0 < \mathbb{E}[y_t|x_t] < \frac{b_1 + m_1|x_t'\beta|}{1 - m_1(\sum_{i=1}^{p}|\phi_i| + \sum_{j=1}^{q}|\psi_j|)}. \tag{3.12}$$

For $x_t \overset{iid}{\sim} F(x_t)$:

$$0 < \mathbb{E}[y_t] < \frac{b_1 + m_1\mathbb{E}[|x_t'\beta|]}{1 - m_1(\sum_{i=1}^{p}|\phi_i| + \sum_{j=1}^{q}|\psi_j|)}. \tag{3.13}$$

Proof in Appendix. As can be seen in the appendix, the parameters b and m are chosen to ensure the right side of all inequalities are positive and finite and $b+m|x| > \lambda(x)$. Under the conditions $x_t = 1$, $\beta_0 > 0$ and sufficiently large, $\{\phi_i\} > 0,\ \{\psi_j\} > 0$, $\sum_{i=1}^{p} \phi_i + \sum_{j=1}^{q} \psi_j < 1$, we can set $b_1 = 0$, $m_1 = 1$, and the bound becomes the stationary mean of a linear ARMA process with the same parameters.

Proposition 3.2

For $x_t = 1$:

$$0 \leq \mathbb{E}[y_t^n] < \frac{b_n + m_n|\beta_0|^n(p+q+1)^{n-1} + \sum_{l=1}^{n-1} \left\{ {n \atop l} \right\} \mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}(\sum_{i=1}^{p} |\phi_i|^n + \sum_{j=1}^{q} |\psi_j|^n)}. \tag{3.14}$$

For $x_t \in \mathbb{R}^{d_x}$:

$$0 \leq \mathbb{E}[y_t^n|x_t] < \frac{b_n + m_n|x_t'\beta|^n(p+q+1)^{n-1} + \sum_{l=1}^{n-1} \left\{ {n \atop l} \right\} \mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}(\sum_{i=1}^{p} |\phi_i|^n + \sum_{j=1}^{q} |\psi_j|^n)}. \tag{3.15}$$

For $x_t \overset{iid}{\sim} F(x_t)$:

$$0 \leq \mathbb{E}[y_t^n] < \frac{b_n + m_n\mathbb{E}[|x_t'\beta|^n](p+q+1)^{n-1} + \sum_{l=1}^{n-1} \left\{ {n \atop l} \right\} \mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}(\sum_{i=1}^{p} |\phi_i|^n + \sum_{j=1}^{q} |\psi_j|^n)}. \tag{3.16}$$

Proof in Appendix. As in Proposition 3.1, the parameters b_n and m_n are chosen to ensure the right side of all inequalities are positive and finite and $b_n + m_n|x|^n > \lambda(x)^n$. $\left\{ {n \atop l} \right\}$ are Stirling numbers of the second kind.

Proposition 3.3

For $x_t = 1$ or $x_t \overset{iid}{\sim} F(x_t)$:

$$-(\mathbb{E}[y_t])^2 < Cov[y_t, y_{t-s}] < Var[y_t] < \mathbb{E}[y_t^2]. \tag{3.17}$$

For $x_t \in \mathbb{R}^{d_x}$:

$$-(\mathbb{E}[y_t|x_t])^2 < Cov[y_t, y_{t-s}|x_t] < Var[y_t|x_t] < \mathbb{E}[y_t^2|x_t]. \tag{3.18}$$

Proof in Appendix.

The Conditional Negative Binomial Distribution

Proposition 3.4 is identical to Proposition 3.1, as is its proof.

Proposition 3.4

For $x_t = 1$:

$$0 \leq \mathbb{E}[y_t^n] < \frac{\gamma^{-n}\gamma^{(n-1)}(b_n + m_n|\beta_0|^n(p+q+1)^{n-1}) + \sum_{l=1}^{n-1}\left\{{n \atop l}\right\}\gamma^{-l}\gamma^{(l-1)}\mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}\gamma^{-n}\gamma^{(n-1)}(\sum_{i=1}^{p}|\phi_i|^n + \sum_{j=1}^{q}|\psi_j|^n)}. \tag{3.19}$$

For $x_t \in \mathbb{R}^{d_x}$:

$$0 \leq \mathbb{E}[y_t^n|x_t] < \frac{\gamma^{-n}\gamma^{(n-1)}(b_n + m_n|x_t'\beta|^n(p+q+1)^{n-1}) + \sum_{l=1}^{n-1}\left\{{n \atop l}\right\}\gamma^{-l}\gamma^{(l-1)}\mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}\gamma^{-n}\gamma^{(n-1)}(\sum_{i=1}^{p}|\phi_i|^n + \sum_{j=1}^{q}|\psi_j|^n)}. \tag{3.20}$$

For $x_t \overset{iid}{\sim} F(x_t)$:

$$0 \leq \mathbb{E}[y_t^n] < \frac{\gamma^{-n}\gamma^{(n-1)}(b_n + m_n\mathbb{E}[|x_t'\beta|]^n(p+q+1)^{n-1}) + \sum_{l=1}^{n-1}\left\{{n \atop l}\right\}\gamma^{-l}\gamma^{(l-1)}\mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}\gamma^{-n}\gamma^{(n-1)}(\sum_{i=1}^{p}|\phi_i|^n + \sum_{j=1}^{q}|\psi_j|^n)}. \quad (3.21)$$

Proof in Appendix. $\gamma^{(n)}$ is an ascending factorial.

Proposition 3.6 is identical to Proposition 3.3, as is its proof.

Strict Stationarity and Ergodicity

The existence of a stationary value for every raw moment implies the existence of a stationary moment generating function and thus a unique stationary distribution $f(y_t)$. Since $\mathbb{E}[\lambda_t^n] < \mathbb{E}[y_t^n]$, there also exists a stationary distribution $f(\lambda_t)$. Let $Y_{t:t+h} = \{y_t, \ldots, y_{t+h}\}$, $\Lambda_{t:t+h} = \{\lambda_t, \ldots, \lambda_{t+h}\}$, $X_{t:t+h} = \{x_t, \ldots, x_{t+h}\}$. The same approach used in Propositions 1-6 can be used to show the existence of $\mathbb{E}[y_t^n|Y_{t-1:t-i}, \Lambda_{t-1:t-j}]$, and thus $f[y_t|Y_{t-i:t-1}, \Lambda_{t-j:t-1}]$, $i = 1, \ldots, p$, $j = 1, \ldots, q$. There must then exist a density $f(Y_{t-p:t}, \Lambda_{t-q:t})$. Since $f(Y_{t:t+h}, \Lambda_{t:t+h}|Y_{t-p:t-1}, \Lambda_{t-q:t-1})$ is just a product of known densities, we can then conclude that $f(Y_{t:t+h})$ depends only on the value h.

A Markov chain for a random variable W_t is said to be ergodic if

$$\frac{1}{T}\sum_{t=1}^{T} f(W_t) \xrightarrow{a.s.} \int f(W)\pi(W)dW, \quad (3.22)$$

where $\pi(W)$ is the stationary distribution of W_t. Using the strict stationarity of y_t, we can

now show that Poisson and Negative Binomial AR(p) models are ergodic. The autoregressive processes described 3.2.1 and 3.2.2 are Markov chains on a countably infinite state space. For an AR(1) model, the Markov random variable is simply $W_t = y_t$. For an AR(p) process, the Markov random variable is a vector, $W_t = Y_{t:t+p}$. The Markov transition kernel $P(W_t, W_{t+1})$ is the probability of moving from W_t to W_{t+1}. An invariant measure π is a solution to the integral equation $\int P(W_t, W_{t+1})\pi(W_t)dW_t = \pi(W_{t+1})$. The existence of a finite-summable invariant measure is guaranteed by the strict stationarity of y_t. $P(W_t, W_{t+1})$ is either a Poisson or Negative Binomial Distribution in the AR(1) case or a product of these distributions when $p > 1$. Since these distributions assign positive probability to every nonnegative integer, P is π-irreducible and aperiodic. We can thus conclude that the Markov Chain is ergodic (Tierney, 1994).

Likelihood Computation

The likelihood for an AR(p) model of data $Y_{1:T}$ can be written as

$$f(Y_{1:T}|\theta, X) = f(Y_{1:p}|X_{1:p}) \prod_{t=p+1}^{T} f(y_t|\theta, x_t, Y_{t-1:t-p}). \tag{3.23}$$

The absence of a closed form expression for the stationary distribution of y_t leaves us with two options: use the conditional likelihood $f(Y_{p+1:T}|X, Y_p)$ or estimate the probability $f(Y_{1:p}|\theta, X_{1:p})$. The conditional likelihood is a reasonable choice in models where one is primarily interested in the effects of the covariates x_t. One may simply include a single lag of y_t to control for any possible autocorrelation. This becomes less attractive when analyzing data that one suspects is mostly or entirely driven by autoregressive dynamics. In this situation, objective selection of the number of lags is crucial. Common methods for lag selection include Bayes factors and information criteria. Both of these approaches require the calculation of $f(Y_{1:T}|\theta, X)$. The conditional likelihood can now only be used if we condition each estima-

tion on the first $\bar{p}$ observations, the maximum number of lags considered. When the sample size is small, this can cost precious degrees of freedom. One of our empirical applications will compare results obtained from using either conditional or stationary likelihoods.

The likelihood component $f(Y_{1:p}|\theta, X_{1:p})$ can be estimated via Markov Chain Monte Carlo (MCMC) integration as in Jeliazkov and Lee (2010). The invariance property of the Markov chain tells us that $\int f(Y_{1:p}|Y_{-p:0}, \theta) f(Y_{1:p}|\theta) dY_{-p:0} = f(Y_{1:p}|\theta)$. Using B MCMC samples from $f(Y_{1:p}|\theta)$, we can estimate the stationary probability as

$$\hat{f}(Y_{1:p}|\theta) = \frac{1}{B} \sum_{b=1}^{B} f(Y_{1:p}|Y_{-p:0}^{(b)}). \tag{3.24}$$

3.3 Empirical Application 1: Daylight Savings Time and Fatal Car Crashes

A recent strand of empirical literature has focused on the social costs that Daylight Savings Time (DST) places on Americans. In particular, the one-hour shift of daylight during the spring transition has been shown to have both beneficial and detrimental consequences. For example, Barnes and Wagner (2009) find that Americans sleep 40 minutes less on the night of the spring transition, while Doleac and Sanders (2015) find that increased ambient light in the evening following the shift to DST decreases robberies by 7%. Moreover, Smith (2016) finds that the spring transition into DST increases fatal car crash risk by 5.0%-6.5% (translating to an annual increase of over 30 deaths from 2002-2011).

To demonstrate how our new conditional mean function may be applied to count data, our first empirical application replicates the results from Table 5 of Smith (2016). However, our model differs from Smith (2016) in a few respects. While Smith (2016) chooses to model

the natural logarithm of traffic fatalities as the dependent variable in a linear fixed effects model, we instead model traffic fatalities directly as the outcome of a conditional Poisson model. Our approach allows for direct estimation of the marginal effect of DST, as well as the inclusion of lags of the dependent variable. Lagged dependent variables help partially capture the effects of aperiodic, sustained events like weather.

The shape and scale parameters are not sufficiently identified such that they can be estimated accurately. k is poorly identified if the data is generated primarily from the convex or concave regions of the conditional mean function. c and the regression constant (β_0) are not jointly identified when $x_i'\beta$ is mostly negative, making the conditional mean function approximately exponential. We instead estimate a set of 35 models over a grid of five c values and seven k values via maximum likelihood. We then choose the values of c and k that yield the highest likelihood. Maximum likelihood estimates are obtained using the Berndt et al. (1974) (BHHH) algorithm. This algorithm is the natural choice because the inclusion of year, day-of-week, and day-of-year fixed effects creates a very high dimensional parameter vector.

Parameter estimates are provided in Table 3.3. We see complete agreement in terms of sign and significance of covariate effects with the results of Smith (2016). In addition, lagged values of the dependent variable are highly significant in all but one of the specifications. Figure 3.3 shows the shape of the conditional mean function for all four dependent variables using the results from columns (4)-(7) of Table 3.3. We see that the conditional mean function has a convex, slightly exponential shape when the outcome variable is fatalities in all hours or in the least light impacted times. It becomes linear when we examine fatal car crashes in the morning or evening hours. That the conditional mean functions are neither strongly exponential nor strongly concave demonstrates that both a conditional Poisson model with an exponential conditional mean function and a log-linear model are ill-suited to the data.

	All Hours				Least Light Impacted	Morning	Evening
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
y_{t-1}	0.0021***	0.002***	0.002***	0.002***	0.0025***	0.0074	0.0109***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0004)	(0.0053)	(0.0026)
Spring DST	0.0384**	0.037**					
	(0.0168)	(0.0168)					
First six			0.0642***	0.0634***	0.083**	0.5756***	-0.0249
days of DST			(0.0208)	(0.0209)	(0.0336)	(0.1881)	(0.1076)
Next eight			0.0286	0.027	0.0373	0.4187***	-0.1169
days of DST			(0.0228)	(0.0228)	(0.0349)	(0.1551)	(0.103)
Remainder of			0.0204	0.0181	0.0152	0.3602**	-0.0668
Spring DST			(0.0208)	(0.0208)	(0.0323)	(0.1519)	(0.0993)
Fall DST	0.0176	0.0178	0.0176	0.0177	0.0622	0.8015***	-0.4476***
	(0.0269)	(0.027)	(0.0269)	(0.027)	(0.041)	(0.2114)	(0.16)
ln(gas price)		-0.0362*		-0.0378*	-0.0444	-0.2455	-0.0512
		(0.0216)		(0.0215)	(0.035)	(0.1852)	(0.1148)
Observations	3,341	3,341	3,341	3,341	3,341	3,341	3,341
c	500	500	500	500	100	5	10
k	1.1	1.1	1.1	1.1	1.1	1.1	1.1

Table 3.3: Parameter Estimates for Daylight Savings Models

Notes: The dependent variable (y_t) is the daily number of fatal car crashes. Maximum likelihood estimates are reported with asymptotic standard errors in parentheses. All specifications include year, day-of-year, and day-of-week fixed effects. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

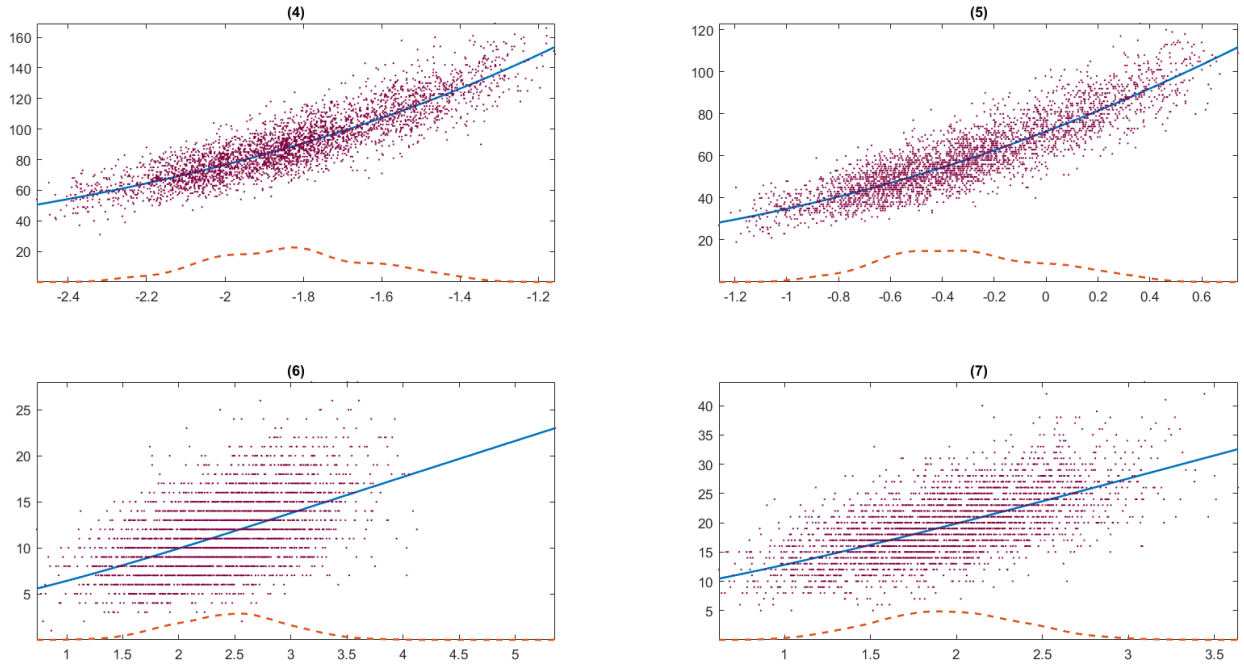


Figure 3.3: Daylight Savings Conditional Mean Functions

Notes: The solid line is the conditional mean function. The dots are the observed outcomes plotted against $x'\hat{\beta}_{MLE}$. The dotted line is the empirical density of $x'\hat{\beta}_{MLE}$.

3.4 Empirical Application 2: Slot Controls and Flight Delays

Our second empirical application examines how regulatory policy restricting the number of flights at capacity constrained airports (i.e., slot controls) affects flight delays. Slot controls are currently implemented at over 200 capacity constrained airports worldwide to mitigate persistent congestion and delays. However, evidence is limited on the effectiveness of these policies. Although it is expected that the decline in flight traffic resulting from slot controls should decrease delays, the findings in Ater (2012) suggest that policies aimed at reducing congestion at highly concentrated airports will only have a limited impact because dominant airlines already internalize congestion when scheduling flights.

In the United States, slot controls are currently in place at Washington National, LaGuardia, and John F. Kennedy airports. Slot controls have previously been implemented at Chicago O'Hare and Newark (EWR) airports. In our empirical application, we evaluate how the introduction of slot controls at EWR on June 20$^{\text{th}}$, 2008 and the removal of these controls on October 30$^{\text{th}}$, 2016 affected the daily number of delayed flights.

EWR provides an interesting case study for two reasons. Foremost, EWR is a heavily concentrated airport. At the time slot controls were implemented, Continental (now United) accounted for 72% of flight traffic at EWR. Second, EWR is consistently one of the most delay-prone airports in the United States. Prior to the implementation of slot controls, the percentage of on-time gate arrivals decreased from 70.66% in 2000, to 63.97% in 2006 and 61.71% in 2007. Over the same period, the average daily count of flights with arrival delays greater than one hour were 54 in 2000, 79 in 2006, and 93 in 2007 (Transportation (DOT), 2008).

3.4.1 Data

The data we use to evaluate the effectiveness of slot controls at EWR are derived from the Bureau of Transportation Statistics (BTS). Since 1987, airlines with at least one percent of total domestic traffic have been required to report on-time performance (OTP) data to the BTS. For each domestic flight, the OTP data includes the scheduled arrival and departure time, the actual arrival and departure time, whether the flight was canceled or diverted, and indicators specifying whether the flight departed or arrived fifteen or more minutes past the scheduled departure or arrival time. Notably, the Department of Transportation defines a late arrival (departure) as any flight that arrives (departs) fifteen or more minutes past the scheduled arrival (departure) time. In other words, a flight that arrives (departs) fourteen minutes past the scheduled arrival (departure) is considered "on-time."

From the OTP data, we construct a fifteen-year panel of daily observations encompassing the period from January 1st, 2005 to December 31st, 2019. For each day, we compute four delay measures: the count of flights arriving late, the count of flights departing late, the count of flights arriving sixty or more minutes late, and the count of flights departing sixty or more minutes late. During our sample period, slot controls were in effect at EWR from June 20th, 2008 to October 30th, 2016. Our sample period provides us with approximately 3.5 years of data prior to the implementation of slot controls, 8.3 years of data when the slot controls are in effect, and 3.2 years of data after the slot controls were removed.

Recognizing that adverse weather contributes to flight delays, we supplemented the OTP data with daily weather measures derived from the National Oceanographic and Atmospheric Administration. In our empirical application, we control for adverse weather by including a series of indicator variables that account for the presence of heavy rain, snowfall, high winds, thunder, and fog.

3.4.2 Results

Consistent with our DST estimations in Section 3.3, we estimated a set of models over a grid of five c values and seven k values. The maximum likelihood estimates for these models are provided in Tables 3.4-3.7. All models were estimated with year, month, and day-of-week fixed effects to account for seasonality in flight traffic. Table 3.4 presents estimates when the count of late arrivals is the dependent variable, Table 3.5 when the count of flights arriving 60 or more minutes late is the dependent variable, Table 3.6 when the count of late departures is the dependent variable, and Table 3.7 when the count of flights departing 60 or more minutes late is the dependent variable. In each table, our preferred specification is provided in column 4. This specification includes weather controls, a quadratic time trend, and a variable controlling for the daily number of flights at EWR (i.e., the total number

of arrivals and departures). Accordingly, the coefficient on Slot Control Period in column 4 of each table indicates the effect of slot controls on the daily number of delayed flights at EWR, after conditioning on total flight traffic.

Slot Control Period	-0.2023***	-0.4831***	-0.5233***	-0.2535***
(June 20th, 2008 - October 29th, 2016)	(0.0017)	(0.0038)	(0.0041)	(0.0040)
Flights				0.0057***
				(0.0000)
Observations	5,475	5,475	5,475	5,475
c	500	50	50	50
k	1.1	1.1	1.1	1.1
Quadratic Time Trend	No	Yes	Yes	Yes
Weather Controls	No	No	Yes	Yes

Table 3.4: The Impact of Slot Controls on Late Arrivals

Notes: The dependent variable is the daily number of flights arriving 15 or more minutes late at Newark airport. Maximum likelihood estimates are reported with asymptotic standard errors in parentheses. All specifications include year, month-of-year, and day-of-week fixed effects. The sample period is January 1st, 2005 through December 31st, 2019. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Across all four flight delay measures, the coefficient on Slot Control Period in column 4 of Tables 3.4-3.7 is negative and statistically significant, indicating that slot controls were effective at reducing the daily number of delayed arriving and departing flights at EWR. Using the estimates from column 4 of each table, conditional mean functions for each dependent variable are shown in Figure 3.4. To illustrate the importance of a flexible conditional mean function, we also estimated the specifications displayed in Figure 3.4 using an exponential conditional mean function. The conditional mean functions using an exponential conditional mean are plotted in Figure 3.5. Compared to our new flexible conditional mean function in Figure 3.4, the exponential conditional mean function in Figure 3.5 offers a substantially worse fit to the data. The variables arrivals, arrivals-60, departures, and departures-60 refer to the number of late arrivals, the number of arrivals late by more than 60 minutes, the number of late departures, and the number of departures late by more than 60 minutes, respectively.

Slot Control Period	-0.5371***	-1.0191***	-1.1654***	-1.3224***
(June 20th, 2008 - October 29th, 2016)	(0.0062)	(0.0105)	(0.0122)	(0.0126)
Flights				-0.0033***
				(0.0000)
Observations	5,475	5,475	5,475	5,475
c	100	10	10	50
k	3	1.1	1.1	1.1
Quadratic Time Trend	No	Yes	Yes	Yes
Weather Controls	No	No	Yes	Yes

Table 3.5: The Impact of Slot Controls on Late Arrivals (60 Minutes or More)

Notes: The dependent variable is the daily number of flights arriving 60 or more minutes late at Newark airport. Maximum likelihood estimates are reported with asymptotic standard errors in parentheses. All specifications include year, month-of-year, and day-of-week fixed effects. The sample period is January 1st, 2005 through December 31st, 2019. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

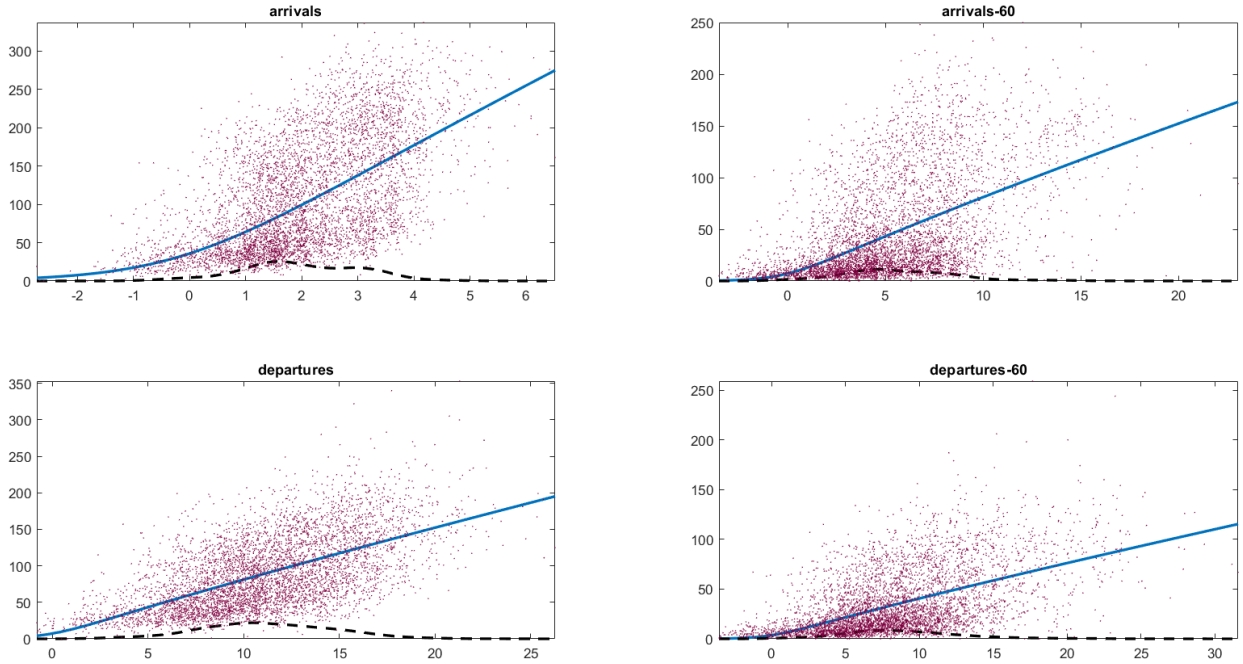


Figure 3.4: Slot Control Conditional Mean Functions (LW Conditional Mean Function)

Notes: The solid line is the conditional mean function. The dots are the observed outcomes plotted against $x'\hat{\beta}_{MLE}$. The dotted line is the empirical density of $x'\hat{\beta}_{MLE}$.

Slot Control Period	-0.3129***	-2.3886***	-0.5368***	-1.9463***
(June 20th, 2008 - October 29th, 2016)	(0.0035)	(0.0259)	(0.0055)	(0.0280)
Flights				0.0127***
				(0.0000)
Observations	5,475	5,475	5,475	5,475
c	100	10	50	10
k	1.1	1.1	1.1	1.1
Quadratic Trend	No	Yes	Yes	Yes
Weather Controls	No	No	Yes	Yes

Table 3.6: The Impact of Slot Controls on Late Departures

Notes: The dependent variable is the daily number of flights departing 15 or more minutes late from Newark airport. Maximum likelihood estimates are reported with asymptotic standard errors in parentheses. All specifications include year, month-of-year, and day-of-week fixed effects. The sample period is January 1st, 2005 through December 31st, 2019. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

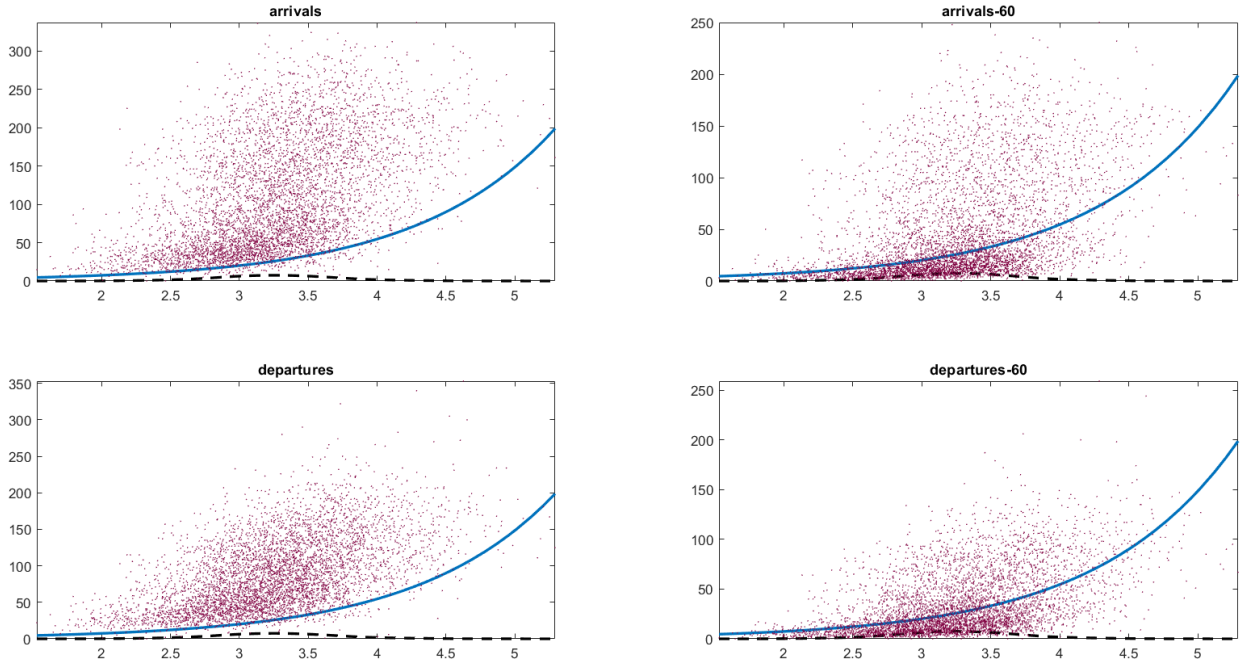


Figure 3.5: Slot Control Conditional Mean Functions (Exponential Conditional Mean Function)

Notes: The solid line is the conditional mean function. The dots are the observed outcomes plotted against $x'\hat{\beta}_{MLE}$. The dotted line is the empirical density of $x'\hat{\beta}_{MLE}$.

Slot Control Period	-0.3599***	-2.4580***	-2.7461***	-2.6489***
(June 20th, 2008 - October 29th, 2016)	(0.0045)	(0.0256)	(0.0314)	(0.0310)
Flights				-0.0005***
				(0.0001)
Observations	5,475	5,475	5,475	5,475
c	50	5	1000	5
k	1.1	1.1	10	1.1
Quadratic Trend	No	Yes	Yes	Yes
Weather Controls	No	No	Yes	Yes

Table 3.7: The Impact of Slot Controls on Late Departures (60 Minutes or More)

Notes: The dependent variable is the daily number of flights departing 60 or more minutes late from Newark airport. Maximum likelihood estimates are reported with asymptotic standard errors in parentheses. All specifications include year, month-of-year, and day-of-week fixed effects. The sample period is January 1st, 2005 through December 31st, 2019. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

Since the conditional Poisson model is nonlinear, parameter estimates are not directly interpretable as marginal effects. To aid our understanding of how slot controls affect the count of delayed flights at EWR, we examined the distribution of covariate effects using the procedure described in Jeliazkov and Vossmeyer (2018). Histograms of the impact of slot controls on the four dependent variables are presented in Figure 3.6. The distributions account for both variation in the data and parameter uncertainty. Due to the relatively high computational cost of evaluating the likelihood, parameters are sampled from their asymptotic distributions. Because the sample size is rather large, this procedure should produce a reasonable approximation of the posterior parameter distribution.

The distributions presented in Figure 3.6 indicate that slot controls were very effective at reducing the daily number of delayed flights at EWR. For example, the mode of the distribution of late arrivals in the top left panel of Figure 3.6 indicates that slot controls reduced the daily number of flights that arrived late by an average of 10 during the slot control period. Similarly, the mode of the distribution of late departures in the bottom left panel of Figure 3.6 indicates that slot controls reduced the daily number of flights that departed late by an

average of 15 during the slot control period.

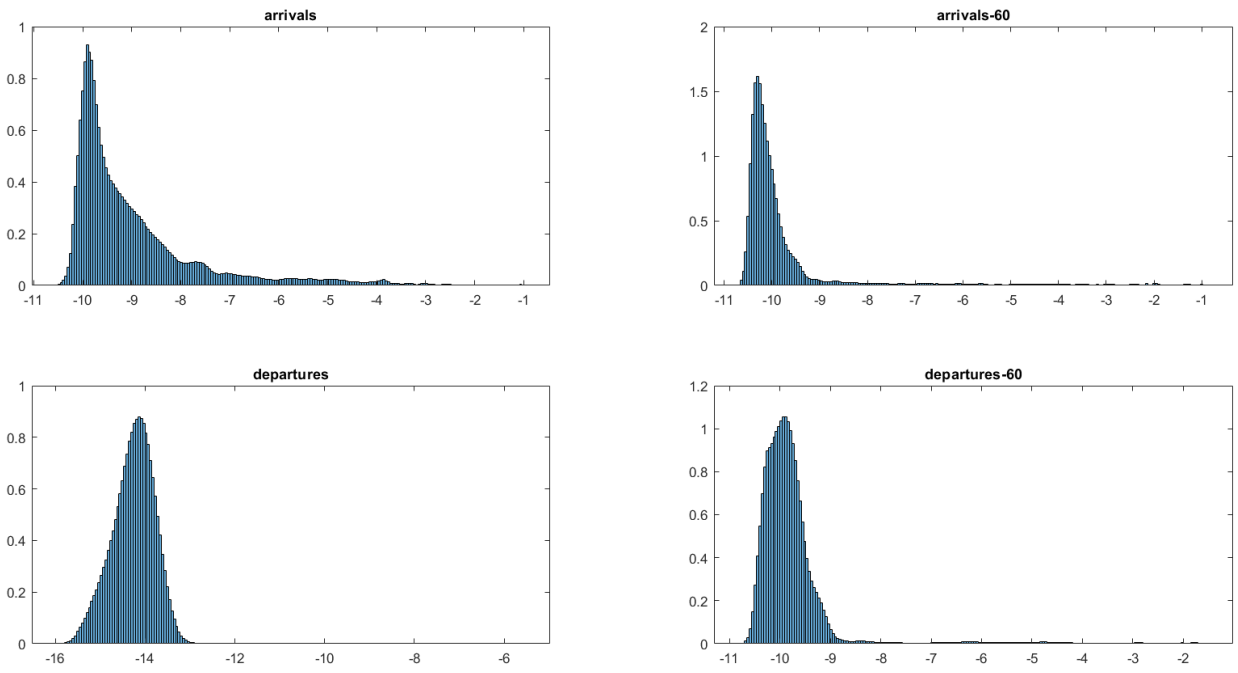


Figure 3.6: The Distributions of Marginal Effects

3.5 Conclusion

This paper proposed a new conditional mean function for count data modelling. The proposed function avoids the problems of explosive covariate effects, model misspecification, and nonstationarity that occur with the exponential conditional mean function. The LW function was used to construct time series count processes from conditional Poisson and Negative Binomial distributions. The processes were shown to be strictly stationary and ergodic under any parameter values. The LW function was used in three applications. We showed that the Spring Daylight Savings Time change increases traffic fatalities and slot controls positively impacted airline efficiency.

There are several avenues for further research. While we have focused on count data, trans-

formations involving the exponential function appear broadly in econometrics. The exponential function could be replaced by our proposed function in conditional heteroskedasticity models, gravity models of trade, and any other model with nonnegative data or parameters.

Bibliography

Ahn, Seung C., and Alex R. Horenstein. 2013. "Eigenvalue Ratio Test for the Number of Factors." *Econometrica* 81 (3): 1203–1227.

Albert, James H., and Siddhartha Chib. 1993. "Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts." *Journal of Business Economic Statistics* 11 (1): 1–15.

Ater, Itai. 2012. "Internalization of Congestion at US Hub Airports." *Journal of Urban Economics* 72 (2-3): 196–209.

Auestad, Bjørn, and Dag Tjøstheim. 1990. "Identification of Nonlinear Time Series: First Order Characterization and Order Determination." *Biometrika* 77 (4): 669–687.

Babak, Shahbaba. 2009. "Discovering Hidden Structures Using Mixture Models: Application to Nonlinear Time Series Processes." *Studies in Nonlinear Dynamics Econometrics* 13 (2): 1–21.

Bai, Jushan, and Serena Ng. 2002. "Determining the Number of Factors in Approximate Factor Models." *Econometrica* 70 (1): 191–221.

———. 2006. "Evaluating Latent and Observed Factors in Macroeconomics and Finance." *Journal of Econometrics* 131 (1): 507–537.

Bai, Jushan, and Serena Ng. 2007. "Determining the Number of Primitive Shocks in Factor Models." *Journal of Business & Economic Statistics* 25 (1): 52–60.

Barnes, Christopher M., and David T. Wagner. 2009. "Changing to Daylight Saving Time Cuts into Sleep and Increases Workplace Injuries." *Journal of Applied Psychology* 94 (5): 1305.

Baum, Leonard E., and Ted Petrie. 1966. "Statistical Inference for Probabilistic Functions of Finite State Markov Chains." *The Annals of Mathematical Statistics* 37 (6): 1554–1563.

Beaudry, Paul, and Gary Koop. 1993. "Do Recessions Permanently Change Output?" *Journal of Monetary Economics* 31 (2): 149–163.

Belviso, Francesco, and Fabio Milani. 2006. "Structural Factor-Augmented VARs (SFAVARs) and the Effects of Monetary Policy." *Topics in Macroeconomics* 6 (3).

Bernanke, Ben S., Jean Boivin, and Piotr Eliasz. 2005. "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach." *The Quarterly Journal of Economics* 120, no. 1 (February): 387–422.

Berndt, Ernst R., Bronwyn H. Hall, Robert E. Hall, and Jerry A. Hausman. 1974. "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement* 3:653–665.

Beyeler, Simon, and Sylvia Kaufmann. 2021. "Reduced-Form Factor Augmented VAR - Exploiting Sparsity to Include Meaningful Factors." *Journal of Applied Econometrics* 36 (7): 989–1012.

Binks, Rachel L., Sarah E. Heaps, Mariella Panagiotopoulou, Yujiang Wang, and Darren J. Wilkinson. 2023. *Bayesian Inference on the Order of Stationary Vector Autoregressions.* arXiv: 2307.05708.

Blundell, Richard, Rachel Griffith, and John Van Reenen. 1995. "Dynamic Count Data Models of Technological Innovation." *The Economic Journal* 105 (42): 333–344.

Böhning, Dankmar, Ekkehart Dietz, Peter Schlattmann, Lucia Mendonça, and Ursula Kirchner. 1999. "The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162:195–209.

Boivin, Jean, Marc P Giannoni, and Dalibor Stevanović. 2013. "Dynamic Effects of Credit Shocks in a Data-Rich Environment." *FRB of New York Staff Report,* no. 615.

Boldin, Michael. 1996. "A Check on the Robustness of Hamilton's Markov Switching Model Approach to the Economic Analysis of the Business Cycle." *Studies in Nonlinear Dynamics Econometrics* 1 (1): 1–14.

Brännäs, Kurt, and Alastair Hall. 2001. "Estimation in Integer-Valued Moving Average Models." *Applied Stochastic Models in Business and Industry* 17:277–291.

Cameron, A. Colin, and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data.* Cambridge: Cambridge University Press.

Cameron, A. Colin, Pravin K. Trivedi, Frank Milne, and John Piggott. 1988. "A Microeconometric Model of the Demand for Health Care and Health Insurance in Australia." *Review of Economic Studies* 55:85–106.

Carvalho, Carlos M., Jeffrey Chang, Joseph E. Lucas, Joseph R. Nevins, Quanli Wang, and Mike West. 2008. "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics." *Journal of the American Statistical Association* 103 (484): 1438–1456.

Chauvet, Marcelle. 1998. "An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switching." *International Economic Review* 39 (4): 969–96.

Chauvet, Marcelle, and James D. Hamilton. 2006. "Dating Business Cycle Turning Points." In *Nonlinear Time Series Analysis of Business Cycles,* 1–54. Emerald Group Publishing Limited.

Chen, Nai-Fu, Richard Roll, and Stephen A. Ross. 1986. "Economic Forces and the Stock Market." *The Journal of Business* 59 (3): 383–403.

Chib, Siddhartha. 1995. "Marginal Likelihood from the Gibbs Output." *Journal of the American Statistical Association* 90 (432): 1313–1321.

———. 1996. "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models." *Journal of Econometrics* 75 (1): 79–97.

———. 1998. "Estimation and Comparison of Multiple Change-Point Models." *Journal of Econometrics* 86 (2): 221–241.

Chib, Siddhartha, and Michael Dueker. 2004. *Non-Markovian Regime Switching with Endogenous States and Time-Varying State Strengths.* Working Papers 2004-030. Federal Reserve Bank of St. Louis.

Chib, Siddhartha, and Ivan Jeliazkov. 2001. "Marginal Likelihood From the Metropolis - Hastings Output." *Journal of the American Statistical Association* 96 (453): 270–281.

Cochrane, John H. 2011. "Presidential Address: Discount Rates." *The Journal of Finance* 66 (4): 1047–1108.

Corrado, Carol, and Joe Mattey. 1997. "Capacity Utilization." *Journal of Economic Perspectives* 11, no. 1 (March): 151–167.

Davis, Richard A., William T. M. Dunsmuir, and Stephen B. Streett. 2003. "Observation-Driven Models for Poisson Counts." *Biometrika* 90:777–790.

Davis, Richard A., William T. M. Dunsmuir, and Yi Wang. 2000. "On Autocorrelation in a Poisson Regression Model." *Biometrika* 87:491–505.

Delatola, Eleni Ioanna, and Jim Griffin. 2011. "Bayesian Nonparametric Modelling of the Return Distribution with Stochastic Volatility." *Bayesian Analysis* 6 (August).

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the EM Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22.

Diebold, Francis, Joon-Haeng Lee, and Gretchen Weinbach. 1993. "Regime Switching with Time-Varying Transition Probabilities." *Nonstationary Time Series Analysis and Cointegration* (February).

Doleac, Jennifer L., and Nicholas J. Sanders. 2015. "Under the Cover of Darkness: How Ambient Light Influences Criminal Activity." *Review of Economics and Statistics* 97 (5): 1093–1103.

Dueker, Michael. 1997. "Markov Switching in GARCH Processes and Mean-Reverting Stock-Market Volatility." *Journal of Business Economic Statistics* 15 (1): 26–34.

Dugas, Claude, Yoshua Bengio, Frédéric Bélisle, Claude Nadeau, and Roberto Garcia. 2001. "Incorporating Second-Order Functional Knowledge for Better Option Pricing." In *Advances in neural information processing systems,* 472–478.

Fama, Eugene F., and Kenneth R. French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of financial economics* 33 (1): 3–56.

———. 2015. "A five-factor asset pricing model." *Journal of Financial Economics* 116 (1): 1–22.

Fernald, John G, Mark M Spiegel, and Eric T Swanson. 2014. "Monetary Policy Effectiveness in China: Evidence From a FAVAR model." *Journal of International Money and Finance* 49:83–103.

Filardo, Andrew. 1994. "Business-Cycle Phases and Their Transitional Dynamics." *Journal of Business Economic Statistics* 12 (3): 299–308.

Filardo, Andrew, and Stephen Gordon. 1998. "Business Cycle Durations." *Journal of Econometrics* 85 (1): 99–123.

Fruhwirth-Schnatter, Sylvia. 2001. "Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models." *Journal of the American Statistical Association* 96 (453): 194–209.

Frühwirth-Schnatter, Sylvia, and Hedibert Freitas Lopes. 2009. *Parsimonious Bayesian Factor Analysis When the Number of Factors is Unknown.* Technical report. University of Chicago Booth School of Business.

Ghysels, Eric, Robert E. McCulloch, and Ruey S. Tsay. 1998. "Bayesian Inference for Periodic Regime-Switching Models." *Journal of Applied Econometrics* 13 (2): 129–143.

Goldfeld, Stephen M., and Richard E. Quandt. 1973. "A Markov Model for Switching Regressions." *Journal of Econometrics* 1 (1): 3–15.

Guidolin, Massimo, and Allan Timmermann. 2005. "Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns." *Economic Journal* 115 (500): 111–143.

Haas, M., S. Mittnik, and M. S. Paolella. 2004. "A New Approach to Markov-Switching GARCH Models." *Journal of Financial Econometrics* 2, no. 4 (September): 493–530.

Hallin, Marc, and Roman Liška. 2007. "Determining the Number of Factors in the General Dynamic Factor Model." *Journal of the American Statistical Association* 102 (478): 603–617.

Hamilton, James. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57 (2): 357–84.

Hamilton, James. 2001. "A Parametric Approach to Flexible Nonlinear Inference." *Econometrica* 69 (3): 537–73.

Härdle, W., A. Tsybakov, and L. Yang. 1998. "Nonparametric Vector Autoregression." Nonlinear Time Series Models, Part 2, *Journal of Statistical Planning and Inference* 68 (2): 221–245.

Heaps, Sarah E. 2023. "Enforcing Stationarity Through the Prior in Vector Autoregressions." *Journal of Computational and Graphical Statistics* 32 (1): 74–83.

Heinen, Andreas. 2003. *Modelling Time Series Count Data: an Autoregressive Conditional Poisson Model.* Technical report. Core Discussion Paper No. 2003-63.

Huber, Florian, and Luca Rossini. 2021. *Inference in Bayesian Additive Vector Autoregressive Tree Models.* arXiv: 2006.16333.

Hwu, Shih-Tang, and Chang-Jin Kim. 2023. *Studies in Nonlinear Dynamics Econometrics.*

Hwu, Shih-Tang, Chang-Jin Kim, and Jeremy Piger. 2021. "An N-State Endogenous Markov-Switching Model with Applications in Macroeconomics and Finance." *Macroeconomic Dynamics* 25 (8): 1937–1965.

Jeliazkov, Ivan. 2013. "Nonparametric Vector Autoregressions: Specification, Estimation, and Inference." In *VAR Models in Macroeconomics – New Developments and Applications: Essays in Honor of Christopher A. Sims,* 32:327–359. Emerald Group Publishing Limited.

Jeliazkov, Ivan, and Esther Hee Lee. 2010. "MCMC Perspectives on Simulated Likelihood Estimation." In *Maximum Simulated Likelihood Methods and Applications,* 3–39. Emerald Group Publishing Limited.

Jeliazkov, Ivan, and Angela Vossmeyer. 2018. "The Impact of Estimation Uncertainty on Covariate Effects in Nonlinear Models." *Statistical Papers* 59 (3): 1031–1042.

Jensen, Mark, and John Maheu. 2010. "Bayesian Semiparametric Stochastic Volatility Modeling." *Journal of Econometrics* 157 (2): 306–316.

Jin, Sainan, Ke Miao, and Liangjun Su. 2021. "On Factor Models with Random Missing: EM Estimation, Inference, and Cross Validation." *Journal of Econometrics* 222 (1, Part C): 745–777.

Jung, Robert, Martin Kukuk, and Roman Liesenfeld. 2006. "Time Series of Count Data: Modeling, Estimation and Diagnostics." *Computational Statistics & Data Analysis* 51 (4): 2350–2364.

Kalli, Maria, and Jim Griffin. 2018. "Bayesian Nonparametric Vector Autoregressive Models." *Journal of Econometrics* 203 (2): 267–282.

Kang, Kyu H. 2014. "Estimation of State-Space Models with Endogenous Markov Regime-Switching Parameters." *Econometrics Journal* 17 (1): 56–82.

Kass, Robert E., and Adrian E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90 (430): 773–795.

Kim, Chang-Jin. 1994. "Dynamic Linear Models with Markov-Switching." *Journal of Econometrics* 60 (1-2): 1–22.

Kim, Chang-Jin, and Charles R. Nelson. 1999. *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications.* Vol. 1. MIT Press Books 0262112388. The MIT Press, December.

Kim, Chang-Jin, Jeremy Piger, and Richard Startz. 2008. "Estimation of Markov Regime-Switching Regression Models with Endogenous Switching." *Journal of Econometrics* 143 (2): 263–273.

Kim, Young Min, and Kyu Ho Kang. 2022. "Bayesian Inference of Multivariate Regression Models with Endogenous Markov Regime-Switching Parameters*." *The Journal of Financial Econometrics* 20 (3): 391–436.

Knowles, David, and Zoubin Ghahramani. 2011. "Nonparametric Bayesian Sparse Factor Models with Application to Gene Expression Modeling." *The Annals of Applied Statistics* 5 (2B): 1534–1552.

Koop, Gary, and Dimitris Korobilis. 2013. "Large Time-Varying Parameter VARs." *Journal of Econometrics* 177 (2): 185–198.

Koop, Gary, and Simon M. Potter. 1999. "Bayes Factors and Nonlinearity: Evidence from Economic Time Series." *Journal of Econometrics* 88 (2): 251–281.

Litterman, Robert. 1986. "Forecasting with Bayesian Vector Autoregressions-Five Years of Experience." *Journal of Business Economic Statistics* 4 (1): 25–38.

Liu, Chuanhai, and Donald B. Rubin. 1994. "The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence." *Biometrika* 81 (4): 633–648. Accessed August 8, 2023.

Liu, Chuanhai, Donald B. Rubin, and Ying Nian Wu. 1998. "Parameter Expansion to Accelerate EM: The PX-EM Algorithm." *Biometrika* 85 (4): 755–770.

Luo, Jiayi, and Cindy Long Yu. 2021. "Determining Number of Factors in Dynamic Factor Models Contributing to GDP Nowcasting." *Mathematics* 9 (22).

McAlinn, Kenichiro, Veronika Ročková, and Enakshi Saha. 2018. *Dynamic Sparse Factor Analysis.* arXiv: 1812.04187.

McCracken, Michael, and Serena Ng. 2016. "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business & Economic Statistics* 34 (4): 574–589.

McCracken, Michael, and Serena Ng. 2020. *FRED-QD: A Quarterly Database for Macroeconomic Research.* Technical report. National Bureau of Economic Research.

Al-Osh, M. A., and A. A. Alzaid. 1987. "First-Order Integer-Valued Autoregressive (INAR(1)) Process." *Journal of Time Series Analysis* 8:261–275.

Paccagnini, Alessia. 2017. *Forecasting with FAVAR: Macroeconomic Versus Financial Factors.* NBP Working Papers 256. Narodowy Bank Polski.

Parker, Jason, and Donggyu Sul. 2016. "Identification of Unknown Common Factors: Leaders and Followers." *Journal of Business & Economic Statistics* 34 (2): 227–239.

Primiceri, Giorgio. 2005. "Time Varying Structural Vector Autoregressions and Monetary Policy." *Review of Economic Studies* 72 (3): 821–852.

Ročková, Veronika, and Edward I. George. 2016. "Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity." *Journal of the American Statistical Association* 111 (516): 1608–1622.

Ruud, Paul A. 1991. "Extensions of Estimation Methods Using the EM Algorithm." *Journal of Econometrics* 49 (3): 305–341.

Schoenmaker, Dirk. 1996. "Contagion Risk in Banking." LSE Financial Markets Group.

Sims, Christopher. 1980. "Macroeconomics and Reality." *Econometrica* 48 (1): 1–48.

Sims, Christopher, and Tao Zha. 2006. "Were There Regime Switches in U.S. Monetary Policy?" *American Economic Review* 96 (1): 54–81.

Smith, Austin. 2016. "Spring Forward at Your Own Risk: Daylight Saving Time and Fatal Vehicle Crashes." *American Economic Journal: Applied Economics* 8 (2): 65–91.

Stock, James H, and Mark W Watson. 2002. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business & Economic Statistics* 20 (2): 147–162.

Stock, James H, and Mark W Watson. 2016. "Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics." Chap. Chapter 8, 2:415–525. Elsevier.

Tierney, Luke. 1994. "Markov Chains for Exploring Posterior Distributions." *The Annals of Statistics* 22 (4): 1701–1728.

Tracy, Kevin. 2022. *A Square-Root Kalman Filter Using Only QR Decompositions.* arXiv: 2208.06452.

Transportation (DOT), Department of. 2008. *Federal Register March 18, 2008,* 53.

Uribe, Paloma W., and Hedibert F. Lopes. 2020. *Dynamic Sparsity on Dynamic Regression Models.* arXiv: 2009.14131.

Vigfusson, Robert. 1997. "Switching between Chartists and Fundamentalists: A Markov Regime-Switching Approach." *International Journal of Finance & Economics* 2, no. 4 (October): 291–305.

Watson, Mark W., and Robert F. Engle. 1983. "Alternative Algorithms for the Estimation of Dynamic Factor, Mimic and Varying Coefficient Regression Models." *Journal of Econometrics* 23 (3): 385–400.

Witten, Daniela M., Robert Tibshirani, and Trevor Hastie. 2009. "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis." *Biostatistics* 10, no. 3 (April): 515–534.

Wu, C. F. Jeff. 1983. "On the Convergence Properties of the EM Algorithm." *The Annals of Statistics* 11 (1): 95–103.

Zeger, Scott L. 1988. "A Regression Model for Time Series of Counts." *Biometrika* 75:621–629.

Zou, Hui, Trevor Hastie, and Robert Tibshirani. 2006. "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics* 15 (2): 265–286.

Appendix A

Chapter 1

A.1 Proof of Proposition 1.1

A different route to efficient evaluation of the gradient can be seen by noting that the integrated likelihood for a DFM is equivalent to that of a DFM that also includes presample instances of the state variable. If we let $F^{\dagger} = (f_0, f_{-1}, \ldots, f_{-\tau+1})$, the likelihood can then be expressed as

$$f(X|\theta) = \int f(X|F,\theta)\pi(F|\theta)dF = \int\int f(X|F,\theta)\pi(F|F^{\dagger},\theta)\pi(F^{\dagger}|\theta)dFdF^{\dagger}. \tag{A.1}$$

Since the model with presample factors is also valid, it is also amenable to the construction of an EM algorithm. Define $Q_{\tau}(\theta|\theta_n) \equiv \mathbb{E}[\ln f(X, F, F^{\dagger}|\theta)]$. Using the same result from Ruud (1991), we know that

$$\nabla \ln f(X|\theta_n) = \nabla Q_{\tau}(\theta_n|\theta_n). \tag{A.2}$$

I will now show that for τ sufficiently large, we can calculate the gradient using only the conditional terms in $Q_\tau(\theta_n|\theta_n)$ and omit any terms that involve the stationary distribution of the factors.

Proposition 1.1

Let $F^\dagger = (f_0, f_{-1}, \ldots, f_{-\tau+1})$, $Q_{\tau-p}(\theta|\theta_n) \equiv \mathbb{E}[\ln f(X, F, f_0, f_{-1}, \ldots, f_{-\tau+p+1}|$ $f_{-\tau+p}, \ldots, f_{-\tau+1}, \theta)|X, \theta_n]$ and assume θ_n is an interior point of the parameter space.

$$\lim_{\tau\to\infty} \nabla Q_{\tau-p}(\theta_n|\theta_n) = \nabla \ln f(X|\theta_n).$$

Proof. As $\tau \to \infty$, $\nabla Q_\tau(\theta_n|\theta_n)$ becomes an infinite sum. Since $\nabla \ln f(X|\theta_n) = \nabla Q_\tau(\theta_n|\theta_n)$, we know that this sum must converge to the desired gradient. All that remains is to show that the terms involving the stationary distribution go to 0. As the Kalman smoother is iterated backwards, the smoothed moments of the factors will converge to the stationary moments: $\mathbb{E}[g_t|X, \theta_n] \to \mathbb{E}[g_t|\theta_n] = 0$, $\mathbb{E}[(g_t - \hat{g}_t)(g_t - \hat{g}_t)'|X, \theta_n] \to P_0$. By Gibb's Inequality, $\mathbb{E}[\nabla \ln \pi(g_t|\theta_n)|\theta_n] = 0$ for any θ_n in the interior of the parameter space. We can thus conclude that

$$\begin{aligned}
\lim_{\tau\to\infty} \nabla Q_\tau(\theta_n|\theta_n) - \nabla Q_{\tau-p}(\theta_n|\theta_n) &= \lim_{\tau\to\infty} \mathbb{E}[\nabla \ln \pi(g_{-\tau+p}|\theta_n)|X, \theta_n] \\
&= \mathbb{E}[\nabla \ln \pi(g_t|\theta_n)|\theta_n] \\
&= 0.
\end{aligned}$$

□

For an accurate calculation of the gradient, τ should be chosen so that the smoothed moments

converge to the stationary moments. This will obviously depend on the persistence of shocks in the model. For highly persistent models, simulation results suggest that $\tau = 5,000$ is sufficiently large.

Appendix B

Chapter 3

B.1 Proof of Proposition 3.1

We will prove stationarity in the case of $x_t = 1$. Results can then be extended by replacing x_t with any exogenous covariates. We show that $\mathbb{E}[y_t]$ is bounded from below by using the law of iterated expectations. Since $\lambda_t > 0$,

$$\mathbb{E}[y_t] = \mathbb{E}[\lambda_t] > 0. \tag{B.1}$$

We will continue by showing that $\mathbb{E}[y_t]$ is bounded from above by a function of constant parameters. We can choose b_1 and m_1 to satisfy

$$\begin{aligned} &1.\ b_1 + m_1|x| > \lambda(|x|) \\ &2.\ 0 < m_1 < (\sum_{i=1}^{p} |\phi_i| + \sum_{j=1}^{q} |\psi_j|)^{-1} \end{aligned} \tag{B.2}$$

One way of finding such a pair, which is by no means optimal, is to pick a value for m_1 that satisfies condition 2 and then setting $b_1 = m_1 max\{g_0\}$, where g_0 are the roots of $g(|x|) = m_1|x| - \lambda(|x|)$. This then allows us to state

$$\begin{aligned}
\mathbb{E}[y_t] &= \mathbb{E}[\lambda(z_t'\delta)] \\
&\leq \mathbb{E}[\lambda(|z_t'\delta)|)] \\
&\leq b_1 + m_1\mathbb{E}[|\beta_0 + \sum_{i=1}^{p}\phi_i y_{t-i} + \sum_{j=1}^{q}\psi_j\lambda_{t-j}|] \\
&\leq b_1 + m_1\mathbb{E}[|\beta_0| + \sum_{i=1}^{p}|\phi_i y_{t-i}| + \sum_{j=1}^{q}|\psi_j\lambda_{t-j}|] \\
&= b_1 + m_1\mathbb{E}[|\beta_0| + \sum_{i=1}^{p}|\phi_i|y_{t-i} + \sum_{j=1}^{q}|\psi_j|\lambda_{t-j}] \\
&= b_1 + m_1|\beta_0| + m_1(\sum_{i=1}^{p}|\phi_i|\mathbb{E}[y_t] + \sum_{j=1}^{q}|\psi_j|\mathbb{E}[\lambda_t]) \\
&= b_1 + m_1|\beta_0| + m_1(\sum_{i=1}^{p}|\phi_i| + \sum_{j=1}^{q}|\psi_j|)\mathbb{E}[y_t].
\end{aligned} \tag{B.3}$$

Rearranging terms yields

$$0 < \mathbb{E}[y_t] < \frac{b_1 + m_1|\beta_0|}{1 - m_1(\sum_{i=1}^{p}|\phi_i| + \sum_{j=1}^{q}|\psi_j|)}. \tag{B.4}$$

□

B.2 Proof of Proposition 3.2

We will prove stationarity in the case of $x_t = 1$. Results can then be extended by replacing x_t with any exogenous covariates. We show that $\mathbb{E}[y_t^n]$ is bounded from below by using the

law of iterated expectations and the formula for raw moments of a Poisson random variable. Since $\lambda_t > 0$,

$$\mathbb{E}[y_t^n] = \mathbb{E}[\sum_{l=1}^{n} \left\{ {n \atop l} \right\} \lambda_t^l] > 0. \tag{B.5}$$

We will continue by showing that $\mathbb{E}[y_t^n]$ is bounded from above by a function of constant parameters. We can choose b_n and m_n to satisfy

$$\begin{aligned} &1.\ b_n + m_n|x|^n > \lambda(|x|)^n \\ &2.\ 0 < m_n < ((p+q+1)^{n-1}(\sum_{i=1}^{p} |\phi_i|^n + \sum_{j=1}^{q} |\psi_j|^n)^{-1} \end{aligned} \tag{B.6}$$

As in the first proof, we can pick a value for m_n that satisfies condition 2 and then setting $bn = m_1 max\{g_0\}^n$, where g_0 are the roots of $g(|x|) = m_n|x|^n - \lambda(|x|)^n$. We will first find a bound for $\mathbb{E}[\lambda_t^n]$:

$$
\begin{aligned}
\mathbb{E}[\lambda_t^n] &\leq \mathbb{E}[\lambda(|z_t'\delta)|)^n] \\
&\leq b_n + m_n \mathbb{E}[|\beta_0 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \psi_j \lambda_{t-j}|^n] \\
&\leq b_n + m_n (p+q+1)^{n-1} \mathbb{E}[|\beta_0|^n + \sum_{i=1}^{p} |\phi_i y_{t-i}|^n + \sum_{j=1}^{q} |\psi_j \lambda_{t-j}|^n] \\
&= b_n + m_n (p+q+1)^{n-1} \mathbb{E}[|\beta_0|^n + m_n (p+q+1)^{n-1} \sum_{i=1}^{p} |\phi_i|^n y_{t-i}^n \\
&\qquad + m_n (p+q+1)^{n-1} \sum_{j=1}^{q} |\psi_j|^n \lambda_{t-j}^n] \\
&= b_n + m_n (p+q+1)^{n-1} |\beta_0|^n + m_n (p+q+1)^{n-1} \sum_{i=1}^{p} |\phi_i|^n \mathbb{E}[y_t^n] \\
&\qquad + m_n (p+q+1)^{n-1} \sum_{j=1}^{q} |\psi_j|^n \mathbb{E}[\lambda_t^n] \\
&\leq b_n + m_n (p+q+1)^{n-1} |\beta_0|^n + m_n (p+q+1)^{n-1} (\sum_{i=1}^{p} |\phi_i|^n + \sum_{j=1}^{q} |\psi_j|^n) \mathbb{E}[y_t^n].
\end{aligned}
\tag{B.7}
$$

The second inequality is given by condition 1. The third inequality results from the convexity of $|x|^n$. The final inequality can be seen by writing

$$
\begin{aligned}
\mathbb{E}[\lambda_t^n] &= \mathbb{E}[y_t^n] - \sum_{l=1}^{n-1} \left\{ {n \atop l} \right\} \mathbb{E}[\lambda_t^l] \\
&< \mathbb{E}[y_t^n].
\end{aligned}
\tag{B.8}
$$

Using the bounds in B.7, B.8, and the Poisson moment formula, we arrive at the inequality

$$
\mathbb{E}[y_t^n] < b_n + m_n (p+q+1)^{n-1} |\beta_0|^n + m_n (p+q+1)^{n-1} (\sum_{i=1}^{p} |\phi_i|^n + \sum_{j=1}^{q} |\psi_j|^n) \mathbb{E}[y_t^n] + \sum_{l=1}^{n-1} \left\{ {n \atop l} \right\} \mathbb{E}[y_t^l].
\tag{B.9}
$$

Rearranging terms give the result

$$\mathbb{E}[y_t^n] < \frac{b_n + m_n|\beta_0|^n(p+q+1)^{n-1} + \sum_{l=1}^{n-1}\left\{{n \atop l}\right\}\mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}(\sum_{i=1}^{p}|\phi_i|^n + \sum_{j=1}^{q}|\psi_j|^n)}. \tag{B.10}$$

□

B.3 Proof of Proposition 3.3

By the nonnegativity of y_t, $\mathbb{E}[y_t y_{t-s}] \geq 0$. By the Cauchy-Schwarz Inequality,

$$Cov[y_t, y_{t-s}] = \mathbb{E}[y_t y_{t-s}] - \mu^2 \leq Var[y_t] < \mathbb{E}[y_t^2]. \tag{B.11}$$

Using these two facts, we can determine

$$-\mu^2 < Cov[y_t, y_{t-s}] \leq Var[y_t] < \mathbb{E}[y_t^2]. \tag{B.12}$$

□

B.4 Proof of Proposition 3.5

The proof outlined above extends to the case of a conditional Negative Binomial distribution with only minor modifications. The only difference is that we use the moment formula for the

Negative Binomial distribution. After taking expectations with respect to λ_t, the formula becomes

$$\mathbb{E}[y_t^n] = \sum_{l=1}^{n} \left\{ {n \atop l} \right\} \gamma^{-l} \gamma^{(l)} \mathbb{E}[\lambda_t^l]. \tag{B.13}$$

Inequality B.7 is unchanged and inequality B.8 becomes

$$\begin{aligned} \mathbb{E}[\lambda_t^n] &= (\gamma^{-n}\gamma^{(n)})^{-1}(\mathbb{E}[y_t^n] - \sum_{l=1}^{n-1} \left\{ {n \atop l} \right\} \gamma^{-l}\gamma^{(l)}\mathbb{E}[\lambda_t^l]) \\ &< \mathbb{E}[y_t^n]. \end{aligned} \tag{B.14}$$

Simple substitution and gathering of terms leads to

$$0 \leq \mathbb{E}[y_t^n] < \frac{\gamma^{-n}\gamma^{(n-1)}(b_n + m_n|\beta_0|^n(p+q+1)^{n-1}) + \sum_{l=1}^{n-1} \left\{ {n \atop l} \right\} \gamma^{-l}\gamma^{(l-1)}\mathbb{E}[y_t^l]}{1 - m_n(p+q+1)^{n-1}\gamma^{-n}\gamma^{(n-1)}(\sum_{i=1}^{p}|\phi_i|^n + \sum_{j=1}^{q}|\psi_j|^n)}. \tag{B.15}$$

□