Dynamic State-Space Models*

Paul Karapanagiotidis†

Draft 6
June 3, 2014

Abstract

A review of the general state-space modeling framework. The discussion focuses heavily on the three prediction problems of forecasting, filtering, and smoothing within the state-space context. Numerous examples are provided detailing special cases of the state-space model and its use in solving a number of modeling issues. Independent sections are also devoted to both the topics of Factor models and Harvey's Unobserved Components framework.

Keywords: state-space models, signal extraction, unobserved components.
JEL: C10, C32, C51, C53, C58

1 Introduction

The *dynamic state-space model* was developed in the control systems literature, where physical systems are described mathematically as sets of inputs, outputs, and state variables, related by difference equations. The following, Section 2, describes the various versions of the linear state-space framework, discusses the relevant assumptions imposed, and provides examples encountered in economics and finance. Subsequently, Section 3 provides the analogous description of the more general nonlinear state-space framework. Section 4 then discusses some of the common terminologies related to the state-space framework within the different contexts they are encountered. Section 5 follows by discussing the general problems of state-space prediction, including forecasting, filtering, and smoothing. Moreover, it provides a number of simple applications to chosen models from Section 2. Sections 6 and 7 then go into more detail: Section

*I'd like to thank Christian Gourieroux for his helpful comments and suggestions.
†University of Toronto, Department of Economics, p.karapanagiotidis@utoronto.ca

6 briefly discusses the problem of prediction in the frequency domain and Section 7 outlines in detail the solutions to the forecasting, smoothing, and filtering problems, within the time domain. In particular, we interpret the solutions in terms of an orthogonal basis, and provide the MA and AR representations of the state-space model. Section 8 then details estimation of the state-space model parameters in the time domain. Finally, Section 9 discusses the equivalent Factor model representation, including the relationship between this representation, the VARMA, and the VECM models. It also discusses in more detail the Unobserved Components framework popularized by Harvey (1984,89).

2 Linear dynamic state-space model

2.1 The models

2.1.1 Weak linear state-space model

The weak form of the linear *dynamic state-space model* is as follows:

$$\boldsymbol{x}_t = \boldsymbol{F}_t \boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{1a}$$

$$\boldsymbol{y}_t = \boldsymbol{H}_t \boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{1b}$$

with the moment restrictions:

$$\begin{aligned} E[\boldsymbol{\epsilon}_t] = \mathbf{0}, \quad & Cov(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-s}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}t} \mathbb{1}_{s=0}, \\ E[\boldsymbol{\eta}_t] = \mathbf{0}, \quad & Cov(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t-s}) = \boldsymbol{\Sigma}_{\boldsymbol{\eta}t} \mathbb{1}_{s=0}, \\ E[\boldsymbol{\epsilon}_{t-j} \boldsymbol{\eta}'_{t-s}] = \mathbf{0}, \quad & \forall j, s \in \mathbb{Z}, \\ \text{and} \quad Cov(\boldsymbol{x}_0, \boldsymbol{\epsilon}_t) = Cov(\boldsymbol{x}_0, \boldsymbol{\eta}_t) = \mathbf{0}, \quad & \forall t > 0, \end{aligned} \tag{2a}$$

where:

- $\boldsymbol{y}_t$ is a $n \times 1$ vector of the observed values at time t.

- $\boldsymbol{x}_t$ is a $p \times 1$ vector of state process values at time t.[1]

- $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are assumed uncorrelated with each other across all time lags, and their covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}t}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_t}$, depend on t.

- $\boldsymbol{F}_t$ is called the "system matrix" and $\boldsymbol{H}_t$ the "observation matrix." These matrices are assumed to be *non-stochastic* where $\boldsymbol{F}_t$ is $p \times p$, and if we allow for more state processes than observed ones, $\boldsymbol{H}_t$ is $n \times p$ where $n \geq p$.

- Equation (1a) is called the "state transition" equation and (1b) is called the "observation" or "measurement" equation.

- The state initial condition, $\boldsymbol{x}_0$, is assumed stochastic with second order moments denoted $E[\boldsymbol{x}_0] = \boldsymbol{\mu}$ and $Var(\boldsymbol{x}_0) = \boldsymbol{\Sigma}_{\boldsymbol{x}_0}$. Finally, there exists zero correlation between the initial state condition and the observation and state error terms, for all dates $t > 0$.

2.1.2 The Gaussian linear state-space model

In the weak version of the linear dynamic state-space model, the assumptions concern only the first and second-order moments of the noise processes and initial state, or equivalently the first and second-order moments of the joint process $[(\boldsymbol{x}_t', \boldsymbol{y}_t')']$. We can also introduce a more restrictive version of the model by assuming, independent, and identically distributed, Gaussian white noises (IIN) for the errors of the state and measurement equations. The Gaussian linear state

[1] Other terminology include the "signal" in the engineering context, "control variable" in the systems control literature, "latent factor" in the factor models approach, or "stochastic parameter" in the Bayesian context. See below for more details.

space model is therefore defined as:

$$\boldsymbol{x}_t = \boldsymbol{F}_t \boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{1a}$$

$$\boldsymbol{y}_t = \boldsymbol{H}_t \boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{1b}$$

$$\text{where} \quad \begin{pmatrix} \boldsymbol{\epsilon}_t \\ \boldsymbol{\eta}_t \end{pmatrix}_{t \geq 1} \sim IIN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon} t} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\eta}_t} \end{bmatrix} \right), \quad \boldsymbol{x}_0 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}_0}), \tag{3a}$$

$$\text{and } \boldsymbol{x}_0 \text{ and the joint process } (\boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t) \text{ are independent.}$$

The Gaussian version of the state-space model is often used as a convenient intermediary tool. Indeed, under the assumption of Gaussian noise and initial state, we know that the joint process $[(\boldsymbol{x}_t', \boldsymbol{y}_t')']$ is Gaussian. This implies that all marginal and conditional distributions concerning the components of these processes are also Gaussian. If the distributions are easily derived, we get as a by-product the expression of the associated linear regressions and residual variances. Since these linear regressions and residual variances are functions of the first and second-order moments only, their expressions are valid even if the noises are not Gaussian–that is, for the weak linear state space model.

2.2 Examples

We will now discuss various examples of the state-space model. The first examples from 2.2.1-2.2.3 are descriptive models used for predicting the future; the second set of examples, 2.2.4-2.2.9 introduces some structure on the dynamics to capture measurement error, missing data, or aggregation. Finally, the last examples, 2.2.10-2.2.12, come from economic and financial applications.

2.2.1 The Vector Autoregressive model of order 1, VAR(1)

The (weak) Vector AutoRegressive model of order 1, $VAR(1)$, is defined as:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{with} \quad \boldsymbol{\epsilon}_t \sim WWN(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \tag{4a}$$

$$\boldsymbol{y}_t = \boldsymbol{x}_t, \tag{4b}$$

with the condition of no correlation between initial state, $\boldsymbol{x}_0$, and the error terms, $\boldsymbol{\epsilon}_t$, satisfied. Furthermore, WWN denotes "weak white noise" or a process with finite, constant, first and second-order moments which exhibits no serial correlation.

In this case the observation process coincides with the state process. This implies that:

$$\boldsymbol{y}_t = \boldsymbol{F}\boldsymbol{y}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{with} \quad \boldsymbol{\epsilon}_t \sim WWN(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \tag{5}$$

which is the standard definition of the (weak) $VAR(1)$ process.

2.2.2 The univariate Autoregressive model of order p, AR(p)

The (weak) univariate AutoRegressive model of order p, $AR(p)$, is defined as:

$$x_t + b_1 x_{t-1} + \cdots + b_p x_{t-p} = \epsilon_t, \quad \text{with} \quad \epsilon_t \sim WWN(0, \sigma_\epsilon^2), \tag{6}$$

The model can be written in state-space form as:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{7a}$$

$$y_t = \boldsymbol{H}\boldsymbol{x}_t. \tag{7b}$$

The state vector includes the current and first $p-1$ lagged values of $\boldsymbol{x}_t$:

$$\boldsymbol{x}_t = \begin{bmatrix} x_t & x_{t-1} & \dots & x_{t-p+1} \end{bmatrix}'_{p\times 1}, \tag{8}$$

with system matrices are given as:

$$\boldsymbol{F} = \begin{bmatrix} -b_1 & -b_2 & \dots & -b_{p-1} & -b_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}_{p \times p}, \tag{9a}$$

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}_{1 \times p}, \tag{9b}$$

$$\text{and} \quad \boldsymbol{\epsilon_t} = \begin{bmatrix} \epsilon_t & 0 & \dots & 0 \end{bmatrix}'_{p \times 1}. \tag{9c}$$

Since the $AR(p)$ process is completely observed, $\boldsymbol{\eta}_t = 0$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}_t} = 0$ for all t. Moreover, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon} t}$ is a singular matrix with zeroes in each element except the top diagonal element, which is equal to σ_ϵ^2 for all t.

2.2.3 The univariate Autoregressive-Moving Average model of order (p, q), ARMA(p,q)

The (weak) univariate AutoRegressive-Moving Average model of order (p, q), $ARMA(p, q)$, is defined as:

$$x_t + b_1 x_{t-1} + \cdots + b_p x_{t-p} = \epsilon_t + a_1 \epsilon_{t-1} + \cdots + a_q \epsilon_{t-q}, \quad \text{where} \quad \epsilon_t \sim WWN(0, \sigma_\epsilon^2), \tag{10}$$

There are a number of possible state-space representations of an ARMA process. In the language of Akaike (1975), a "minimal" representation is a representation whose state vector elements represent the minimum collection of variables which contain all the information needed to produce forecasts given some forecast origin t. For ease of exposition in what follows we provide a non-minimal state-space representation, although the interested reader can consult Gourieroux (1997, p. 607) for a minimal one.

Let the dimension of the state $\boldsymbol{x}_t$ be $m = p + q$. We have:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{G}\boldsymbol{\epsilon}_t, \tag{11a}$$

$$y_t = \boldsymbol{H}\boldsymbol{x}_t, \tag{11b}$$

The state vector is given as:

$$\boldsymbol{x}_t = \begin{bmatrix} x_t & x_{t-1} & \dots & x_{t-(p-2)} & x_{t-(p-1)} & \epsilon_t & \epsilon_{t-1} & \dots & \epsilon_{t-(q-2)} & \epsilon_{t-(q-1)} \end{bmatrix}'. \tag{12a}$$

The system matrices are given as:

$$\boldsymbol{F} = \begin{bmatrix} -b_1 & \dots & \dots & -b_{p-1} & -b_p & a_1 & \dots & \dots & a_{q-1} & a_q \\ 1 & 0 & \dots & 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \dots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \dots & \dots & \vdots & \vdots & 0 & 1 & \dots & 0 & 0 \\ \vdots & \dots & \dots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}_{m \times m}, \tag{13a}$$

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \end{bmatrix}'_{m \times 1}, \tag{13b}$$

$$\boldsymbol{H} = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}_{1 \times m}, \tag{13c}$$

$$\text{and} \quad \boldsymbol{\epsilon_t} \equiv \epsilon_t \sim WWN(0, \sigma_\epsilon^2). \tag{13d}$$

with the condition of no correlation between initial state, $\boldsymbol{x}_0$, and the error terms, $\boldsymbol{\epsilon}_t$, satisfied.

2.2.4 Partially observed VARMA

In the discrete-time multivariate case, suppose that some linear, unobserved, state process $\boldsymbol{x}_t$ is composed of filtered weak white noise $\boldsymbol{\epsilon}_t$. However, what we actually observe is $\boldsymbol{y}_t$, which has been corrupted by additive weak white noise $\boldsymbol{\eta}_t$. This model can be written as:

$$\boldsymbol{x}_t = \sum_{u=0}^{\infty} \boldsymbol{a}_u \boldsymbol{\epsilon}_{t-u} = \boldsymbol{A}(L)\boldsymbol{\epsilon}_t, \tag{14a}$$

$$\text{and} \quad \boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{\eta}_t, \quad \text{for} \quad t = 1, \ldots, T, \tag{14b}$$

where:

- $\boldsymbol{x}_t$ is an unobserved $P \times 1$ vector of state variables.
- the unobserved state $\boldsymbol{x}_t$ (the "signal" in the engineering context) is corrupted with weak, white, additive noise, $\boldsymbol{\eta}_t$, with covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.
- $\boldsymbol{\epsilon}_t$ is an $M \times 1$ vector of weak white noise input processes with covariance $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$. [2]
- $\boldsymbol{y}_t$ is a $P \times 1$ vector of the observed "noisy" output process.
- $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ represent the covariance of the measurement noise and the state process noise, respectively, and are assumed time-invariant.

Note that $\boldsymbol{A}(L)$ is a $P \times M$ matrix infinite series where L denotes the lag operator; that is, $\boldsymbol{A}(L) = \boldsymbol{a}_0 L^0 + \boldsymbol{a}_1 L^1 + \boldsymbol{a}_2 L^2 + \ldots$, where the individual $P \times M$ matrices, $\boldsymbol{a}_u$, collectively represent the impulse response function f where $f : \mathbb{Z} \to \mathbb{R}^{P \times M}$.

Given (14a), the infinite lag distribution makes working with this model in the time domain troublesome. However, the apparent multi-stage dependence can be reduced to first-order autoregressive dependence by means of a matrix representation if we assume that $\boldsymbol{A}(L)$ can be well-approximated by a ratio of finite lag matrix polynomials – the so called "transfer function"

[2] Note there is no loss of generality here since for any (possibly non-white) second-order stationary stochastic input process we might choose, $\boldsymbol{z}_t$, we can always represent it as $\boldsymbol{z}_t = \sum_{j=0}^{\infty} \boldsymbol{\Theta}_j \boldsymbol{\epsilon}_{t-j}$ by Wold's theorem.

models of Box & Jenkins (1970). That is, suppose that we model instead:

$$\begin{aligned}\boldsymbol{C}(L) &= \boldsymbol{B}(L)^{-1}\boldsymbol{A}(L) \\ &= \left(\boldsymbol{I} + \boldsymbol{b}_1 + \boldsymbol{b}_2 L^2 + \cdots + \boldsymbol{b}_n L^n)\right)^{-1}\left(\boldsymbol{a}_0 + \boldsymbol{a}_1 + \boldsymbol{a}_2 L^2 + \cdots + \boldsymbol{a}_{n-1} L^{n-1})\right) \quad (15)\end{aligned}$$

where the inverse of the matrix lag polynomial $\boldsymbol{B}(L)$ is assumed to exist.

The model in (14), with the ratio of finite order matrix lag polynomials $\boldsymbol{C}(L)$ now replacing the infinite series $\boldsymbol{A}(L)$, becomes:

$$\boldsymbol{x}_t + \boldsymbol{b}_1\boldsymbol{x}_{t-1}, \ldots, +\boldsymbol{b}_n\boldsymbol{x}_{t-n} = \boldsymbol{a}_0\boldsymbol{\epsilon}_t +, \ldots, \boldsymbol{a}_{n-1}\boldsymbol{\epsilon}_{t-(n-1)}, \quad \forall t = 1, \ldots, T, \tag{16a}$$

$$\text{and} \quad \boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{16b}$$

(16) can now be reduced to first-order dependence by means of a redefinition of the state vector as:

$$\boldsymbol{x}_t^* = \boldsymbol{T}\hat{\boldsymbol{x}}_{t-1}, \tag{17a}$$

$$\text{where} \quad \boldsymbol{T} = \begin{bmatrix} -\boldsymbol{b}_n & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ -\boldsymbol{b}_{n-1} & -\boldsymbol{b}_n & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ -\boldsymbol{b}_{n-2} & -\boldsymbol{b}_{n-1} & -\boldsymbol{b}_n & \ldots & \boldsymbol{0} \\ \vdots & & & \ddots & \boldsymbol{0} \\ -\boldsymbol{b}_1 & -\boldsymbol{b}_2 & -\boldsymbol{b}_3 & \ldots & -\boldsymbol{b}_n \end{bmatrix}_{Pn \times Pn}, \tag{17b}$$

$$\text{and} \quad \hat{\boldsymbol{x}}'_{t-1} = \begin{bmatrix} \boldsymbol{x}'_{t-1} & \boldsymbol{x}'_{t-2} & \ldots & \boldsymbol{x}'_{t-n} \end{bmatrix}_{1 \times Pn}, \tag{17c}$$

so that the new first-order autoregressive state-space model takes the form:

$$\boldsymbol{x}_t^* = \boldsymbol{F}\boldsymbol{x}_{t-1}^* + \boldsymbol{G}\hat{\boldsymbol{\epsilon}}_t, \tag{18a}$$

$$\text{and} \quad \boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t^* + \boldsymbol{\eta}_t \tag{18b}$$

where we have that:

$$\boldsymbol{F} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & -\boldsymbol{b}_n \\ \boldsymbol{I} & \mathbf{0} & \dots & \mathbf{0} & -\boldsymbol{b}_{n-1} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{I} & -\boldsymbol{b}_1 \end{bmatrix}_{Pn \times Pn}, \tag{19a}$$

$$\boldsymbol{G} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{a}_0 & \boldsymbol{a}_1 & \dots & \boldsymbol{a}_{n-1} \end{bmatrix}_{Pn \times Mn}, \tag{19b}$$

$$\boldsymbol{H} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{I} \end{bmatrix}_{P \times Pn}, \tag{19c}$$

$$\text{and} \quad \hat{\boldsymbol{\epsilon}}_t' = \begin{bmatrix} \boldsymbol{\epsilon}_t' & \boldsymbol{\epsilon}_{t-1}' & \dots & \boldsymbol{\epsilon}_{t-(n-1)}' \end{bmatrix}_{1 \times Mn} \tag{19d}$$

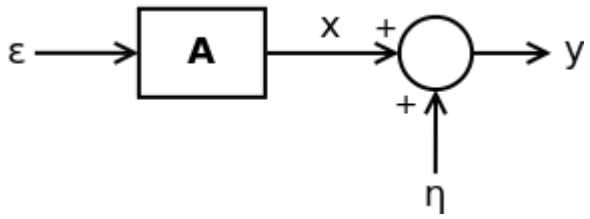
where again $\boldsymbol{F}$ is the "system matrix," $\boldsymbol{G}$ the "input matrix," and $\boldsymbol{H}$ the "observation matrix." Bear in mind that (18) and (19) represent only one possible state-space representation – in fact while the transfer function $\boldsymbol{C}(e^{-i\omega})$ in $\boldsymbol{h}_{xx}(\omega)$ (see Section 2.2.4, (20a)) implies an infinite number of possible state-space representations, any particular state-space representation has only one equivalent transfer function. Additionally, we can immediately see from (18b) that the observed process $\boldsymbol{y}_t$ is a linear function of the unobserved "factors" since $\boldsymbol{y}_t = -\boldsymbol{b}_1\boldsymbol{x}_{t-1} - \boldsymbol{b}_2\boldsymbol{x}_{t-2} - \cdots - \boldsymbol{b}_n\boldsymbol{x}_{t-n} + \boldsymbol{u}_t + \boldsymbol{\eta}_t$, where $\boldsymbol{u}_t$ is equal to the right-hand side of (16a). See Akaike (1974) for a general treatment of finite order linear systems.

i) Spectral properties of the partially observed VARMA process

Note that from (14a), $\boldsymbol{A}(L) = \sum_{u=0}^{\infty} \boldsymbol{a}_u L^u$, so $\boldsymbol{A}(L)\boldsymbol{\epsilon}_t = \sum_{u=0}^{\infty} \boldsymbol{a}_u \boldsymbol{\epsilon}_{t-u}$ More generally, $\boldsymbol{A}(z) = \sum_{u=0}^{\infty} \boldsymbol{a}_u z^u$ where $z \in \mathbb{C}$ is known as the z-transform. Therefore, while $\boldsymbol{A}(L)$ is a polynomial function of an *operator* L, the z-transform, $\boldsymbol{A}(z)$, is a polynomial function of a complex variable. However, since both polynomials admit the same coefficients, we can solve for the *transfer function* of (14a) as $\boldsymbol{A}(z)$ where $z = e^{-i\omega}$, since this represents the Fourier

transform of the impulse response function.[3] (Note that in continuous time this z-transform analogy is unnecessary since there is no need for defining the model in terms of lag operators, L). Therefore, the convolution observed in the time domain in (14a) is equivalent to a multiplication within the frequency domain, so that the Fourier transform of the impulse response, $\boldsymbol{A}(e^{-i\omega})$, disentangles the complicated interdepencies into a simple multiplicative relation between inputs and outputs given any frequency ω. Therefore, working with (14) in the frequency domain is often a useful approach. For clarity the frequency domain relationships are given diagramatically in Figure 1.

Figure 1: Frequency domain relationships of the model in (14)



Since $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are jointly stationary and uncorrelated we have that:

$$\boldsymbol{h}_{yx}(\omega) = \boldsymbol{A}(e^{-i\omega})\boldsymbol{h}_{\epsilon\epsilon}(\omega)\boldsymbol{A}(e^{+i\omega})' = \frac{1}{2\pi}\boldsymbol{A}(e^{-i\omega})\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{A}(e^{+i\omega})' = \boldsymbol{h}_{xx}(\omega), \tag{20a}$$

$$\text{and} \quad \boldsymbol{h}_{y\epsilon}(\omega) = \boldsymbol{A}(e^{-i\omega})\boldsymbol{h}_{\epsilon\epsilon}(\omega) = \frac{1}{2\pi}\boldsymbol{A}(e^{-i\omega})\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \tag{20b}$$

represent the cross-spectral density matrices between $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ and $\boldsymbol{\epsilon}_t$ respectively. Therefore, from (20a) it is clear that $\boldsymbol{x}_t$ represents "filtered" weak white noise, where the flat spectrum of $\boldsymbol{\epsilon}_t$ (i.e. its variance) is given shape by $\boldsymbol{A}(e^{\pm i\omega})$.

Furthermore, the spectral-density matrix of $\boldsymbol{y}_t$ is (from (16)):

$$\boldsymbol{h}_{yy}(\omega) = \boldsymbol{h}_{xx}(\omega) + \boldsymbol{h}_{\eta\eta}(\omega) = \frac{1}{2\pi}\left(\boldsymbol{A}(e^{-i\omega})\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{A}(e^{+i\omega})' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right). \tag{21}$$

[3]The system in (14) is constrained to be "physically realizable" by assuming the impulse response matrices are $\boldsymbol{a}_j = 0, \ \forall j < 0$. This form of impulse response exists, is unique, and is quadratically summable, with no zeros inside the unit circle as long as the integral from $-\pi$ to π of the log of ϵ_t's spectral density is finite – see Doob (1953), as cited in Priesley (1981, pg. 733). Note this condition is a very weak one and is satisfied here – in fact, the mentioned integral can only diverge to $-\infty$ if the spectral density vanishes in some interval in its domain.

Note that for the discrete time process all spectral densities are continuous in the frequency ω and periodic with period 2π.

Finally, a natural non-parametric estimator of the transfer function matrix is given by:

$$\hat{\boldsymbol{A}}(e^{-i\omega}) = \hat{\boldsymbol{h}_{y\epsilon}}(\omega)\hat{\boldsymbol{h}_{\epsilon\epsilon}^{-1}}(\omega) = 2\pi\hat{\boldsymbol{h}_{y\epsilon}}(\omega)\widehat{\boldsymbol{\Sigma_{\epsilon}}^{-1}} \tag{22}$$

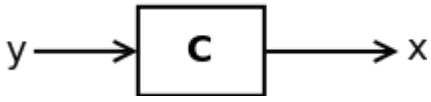
where the spectral densities in (22) can be estimated within the frequency domain. See Priestley (1981, Section 9.5) for more details.

Now, suppose we wish to establish the optimal manner of extracting the signal $\boldsymbol{x}_t$ given only the noisy observations $\boldsymbol{y}_t$. That is, we wish to establish the optimal frequency response, or transfer function $\boldsymbol{C}(\omega)$, in Figure 2. It was Wiener that original solved this frequency domain problem where he established the optimal frequency response as the ratio: [4]

$$\boldsymbol{C}(\omega) = \frac{\boldsymbol{h}_{xy}(\omega)}{\boldsymbol{h}_{yy}(\omega)} = \frac{\boldsymbol{h}_{xx}(\omega)}{\boldsymbol{h}_{xx}(\omega) + \boldsymbol{h}_{\eta\eta}(\omega)}. \tag{23}$$

Therefore, the Wiener filter attenuates those frequencies in which the signal to noise ratio is low and passes through those where it is high.

Figure 2: Wiener filter - the optimal transfer function $\boldsymbol{C}(\omega)$



[4]Noting of course that since $E[\boldsymbol{x}_t\boldsymbol{x}'_{t-s}]$ is symmetric in the time domain for all s, we have that $\boldsymbol{h}_{xx}(\omega)$ is real, and so $\boldsymbol{h}_{xy}(\omega) = \boldsymbol{h}_{xx}(\omega) = \boldsymbol{h}_{yx}(\omega)$ without the need of taking complex conjugates.

2.2.5 The VAR(1) with measurement error

A special case of the partially observed VARMA model in Section 2.2.4 arises as the (weak) VAR(1) with measurement error is defined as:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{24a}$$

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{24b}$$

$$\begin{pmatrix} \boldsymbol{\epsilon}_t \\ \boldsymbol{\eta}_t \end{pmatrix}_{t\geq 1} \sim WWN\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \end{bmatrix} \right), \tag{24c}$$

with the condition of no correlation between initial state, $\boldsymbol{x}_0$, and the error terms process, $(\boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t)$, satisfied.

Therefore, the state-space process is a VAR(1) process, but measured with a multivariate error given by $\boldsymbol{\eta}_t$. The process $(\boldsymbol{y}_t)$ is such that:

$$\begin{aligned} \boldsymbol{y}_t - \boldsymbol{F}\boldsymbol{y}_{t-1} &= \boldsymbol{x}_t + \boldsymbol{\eta}_t - \boldsymbol{F}\left(\boldsymbol{x}_{t-1} + \boldsymbol{\eta}_{t-1}\right) \\ &= \boldsymbol{\epsilon}_t + \boldsymbol{\eta}_t - \boldsymbol{F}\boldsymbol{\eta}_{t-1} \equiv \boldsymbol{v}_t \end{aligned} \tag{25a}$$

The process $(\boldsymbol{v}_t)$ has serial covariances equal to zero for lags larger or equal to 2. Therefore, $\boldsymbol{v}_t$ admits a Vector Moving Average, VMA(1), representation of order 1. Let $\boldsymbol{v}_t = \boldsymbol{u}_t - \boldsymbol{\Theta}\boldsymbol{u}_{t-1}$.[5] We can therefore deduce that the process $\boldsymbol{y}_t$ has a Vector Autoregressive-Moving Average of order (1,1), or VARMA(1,1), representation:

$$\boldsymbol{y}_t - \boldsymbol{F}\boldsymbol{y}_{t-1} = \boldsymbol{u}_t - \boldsymbol{\Theta}\boldsymbol{u}_{t-1}, \quad \text{with} \quad \boldsymbol{u}_t \sim WWN(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{u}}), \tag{26}$$

$\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ are functions of the intial parameters of the state-space representation: $\boldsymbol{F}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, and

[5] And so we have that there exists no correlation between initial state, $\boldsymbol{x}_0$, and the new error terms process, $(\boldsymbol{u}_t)$.

$\boldsymbol{\Sigma_\eta}$. They are related by the matrix equations system:

$$-\boldsymbol{F}\boldsymbol{\Sigma_\eta} = \boldsymbol{\Theta}\boldsymbol{\Sigma_u} \tag{27a}$$

$$\boldsymbol{\Sigma_\eta} + \boldsymbol{F}\boldsymbol{\Sigma_\eta}\boldsymbol{F}' + \boldsymbol{\Sigma_\epsilon} = \boldsymbol{\Sigma_u} + \boldsymbol{\Theta}\boldsymbol{\Sigma_u}\boldsymbol{\Theta}' \tag{27b}$$

which can be solved numerically for values of $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma_u}$.

2.2.6 Static state-space model

The (weak) static state-space model is defined as:

$$\boldsymbol{x}_t = \boldsymbol{\epsilon}_t, \tag{28a}$$

$$\boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{28b}$$

$$\begin{pmatrix} \boldsymbol{\epsilon}_t \\ \boldsymbol{\eta}_t \end{pmatrix}_{t\geq 1} \sim WWN\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\Sigma_\epsilon} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma_\eta} \end{bmatrix} \right), \tag{28c}$$

with the condition of no correlation between initial state, $\boldsymbol{x}_0$, and the error terms, $(\boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t)$, satisfied.

Therefore, from (28) the distribution of the state-space process is such that:

$$\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y_t} \end{bmatrix}_{t\geq 1} = WWN\left[\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} \quad , \quad \begin{pmatrix} \boldsymbol{\Sigma_\epsilon} & \boldsymbol{\Sigma_\epsilon}\boldsymbol{H}' \\ \boldsymbol{H}\boldsymbol{\Sigma_\epsilon} & \boldsymbol{H}\boldsymbol{\Sigma_\epsilon}\boldsymbol{H}' + \boldsymbol{\Sigma_\eta} \end{pmatrix} \right]. \tag{29}$$

In general, the state-space form is equivalent to the *factor model* representation, where we assume that some p factors, $\boldsymbol{x}_t$, influence the n observed processes $\boldsymbol{y}_t$, where $n > p$. Indeed, the goal of the factor model representation is to model the observed processes in terms of a smaller number of factor processes. Therefore, the particular form of the state-space model in (28) is equivalent to the *static* factor model representation, although it is clear that the factor may

instead be formulated in a dynamic manner as in (1a).

We can distinguish two cases in practice:

- Case 1: the factor $\boldsymbol{x}_t$ is unobserved.

  In this case, the unrestricted theoretical model above is unidentifiable. This is because the number of parameters exceeds the number of population moment conditions when $\boldsymbol{x}_t$ is unobserved. Indeed, from (29), we have:

$$Var(\boldsymbol{y}_t)_{n\times n} = \boldsymbol{H}_{n\times p}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}p\times p}\boldsymbol{H}'_{p\times n} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}_{n\times n}}, \tag{30}$$

  and so the $n(n+1)/2$ population second moment conditions are outnumbered by the $np + p(p+1)/2 + n(n+1)/2$ parameters.

  Therefore, moment contraints on the parameters of the factor model are usually introduced to ensure identification. For example, we can assume without loss of generality that the factor covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, is an identity matrix. Indeed, $\boldsymbol{x}_t$ is unobserved and defined up to an invertible linear transform. That is, for any invertible matrix $\boldsymbol{L}$ equation (28b) with $\boldsymbol{x}_t^* = \boldsymbol{L}\boldsymbol{x}_t$ and $\boldsymbol{H}^* = \boldsymbol{H}\boldsymbol{L}^{-1}$ is observationally equivalent. Therefore, for any $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ we can always introduce the transformation $\boldsymbol{x}_t^* = \boldsymbol{D}^{-1/2}\boldsymbol{P}'\boldsymbol{x}_t$, where the matrix $\boldsymbol{P}$ has an orthonormal basis of eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ as its columns and $\boldsymbol{D}$ is a diagonal matrix of the respective eigenvalues, so as to make $Var(\boldsymbol{x}_t^*) = \boldsymbol{I}$. Moreover, it is often assumed in the literature that the observation error covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, is diagonal (or even scalar) so that $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \sigma_\eta^2\boldsymbol{I}$. This additional constraint is considered a "real" constraint since it reduces the model's flexibility in favour of identification.

- Case 2: the factor $\boldsymbol{x}_t$ is observed.

  In this, case the unrestricted theoretical model becomes identifiable. Given the moment condition $Var(\boldsymbol{x}_t)$, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is identified. We can then formulate the theoretical linear regression in (28b) to identify both $\boldsymbol{H}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.

The static factor model is popular in Finance, where the observed variable $\boldsymbol{y}_t$ represents the returns of a set of assets. Under a (weak) efficients market hypothesis the returns are WWN. The observation equation (28b) thus decomposes the returns into a market component, $\boldsymbol{H}\boldsymbol{x}_t$, and a firm specific component, $\boldsymbol{\eta}_t$. Assuming an uncorrelated market component, the unobserved factors, $\boldsymbol{x}_t$, represent the returns on the market and $\boldsymbol{\eta}_t$ represent the firm specific returns, whose variability (or "indiosyncratic" risk) can be reduced through its addition to a well diversied portfolio. Of course, the assumption of uncorrelated market component can be generalized within the dynamic model. For more on the factor model representation, see Section 9.

2.2.7 The state-space model for "data aggregation"

Suppose that we assume the state vector $\boldsymbol{x}_t$ represents some individual level components which we desire to aggregate in some way. In a model for aggregation you have to distinguish between both the behavioural equation which generally includes an error term, and the accounting relationship with no error term.

Therefore, let $\boldsymbol{y}_t$ represent the observed aggregate variable, and let $\boldsymbol{x}_t$ represent some possibly unobserved individual level variables. The state-space formulation defines both the behavioural equation for $\boldsymbol{x}_t$ and the accounting equation for $\boldsymbol{y}_t$ as:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim WWN(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \tag{31a}$$

$$y_t = \boldsymbol{\alpha}'\boldsymbol{x}_t, \tag{31b}$$

where $\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_p \end{bmatrix}'$ is a p vector of size adjustment parameters (which may possibly sum to 1) and so we can model the observed values y_t as the weighted aggregate of individual factors, the elements of $\boldsymbol{x}_t$.

Note that in an accounting relationship you can only add variables with the same unit. Therefore, we have first to transform the elements of $\boldsymbol{x}_t$ into a common unit, which is usually done by considering some measure of value in common units, e.g. dollars.

The aggregation model can also be employed in the Finance context. In this case, as opposed to Section 2.2.6, the observation equation represents an accounting relationship between asset returns and the aggregate portfolio return, not a behavioural relationship. The returns may be weighted according to their contribution to the overall portfolio, where again the returns are written in the same domination, e.g. dollars.

2.2.8 The VAR(1) with "series selection" or "missing series"

The Vector AutoRegressive process of order 1, or VAR(1), with "series selection" or "missing series" is defined as:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim WWN(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \tag{32a}$$

$$\text{and} \quad y_t = x_{i,t}, \tag{32b}$$

where $x_{i,t}$ denotes the i'th element of $\boldsymbol{x}_t$. Therefore the model can be interpreted in two ways, depending on whether or not $\boldsymbol{x}_t$ is observed:

- Case 1: $\boldsymbol{x}_t$ is observed.

  The model is then interpreted as a method of selecting only that series from the state vector $\boldsymbol{x}_t$ that is of interest. Notice that (31b) above is a special case of series selection, when $\boldsymbol{x}_t$ is observed.

- Case 2: $\boldsymbol{x}_t$ is not observed.

  The model is interpreted as the case of "missing series." That is, some of the elements of the series $(\boldsymbol{x}_t)$ are missing.

2.2.9 The VAR(1) with "missing data"

The model in Section 2.2.8, with unobserved $\boldsymbol{x}_t$, can of course be generalized to the cases where not only are series missing, but perhaps individual data elements of some series are missing as

well. We call this the vector autoregressive process of order 1, or VAR(1), for "missing data". The state equation is the same for all cases below:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{\epsilon}_t \sim WWN(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}). \tag{32a}$$

- Case 1: the i'th series is missing and some elements of the j'th series are missing:

$$\boldsymbol{y}_t = \begin{cases} \begin{bmatrix} x_{1,t} & \dots & x_{i-1,t} & x_{i+1,t} & \dots & x_{j,t} & \dots & x_{p,t} \end{bmatrix}' & \text{if} \quad t \neq m \\ \begin{bmatrix} x_{1,t} & \dots & x_{i-1,t} & x_{i+1,t} & \dots & x_{j-1,t} & x_{j+1,t} & \dots & x_{p,t} \end{bmatrix}' & \text{if} \quad t = m \end{cases}. \tag{33}$$

- Case 2: the i'th and j'th series both have missing data but the missing points occur at the same time:

$$\boldsymbol{y}_t = \begin{cases} \boldsymbol{x}_t & \text{if} \quad t \neq m \\ \begin{bmatrix} x_{1,t} & \dots & x_{i-1,t} & x_{i+1,t} & \dots & x_{j-1,t} & x_{j+1,t} & \dots & x_{p,t} \end{bmatrix}' & \text{if} \quad t = m \end{cases}. \tag{34}$$

- Case 3: the i'th and j'th series both have missing data with no inherent pattern.

Where in each case, $m \in 0, \dots, T$, denotes a time period upon which some elements of the vector $\boldsymbol{x}_t$ are missing.

2.2.10 The Unobserved Components model

Consider the special case of the state-space model for aggregation in Section 2.2.7, where the elements of $\boldsymbol{\alpha}$ are all equal to one, and we assume that the p elements of $\boldsymbol{x}_t$ are independent of each other with specified marginal distributions or we at least specify their first two moments:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{35a}$$

$$y_t = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \boldsymbol{x}_t, \tag{35b}$$

the observed series $\boldsymbol{y}_t$ is therefore the sum of various *components*, generally unobserved.

2.2.10.1 The General Stochastic Trend

P.C. Young (2011, p.67) defines the *generalized random walk* or general *stochastic trend* as:

$$\boldsymbol{x}_t = \begin{bmatrix} x_{1,t} \\ \Delta x_{1,t} \end{bmatrix}, \tag{36a}$$

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{G}\boldsymbol{\epsilon}_t, \quad \text{where} \quad \boldsymbol{F} = \begin{bmatrix} \alpha & \beta \\ 0 & \gamma \end{bmatrix}, \quad \text{and} \begin{bmatrix} \delta & 0 \\ 0 & \epsilon \end{bmatrix}. \tag{36b}$$

$$\text{Also} \quad \boldsymbol{\epsilon}_t \sim WWN(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \quad \text{where} \quad \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \quad \text{is diagonal.} \tag{36c}$$

That is, we have defined the state process in such a manner as to allow us to modify the behaviour of the typical random walk in different ways. For example, if $\beta = \gamma = \epsilon = 0$ and $\alpha = \delta = 1$ the model represents the standard random walk. However, if $\alpha = \beta = \gamma = \epsilon = 1$ and $\delta = 0$ we have the *integrated random walk* which is smoother than the standard random walk. Moreover, if $0 < \alpha < 1$ and $\beta = \gamma = \epsilon = 1$ and $\delta = 0$ we have the case of the *smoothed random walk*. Also the case of $\beta = \gamma = \epsilon = 0$ and $0 < \alpha < 1$ and $\delta = 1$ is equivalent to the AR(1) model. Finally, both the *Local Linear Trend* (see Section 2.2.10.2) and *Damped Trend* from Harvey (1984,89) are both given by $\alpha = \beta = \gamma = \epsilon = \delta = 1$ (except in the latter case $0 < \gamma < 1$).

2.2.10.2 Harvey's Unobserved Components models: the "basic structural model"

Harvey (1984,89) attempts to decompose the series $(\boldsymbol{y}_t)$ into a number of unobserved, orthogonal, components representing *trends, seasonals, other cycles, and irregular patterns*, all of which are

informed by the spectral properties of the observed series. For example, consider the model:

$$y_t = T_t + S_t + C_t + I_t \tag{37}$$

where y_t is the observed series, T_t is some trend component, S_t is a seasonal component, C_t is some other cyclical component, and I_t represents the irregular pattern.

Typically the trend component T_t is associated with the slowly changing, low frequency component of y_t (i.e. a spectral frequency close to zero, or equivalently a period close to ∞). It can be modeled by the stochastic counterpart of the linear time trend $\mu_t = \mu_0 + \beta t$, called the *Local Linear Trend* model:

$$T_t \equiv \mu_t = \mu_{t-1} + \beta_{t-1} + v_t, \qquad \text{where} \quad v_t \sim WWN(0, \sigma_v^2), \tag{38a}$$

$$\text{and} \quad \beta_t = \beta_{t-1} + z_t, \qquad \text{where} \quad z_t \sim WWN(0, \sigma_z^2). \tag{38b}$$

Of course, the Local Linear Trend formulation is a special case of the general stochastic trend in Section 2.2.10.1.

Furthermore, the seasonal component S_t can be modeled as dummy intercepts which are constrained to sum to zero (with some small stochastic residual difference, ω). For example, suppose s is the number of "seasons" (say 12 for monthly data) and $z_{j,t}$ for $j = 1, 2, \ldots, s$ is some set of dummy variables that take on the values:

$$z_{j,t} = \begin{cases} 1, & \text{if } t = j, j+s, j+2s, \ldots \\ 0, & \text{if } t \neq j, j+s, j+2s, \ldots \\ -1, & \text{if } t = s, 2s, 3s, \ldots \end{cases} \tag{39}$$

then, if γ_j is the dummy intercept for time j, we have that at $t = xs$ for all $x \in \mathbb{N}^+$:

$$\sum_{j=1}^{s-1} z_{j,t}\gamma_j = -\sum_{j=1}^{s-1} \gamma_j \equiv \gamma_s \tag{40a}$$

$$\Leftrightarrow \sum_{j=1}^{s} \gamma_j = 0 \tag{40b}$$

and given a change in the notation, (40b) can be rewritten as $\sum_{j=0}^{s-1} \gamma_{t-j} = 0$. Adding a disturbance term with zero expectation to the right hand side allows the seasonal effect to vary stochastically:

$$\sum_{j=0}^{s-1} \gamma_{t-j} = \omega_t \quad \text{where} \quad \omega_t \sim WWN(0, \sigma_\omega^2), \tag{41a}$$

$$\Leftrightarrow \left(1 + L + L^2 + \cdots + L^{s-1}\right) \gamma_t = \omega_t. \tag{41b}$$

Finally, the cyclical component C_t can be written as a sum of *stochastic harmonics*, where each component in the sum reflects some particular chosen frequency, $\lambda_j = 2\pi\frac{j}{s}$, where $j \leq \frac{s}{2}$. For example, given monthly data, let s be such that $s \pmod{12} = 0$, and let $j \in \mathbb{N}^+$ be chosen so that s/j represents the desired periodicity of the harmonic function. Therefore, we could choose that $s = 12$ and $j = 1$ so that the period is 12; that is, the cycle repeats every 12 months. Alternatively, if $j = 6$ then the period is 2 and the cycle repeats every 2 months, etc.

The cyclical component C_t can therefore be written as:

$$C_t \equiv \sum_{k \in J} c_{k,t} \tag{42a}$$

$$\text{where} \quad c_{k,t} = \rho_k\{c_{k,t-1}\cos\lambda_k + c^*_{k,t-1}\sin\lambda_k\} + \xi_{k,t}, \tag{42b}$$

$$\text{and} \quad c^*_{k,t} = \rho_k\{-c_{k,t-1}\sin\lambda_k + c^*_{k,t-1}\cos\lambda_k\} + \xi^*_{k,t}, \tag{42c}$$

where J is the set of chosen frequencies, ρ_k is a discount parameter, and ξ_k and ξ^*_k are zero mean WWN processes which are uncorrelated with each other, with common variance $\sigma^2_{\xi,k}$. For more

details on the stochastic harmonic cycles approach, see Hannan, Terrell and Tuckwell (1970).

Finally, the irregular component takes the form of a WWN innovation, $I_t \equiv \eta_t$. Putting all the components together into the state-space form with WWN innovations, we have the observation equation

$$\begin{aligned} y_t &= T_t + S_t + C_t + I_t \\ \Leftrightarrow y_t &= \mu_t + \gamma_t + \sum_{k \in J} c_{k,t} + \eta_t \\ &= \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & \dots & 1 & 0 & 1 & \dots \end{bmatrix} \boldsymbol{x}_t + \eta_t \\ &\equiv \boldsymbol{H}_t \boldsymbol{x}_t + \boldsymbol{\eta}_t, \end{aligned} \tag{43a}$$

and the state transition equation

$$\boldsymbol{x}_t = \begin{bmatrix} \mu_t \\ \beta_t \\ \gamma_t \\ \gamma_{t-1} \\ \vdots \\ c_{1,t} \\ c^*_{1,t} \\ c_{2,t} \\ c^*_{2,t} \\ \vdots \end{bmatrix} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \\ \boldsymbol{0} & \boldsymbol{S} & \boldsymbol{0} & \boldsymbol{0} & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{C}_1 & \boldsymbol{0} & \dots \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{C}_2 & \\ & & \vdots & & \ddots \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \beta_{t-1} \\ \gamma_{t-1} \\ \gamma_{t-2} \\ \vdots \\ c_{1,t-1} \\ c^*_{1,t-1} \\ c_{2,t-1} \\ c^*_{2,t-1} \\ \vdots \end{bmatrix} + \begin{bmatrix} v_t \\ z_t \\ \omega_t \\ 0 \\ \vdots \\ \xi_{1,t} \\ \xi^*_{1,t} \\ \xi_{2,t} \\ \xi^*_{2,t} \\ \vdots \end{bmatrix} \tag{44a}$$

$$\equiv \boldsymbol{F} \boldsymbol{x}_{t-1} + \boldsymbol{\epsilon_t}, \tag{44b}$$

such that

$$\boldsymbol{T} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

$$\boldsymbol{S} = \begin{bmatrix} -1 & -1 & -1 & -1 & \dots \\ 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots & \end{bmatrix},$$

$$\text{and} \quad \boldsymbol{C}_i = \begin{bmatrix} \cos\lambda_i & \sin\lambda_i \\ -\sin\lambda_i & \cos\lambda_i \end{bmatrix}.$$

This state-space representation is known in Harvey (1989, pg.172) as the *Basic Structural Model*.

2.2.11 The CAPM

Another example of the state-space modeling framework is the capital asset pricing model (CAPM) with time-varying coefficients.

Recall that the assumptions of the CAPM model imply that all investments should offer the same reward-to-risk ratio. If the ratio were better for one investment than another, investors would rearrange their portfolios towards the alternative featuring a better tradeoff. Such activity would put pressure on security prices until the ratios were equalized. Within the context of the CAPM, this ratio is known as the "Sharpe ratio" in honor of his pioneering work (Sharpe, 1966) and is defined in terms of excess returns over covariance:

$$\frac{E[R] - R_f}{Cov(R, R_m)} = \frac{E[R_m] - R_f}{\sigma_m^2}. \tag{46}$$

Of course, the Sharpe ratio directly implies a linear relationship between a) the covariance of

an asset's return with the market return; and b) the expected value of the asset's return itself:

$$E[R] - R_f = \frac{Cov(R, R_m)}{\sigma_m^2} \left(E[R_m] - R_f\right) = \beta \left(E[R_m] - R_f\right). \tag{47}$$

However, since it is clear that in the real world the assumptions of the CAPM may hold only approximately, some assets may deviate systematically from the Sharpe ratio relationship, by some amount α:

$$E[R] - R_f = \alpha + \beta \left(E[R_m] - R_f\right). \tag{48}$$

Moreover, each individual asset will be exposed to some form of indiosyncratic "micro" level risk, v, independent of what happens in the market as a whole. It is in fact this indiosyncratic risk that is minimized through the process of diversification. Therefore, we write:

$$E[r] = \alpha + \beta E[r_m] + v \tag{49}$$

where $r \equiv R - R_f$ is the observed excess return on some asset beyond the risk free rate, and $r_m \equiv R_m - R_f$ is the excess return on some market index (assumed to be completely diversified, so that it is orthogonal to the innovation or "indiosyncratic," firm specific risk, v).

Therefore, we can treat the state transition equation as driving the dynamics of the *stochastic parameters* of the model, α_t and β_t. For example, consider the following model, given observations on r_t and $r_{m,t}$ for some $t = 1, \ldots, T$ (which represents a linear regression with unobserved stochastic coefficients):

$$r_t = \alpha_t + \beta_t r_{m,t} + v_t, \qquad \text{where} \quad v_t \sim N(0, \sigma_v^2), \tag{50a}$$

$$\alpha_t = \gamma \alpha_{t-1} + u_t, \qquad \text{where} \quad u_t \sim N(0, \sigma_u^2), \tag{50b}$$

$$\text{and} \quad \beta_t = \mu + \delta \beta_{t-1} + z_t, \qquad \text{where} \quad z_t \sim N(0, \sigma_z^2). \tag{50c}$$

Note that the nature of the equilibrium in the CAPM model suggests some reasonable restrictions on the dynamics of the stochastic parameters, α_t and β_t. First, it is safe to assume that $\frac{\mu}{1-\delta}$

will likely take on some value relatively close to, but not equal to, 1 and will depend directly on the long-run historical covariance between the asset and market returns. Moreover, δ and γ should take on values in the range $0 < x < 1$. That is, they should exhibit *mean reverting* behaviour since in the case of α_t, it is clear that arbitrage opportunities should eventually push α towards zero; and with β_t the relation between r and r_m in (47) should certainly be a bounded one.

Finally, the model in (50) can easily be put into state-space form:

$$\boldsymbol{x}_t \equiv \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} 0 \\ \mu \end{bmatrix} + \begin{bmatrix} \gamma & 0 \\ 0 & \delta \end{bmatrix} \begin{bmatrix} \alpha_{t-1} \\ \beta_{t-1} \end{bmatrix} + \begin{bmatrix} u_t \\ z_t \end{bmatrix} \equiv \boldsymbol{c} + \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{51a}$$

$$\text{and} \quad \boldsymbol{y}_t \equiv r_t = \begin{bmatrix} 1 & r_{m,t} \end{bmatrix} \begin{bmatrix} \alpha_t \\ \beta_t \end{bmatrix} + v_t \equiv \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{51b}$$

where the covariance of the state transition equation is $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \equiv \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_z^2 \end{bmatrix}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} \equiv \sigma_v^2$ is scalar. Given such a state-space representation, and observed values for both r_t and $r_{m,t}$ for all $t = 1, \ldots, T$, we can now employ the techniques outlined in Section 7 to predict the values of the unobserved coefficients, α_t and β_t, across time.

Finally, the dynamic CAPM can also be interpreted as a *dynamic factor model*. Given this interpretation, the *stochastic slope coefficients* β_i now represent trending, unobserved, factors that follow their own stochastic factor dynamics. See Section 9 on "Factor models and common trends" for more details.

Of course, in the special case where $\boldsymbol{x}_t$ is *observed* (through proxy) the above factor model representation is subject to Roll's critique in that any empirical test of (51b) is really a test of the mean-variance efficiency of the proxy chosen for $\boldsymbol{x}_t$.

2.2.12 Stochastic Volatility

A popular method of modeling persistence in the second-moment of financial asset series is the ARCH model of Engle (1982):

$$y_t = \mu + \sigma_t u_t, \quad \text{where} \quad u_t \sim N(0,1), \tag{52a}$$

$$\text{and} \quad \sigma_t^2 = E[\sigma_t^2 u_t^2 | Y_{t-1}] = \alpha + \beta \sigma_{t-1}^2 u_{t-1}^2. \tag{52b}$$

However, in the ARCH framework, the conditional volatility dynamics are driven in a completely deterministic fashion given past observations (that is, they are path dependent given information set $Y_t = \{y_t, y_{t-1}, \ldots, y_1\}$, and the constraint imposed by (52b)). However, these second-moment dynamics may in fact be better modeled by imposing a specification that implies a strictly larger information set than Y_t. That is, we make the conditional volatility dynamics, σ_t^2, stochastic by introducing the exogenous innovations v_t into (52b):

$$\ln(\sigma_t^2) = \ln(E[\sigma_t^2 u_t^2 | \Phi_{t-1}]) = \alpha + \beta \ln(\sigma_{t-1}^2) + v_t, \quad \text{where} \quad v_t \sim N(0, \sigma_v^2), \tag{53}$$

and where we have taken logs to ensure positivity of the conditional volatility process.

Note that by enlarging the information set from Y_t in (52b) to $\Phi_t = \{Y_t, \sigma_t^2\}$ in (53), we are in an intuitive sense "increasing the types of questions" we can ask of our probability measure. That is, we are being more detailed about how outcomes in our probability space map to random variables in our model. However, note that the random variable σ_t^2 is latent or unobserved, and therefore, the information set we actually make inferences from will be an approximate one. In fact, it is this latent variable that makes this model amenable to state-space signal extraction methods.

Note that we can also impose a probability law on the conditional mean process, if we augment Φ_t again, to $\Phi_t = \{Y_t, \sigma_t^2, \mu_t\}$:

$$\mu_t = E[y_t | \Phi_{t-1}] = \gamma + \delta \mu_{t-1} + z_t, \quad \text{where} \quad z_t \sim N(0, \sigma_z^2). \tag{54}$$

Therefore, the entire stochastic volatility and levels model can be written in state-space form, similar to (1), as:

$$\begin{aligned} \boldsymbol{y}_t \equiv \ln(y_t^2) &= \begin{bmatrix} 1 & 1 \end{bmatrix} \boldsymbol{x}_t + \ln(u_t^2) \\ &\equiv \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t, \end{aligned} \tag{55a}$$

$$\begin{aligned} \text{and} \quad \boldsymbol{x}_t \equiv \begin{bmatrix} \mu_t \\ \ln(\sigma_t^2) \end{bmatrix} &= \begin{bmatrix} \gamma \\ \alpha \end{bmatrix} + \begin{bmatrix} \delta & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} \mu_{t-1} \\ \ln(\sigma_{t-1}^2) \end{bmatrix} + \begin{bmatrix} z_t \\ v_t \end{bmatrix} \\ &\equiv \boldsymbol{c} + \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \end{aligned} \tag{55b}$$

where the observation equation in (52a) has been rewritten as $y_t = e^{\mu_t/2}\sigma_t u_t$ so that it is linear in logs. However, note that in this case we now have that $\boldsymbol{\eta_t} \equiv \ln(u_t^2)$ and so the observation equation innovations are not Gaussian. Therefore while the Kalman filter will represent the best unbiased *linear* predictor, it will not be as efficient as a nonlinear filtering method.

i) Factor GARCH

As a second example, we will consider augmenting the information set of the multivariate Factor GARCH model (Engle,1987). Note that this model has much in common with the material discussed in Section 9 which covers latent *dynamic factor models* and the use of principle components analysis to generate orthogonal factors.

First, consider the factor representation:

$$\boldsymbol{y}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{\delta}_t, \tag{56a}$$

$$\begin{aligned} \text{where} \quad \boldsymbol{f}_t | \boldsymbol{F}_{t-1} &\sim WWN(\boldsymbol{0}, \boldsymbol{\Omega}_t), \\ \boldsymbol{\delta}_t &\sim WWN(\boldsymbol{0}, \boldsymbol{\Lambda_\delta}), \\ \text{and} \quad \boldsymbol{F}_{t-1} &= \{\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, \ldots, \boldsymbol{f}_{t-1}, \boldsymbol{f}_{t-2}, \ldots\}, \end{aligned}$$

where $\boldsymbol{B}$ is $n \times k$, and $\boldsymbol{f}_t$ is $k \times 1$ where $k < n$.

From (56a) we have that the conditional covariance of $\boldsymbol{y}_t$ is $\boldsymbol{\Sigma}_{\boldsymbol{y}_t} = \boldsymbol{B}\boldsymbol{\Omega}_t\boldsymbol{B}' + \boldsymbol{\Lambda_\delta}$. Of course,

assuming that $\boldsymbol{\Omega}_t$ is diagonal we can write $\boldsymbol{\Sigma}_{\boldsymbol{y}_t} = \sum_{i=1}^{k} \boldsymbol{b}_i \omega_{i,t} \boldsymbol{b}_i' + \boldsymbol{\Lambda}_{\boldsymbol{\delta}}$, where $\boldsymbol{b}_i$ is the i'th column of $\boldsymbol{B}$, and $\omega_{i,t}$ is the i'th diagonal element of $\boldsymbol{\Omega}_t$.

In order to capture dynamic persistence in the second moment of the factors, $\boldsymbol{\Omega}_t$, we impose the following parsimonious GARCH(1,1) structure on each of the k diagonal elements:

$$\omega_{i,t} = \alpha_i f_{i,t-1}^2 + \beta_i \omega_{i,t-1}, \quad \forall i = 1, \ldots, k, \tag{57}$$

where $\hat{f_{i,t}}$ can be estimated by the i'th element of $\boldsymbol{L}^{*T} \boldsymbol{y}_t$, and $\boldsymbol{L}^*$, $n \times k$, contains the first k columns of $\boldsymbol{L}$ from the spectral decomposition of the unconditional variance of $\boldsymbol{y}_t$. That is, $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}' = \boldsymbol{\Sigma}_{\boldsymbol{y}}$, so $\boldsymbol{L}'\boldsymbol{\Sigma}_{\boldsymbol{y}}\boldsymbol{L}$ is diagonal and the k elements of $\boldsymbol{L}'\boldsymbol{y}_t$ represent an orthogonal set of random factors that account for all the variance of $\boldsymbol{y}_t$. These are the principle components (see Section 9).

Now, subbing (57) into $\boldsymbol{\Sigma}_{\boldsymbol{y}_t} = \sum_{i=1}^{k} \boldsymbol{b}_i \omega_{i,t} \boldsymbol{b}_i' + \boldsymbol{\Lambda}_{\boldsymbol{\delta}}$ above yields:

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_t} = \sum_{i=1}^{k} \alpha_i \boldsymbol{b}_i f_{i,t-1}^2 \boldsymbol{b}_i' + \sum_{i=1}^{k} \beta_i \boldsymbol{b}_i \omega_{i,t-1} \boldsymbol{b}_i' + \boldsymbol{\Lambda}_{\boldsymbol{\delta}} \tag{58a}$$

$$= \sum_{i=1}^{k} \alpha_i \boldsymbol{b}_i \left(\boldsymbol{l}_i^{*'} \boldsymbol{y}_{t-1} \right)^2 \boldsymbol{b}_i' + \sum_{i=1}^{k} \beta_i \boldsymbol{b}_i \left(\boldsymbol{l}_i^{*'} \boldsymbol{\Sigma}_{\boldsymbol{y}_{t-1}} \boldsymbol{l}_i^* \right) \boldsymbol{b}_i' + \boldsymbol{\Lambda}_{\boldsymbol{\delta}}, \tag{58b}$$

where $\boldsymbol{l}_i^*$ is the i'th column of $\boldsymbol{L}^*$. Note that (58b) represents a first-order difference equation for $\boldsymbol{\Sigma}_{\boldsymbol{y}_t}$ and is deterministic given the $\boldsymbol{y}_t$'s. Therefore, signal extraction methods are unnecessary as nothing is unobserved.

However, since the Factor GARCH model implies that the conditional heteroskedasticity is affecting the factors, $\boldsymbol{f}_t$, and not the innovations, $\boldsymbol{\delta}_t$, this is analogous to imposing a GARCH structure on (54) above, but where $\mu_t = z_t$ and $z_t \sim N(0, \sigma_{z,t}^2)$ is now path dependent. Of course, we could always allow for unobserved autoregressive dynamics on $\boldsymbol{f}_t$, implementing state-space framework prediction of this latent "state" variable and avoiding the need for principal components estimation. Another alternative would be to impose a "Factor Stochastic Volatility" specification, with unobserved stochastic processes driving the diagonal elements $\omega_{i,t}$.

3 Nonlinear dynamic state-space models

3.1 The nonlinear state-space model

Generally, the state-space representation requires two assumptions, namely that the process $\boldsymbol{x}_t$ is Markov so that $f(\boldsymbol{x}_t|X_{t-1}, Y_{t-1}) = f(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ and that the conditional distribution of $\boldsymbol{y}_t$ only depends on the current value of the state, $\boldsymbol{x}_t$, or $g(\boldsymbol{y}_t|Y_{t-1}, X_t) = g(\boldsymbol{y}_t|\boldsymbol{x}_t)$, where $Y_t = \{\boldsymbol{y}_t, \ldots, \boldsymbol{y}_0\}$ and $X_t = \{\boldsymbol{x}_t, \ldots, \boldsymbol{x}_0\}$.

Therefore, the general state-space model considers the joint distribution of the process $[(\boldsymbol{x}_t', \boldsymbol{y}_t')']$, $l(\cdot)$:

$$l(\boldsymbol{y}_t, \boldsymbol{x}_t|X_{t-1}, Y_{t-1}) = f(\boldsymbol{y}_t|Y_{t-1}, X_t)g(\boldsymbol{x}_t|X_{t-1}, Y_{t-1}) \tag{59a}$$

$$= f(\boldsymbol{y}_t|\boldsymbol{x}_t)g(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}). \tag{59b}$$

where the initial conditions of the process are defined by the marginal distribution of $\boldsymbol{x}_0$.

3.1.1 Weak nonlinear state-space model

The weak form of the nonlinear dynamic state-space model is as follows:

$$\boldsymbol{x}_t = a(\boldsymbol{x}_{t-1}, \boldsymbol{\epsilon}_t), \tag{60a}$$

$$\boldsymbol{y}_t = c(\boldsymbol{x}_t, \boldsymbol{\eta}_t), \tag{60b}$$

with the moment restrictions:

$$E[\boldsymbol{\epsilon}_t] = \mathbf{0}, \quad Cov(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-s}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}t}\mathbb{1}_{s=0}, \tag{2a}$$

$$E[\boldsymbol{\eta}_t] = \mathbf{0}, \quad Cov(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t-s}) = \boldsymbol{\Sigma}_{\boldsymbol{\eta}t}\mathbb{1}_{s=0},$$

$$E[\boldsymbol{\epsilon}_{t-j}\boldsymbol{\eta}_{t-s}'] = \mathbf{0}, \quad \forall j, s \in \mathbb{Z},$$

$$\text{and} \quad Cov(\boldsymbol{x}_0, \boldsymbol{\epsilon}_t) = Cov(\boldsymbol{x}_0, \boldsymbol{\eta}_t) = \mathbf{0}, \quad \forall t > 0,$$

where:

- $\boldsymbol{y}_t$ is a $n \times 1$ vector of the observed values at time t.
- $\boldsymbol{x}_t$ is a $p \times 1$ vector of state process values at time t.[6]
- $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are assumed uncorrelated with each other across all time lags, and their covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}t}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}t}$, depend on t.
- $a(\cdot)$ and $c(\cdot)$ are some nonlinear functions. These functions are assumed to be *non-stochastic*.
- Equation (1a) is called the "state transition" equation and (1b) is called the "observation" or "measurement" equation.
- The state initial condition, $\boldsymbol{x}_0$, is assumed stochastic with second order moments denoted $E[\boldsymbol{x}_0] = \boldsymbol{\mu}$ and $Var(\boldsymbol{x}_0) = \boldsymbol{\Sigma}_{\boldsymbol{x_0}}$. Finally, there exists zero correlation between the initial state condition and the observation and state error terms, for all dates $t > 0$.

3.1.2 The Gaussian nonlinear state-space model

In the weak version of the nonlinear dynamic state-space model, the assumptions concern only the first and second-order moments of the noise processes, or equivalently the first and second-order moments of the joint process $[(\boldsymbol{x}_t', \boldsymbol{y}_t')']$. As in the case of the linear state-space model, we can introduce the restriction of, independent, and identically distributed, Gaussian white noises (IIN) for the errors of the state and measurement equations. The Gaussian nonlinear state space model is therefore defined as:

[6]Other terminology include the "signal" in the engineering context, "control variable" in the systems control literature, "latent factor" in the factor models approach, or "stochastic parameter" in the Bayesian context. See Section 4 for more details.

$$\boldsymbol{x}_t = a(\boldsymbol{x}_{t-1}, \boldsymbol{\epsilon}_t), \tag{60a}$$

$$\boldsymbol{y}_t = c(\boldsymbol{x}_t, \boldsymbol{\eta}_t), \tag{60b}$$

$$\begin{pmatrix} \boldsymbol{\epsilon}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \sim IIN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\Sigma}_{\epsilon t} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\eta_t} \end{bmatrix} \right) \tag{62a}$$

$$E[\boldsymbol{\epsilon}_{t-j} \boldsymbol{\eta}'_{t-s}] = \mathbf{0}, \ \ \forall j, s \in \mathbb{Z},$$

$$\text{with} \quad \boldsymbol{x}_0 \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}0}), \tag{62b}$$

$$\text{where} \quad \boldsymbol{x}_0 \quad \text{and} \quad (\boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t) \quad \text{are independent.} \tag{62c}$$

However, when the functions $a(\cdot)$ and $c(\cdot)$ are nonlinear, under the assumption of Gaussian noise, it is no longer the case that the joint process $[(\boldsymbol{x}'_t, \boldsymbol{y}'_t)']$ is Gaussian. This implies that all marginal and conditional distributions concerning the components of these processes are also not necessarily Gaussian.

4 Terminologies

It is also interesting to note that given the widespread use of the state-space framework across different disciplines, a wide variety of interpretations have arisen regarding its implementation. For example, the examples illustrated in the previous section employ a number of different terminologies depending on the context:

- $\boldsymbol{y}_t$ is equivalently referred to as the:
  - measure
  - endogenous variable

- ◇ output variable

- Likewise, $\boldsymbol{x}_t$ is equivalently the:

  - ◇ state variable
  - ◇ signal
  - ◇ control variate
  - ◇ factor
  - ◇ latent variable
  - ◇ input variable

- And while, (1b) is typicalled called the "measurement equation," we have that (1a) is equivalently called the:

  - ◇ state equation
  - ◇ factor dynamics

Consequently, the state-space formulation is often also referred to equivalently as the "dynamic factor model" representation, where (1a) defines the "factor dynamics" and the matrix $\boldsymbol{H}_t$ is called the "factor loadings" matrix. See Section 9 for further discussion.

Moreover, the model also has the alternative Bayesian interpretation where $\boldsymbol{x}_t$ represents instead a *stochastic parameter* and where the (1a) defines the dynamics of the prior distribution of the stochastic parameter. See, for example, Section 5.2.3 for more details.

Finally, with regards to the predictive problems from Section 5, we have equivalent terminologies describing the case where the state $\boldsymbol{x}_t$ is unobserved. That is, $\mathcal{L}\left[\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \dots\right]$, where $\mathcal{L}$ is some optimal linear predictor, is equivalently referred to as:

- linear filtering
- linear signal extraction

- linear prediction of the latent factor

- the best linear predictor of $\boldsymbol{x}_t$ given $Y_t = \{\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \dots\}$

5 The prediction problem

The problem of statistical inference can be broken into two equally important exercises. First, there is the problem of *estimation* of the theoretical model parameters, given observed random samples. Second, there is the problem of *prediction*, where we attempt to predict values of stochastic processes conditional on observed random samples, which we refer to as the *information set*. Therefore, in the first case what we are doing is drawing inferences on *fixed* parameters of the model, where in the latter case we are in fact infering values of *random variables*. This section is exclusively concerned with the latter.

Of course, it is not always the case that we have access to random samples of the stochastic processes we have defined in our theoretical model. In this case we say that the stochastic process is *latent* or unobserved. Of course, this does not stop us from infering the value of these stochastic processes any more than we are able to do so with those which we partially observe. In both cases, these inferences fall under the category of prediction, whether we are predicting the future value of a partially observed stochastic process, or we are predicting the value of one we never observed at all.

Therefore, given the state-space representations in Section 2, there are a number of particular prediction problems that can be formulated, which differ only in the choice of process we wish to predict ($\boldsymbol{y}_t$ or $\boldsymbol{x}_t$) and the information set we have available.

The state-space models with $\boldsymbol{x}_t$ unobservable imply the information set $Y_\tau \equiv \{\boldsymbol{y}_\tau, \boldsymbol{y}_{\tau-1}, \dots, \boldsymbol{y}_0\}$. Therefore, we say that the optimal *forecast* at time T, with horizon h, of either $\boldsymbol{y}_{T+h}$ or $\boldsymbol{x}_{T+h}$ employs the information set Y_T; the optimal *filtering* of $\boldsymbol{x}_t$ employs the set Y_t; and finally the optimal *smoothing* of $\boldsymbol{x}_t$ employs the set Y_T, such that $T > t$. Therefore, we have the problems:

1. Prediction of $\boldsymbol{y}_{T+h}|Y_T$ or $\boldsymbol{x}_{T+h}|Y_T$. These are the *forecasting* problems.

2. Prediction of $\boldsymbol{x}_t|Y_t$ for any $t = 1, \ldots, T$. This is the *filtering* problem.

3. Prediction of $\boldsymbol{x}_t|Y_T$. This is the *smoothing* problem.

Notice that since the forecasting problem involves a choice of horizon, h, it implies an entire term structure of predictions as $h \to \infty$.

Moreover, so far we have only concerned ourselves with the *levels* of the processes. Therefore, we may also be interested in predicting how a process may behave, say in how much it varies or whether or not it tends to exhibit positive or negative outliers. Therefore, we may be interested in predicting:

1. Prediction of $(\boldsymbol{y}_{T+h} - \boldsymbol{y}_{T+h}|Y_T)^k|Y_T$ or $(\boldsymbol{x}_{T+h} - \boldsymbol{x}_{T+h}|Y_T)^k|Y_T$. These are the *forecasting* problems.

2. Prediction of $(\boldsymbol{x}_t - \boldsymbol{x}_t|Y_t)^k|Y_t$ for any $t = 1, \ldots, T$. This is the *filtering* problem.

3. Prediction of $(\boldsymbol{x}_t - \boldsymbol{x}_t|Y_T)^k|Y_T$. This is the *smoothing* problem.

for any $k > 1 \in N$. Therefore, any prediction itself can be evaluated both on its "point" accuracy and its "interval" accuracy – thus we have the notions of both point and interval predictions.

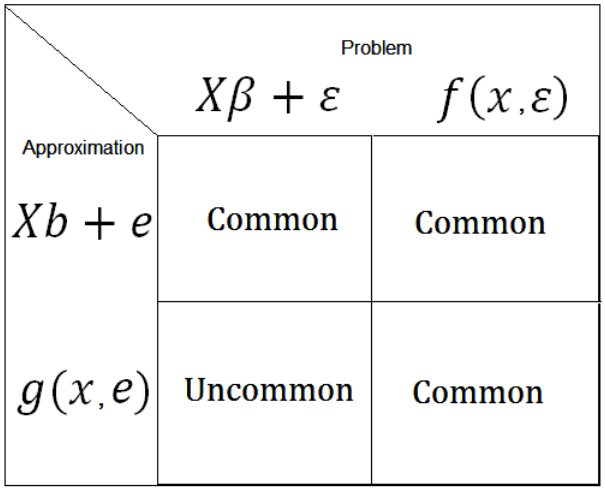
5.1 Types of predictors

Of course, prediction itself can be divided into different types of predictors. Given a particular a problem we may approach the modeling exercise as an approximation, where this approximation may be some linear function, say $x\beta + \epsilon$, or some nonlinear function $f(x, \epsilon)$.

Let us now be more specific about how the prediction problem is formulated, given the general, mean zero, stochastic process $y_t = f(Y_{t-1}, \epsilon_t)$, where again Y_{t-1} represents the information set generated by the stochastic process at time $t - 1$.

It can be shown that in the general case, the optimal predictor $\hat{y}_{t|t-1} = f(Y_{t-1})$ under the minimum mean squared error (MMSE) criterion is the *conditional expectation*, $E[y_t|Y_{t-1}]$. That is, the argmin of the quadratic loss function, the expected mean squared error $E\left[(y_t - f(Y_{t-1}))^2\right]$,

Figure 3: Types of model approximations



can be shown to be $f(Y_{t-1}) = E[y_t|Y_{t-1}]$. See see Priestley (1981, pg. 76) or Hamilton (1994, pg. 72) for proof.

However, suppose instead that we wish to restrict our choice to the class of *linear* predictors. That is, $\hat{y}_{t|t-1} = f(Y_{t-1}) = \boldsymbol{a}'\boldsymbol{y}^*_{t-1}$ where $\boldsymbol{y^*_{t-1}} = \begin{bmatrix} y_{t-1} & y_{t-2} & y_{t-3} & \dots & y_0 \end{bmatrix}'$. Interestingly, the value of $\boldsymbol{a}'$ that satisfies $E\left[\left(y_t - \boldsymbol{a}'\boldsymbol{y}^*_{t-1}\right)\boldsymbol{y}^{*'}_{t-1}\right] = \mathbf{0}'$, that is exhibits zero covariance between forecast error and the independent variable, also minimizes the quadratic loss function above, given the constraint of having to choose amongst linear predictors.

We call this value of $\boldsymbol{a}$ the *linear projection*:

$$\boldsymbol{a}' = E[y_t\boldsymbol{y}^{*'}_{t-1}]E[\boldsymbol{y}^*_{t-1}\boldsymbol{y}^{*'}_{t-1}]^{-1}, \tag{63}$$

and the definition becomes clear when we note that:

$$f(Y_{t-1}) = \boldsymbol{a}'\boldsymbol{y}^*_{t-1} = E[y_t\boldsymbol{y}^{*'}_{t-1}]E[\boldsymbol{y}^*_{t-1}\boldsymbol{y}^{*'}_{t-1}]^{-1}\boldsymbol{y}^*_{t-1} \equiv \mathcal{P}[y_t|Y_{t-1}] \tag{64}$$

and so $\mathcal{P}[y_t|Y_{t-1}]$ represents the scalar projection of y_t onto the Hilbert space spanned by the column $\boldsymbol{y}^*_{t-1}$, with the covariance norm imposed.

Interestingly, if we consider the special case of the linear, Gaussian, stochastic process with uncorrelated innovations, it turns out that the linear projection is equal to the conditional expectation since we know that for Gaussian random variables:

$$\begin{aligned} E[y_t|Y_{t-1}] &= Cov(y_t, Y_{t-1})Cov(Y_{t-1}, Y_{t-1})^{-1}Y_{t-1} \\ &\equiv \mathcal{P}[y_t|Y_{t-1}], \end{aligned} \tag{65a}$$

which is a standard result involving partitioning the set Y_t into y_t and Y_{t-1}, and without loss of generality we assume all the elements of Y_t are mean zero.

However, if the model is not linear in the stochastic variables, or if the stochastic processes are neither weak white nor Gaussian, then there is no reason to believe that the linear projection will prove the best predictor, although it is best amongst the class of linear predictors so that $MSE(\mathcal{P}[y_t|Y_{t-1}]) \geq MSE(E[y_t|Y_{t-1}])$. This is because the linear projection is a function only of the second moment properties of the stochastic processes.

Therefore, to summarize, we have three types of predictors (where one is a special case of the other):

- The linear projection, $\mathcal{P}[y_t|Y_\tau] \quad \tau \in \{1, \ldots, T\}$, which is some linear function of the information set.
- The expectation, $E[y_t|Y_\tau]$, which may possibly be a nonlinear function of the information set.
- The expectation under the assumption of a linear model and Gaussianity co-incides with the linear projection.

5.2 Examples

5.2.1 Prediction of VAR(1) process (weak form)

As a simple example, consider the case where $\boldsymbol{x}_t$, the state process in (1), is observed and that the system matrices, including $\boldsymbol{F}$, are time-invariant. That is, we assume that $\boldsymbol{H} = \boldsymbol{I}$ and we know the covariance matrices $\boldsymbol{\Sigma_\epsilon} >> \mathbf{0}$ and $\boldsymbol{\Sigma_\eta} = \mathbf{0}$. Therefore, the model is a VAR(1) and is a special case of the state-space model in (1) – see equation (4a).

If the eigenvalues of $\boldsymbol{F}$ have modulus strictly less than 1, the VAR(1) can be inverted and written as an VMA(∞):

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t && \text{(66a)} \\ &= \left[\boldsymbol{I} - \boldsymbol{F}L\right]^{-1} \boldsymbol{\epsilon}_t \\ &= \boldsymbol{\epsilon}_t + \boldsymbol{F}\boldsymbol{\epsilon}_{t-1} + \boldsymbol{F}^2\boldsymbol{\epsilon}_{t-2} + \dots && \text{(66b)} \end{aligned}$$

Therefore, we can consider the stochastic process as being a linear function of weak white noise, $\boldsymbol{\epsilon}_t$. That is, the information set is equivalently written as $Y_t \equiv \{\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_{t-1}, \dots\}$.

Now, we would like to try and predict $\boldsymbol{x}_t$ conditional on information set Y_{t-1} under the MMSE (minimum mean squared error) criterion. Let $f(Y_{t-1})$ represent our predictor function. We would like to minimize a quadratic loss function, the expected mean squared error $E\left[\left(\boldsymbol{x}_t - f(Y_{t-1})\right)\left(\boldsymbol{x}_t - f(Y_{t-1})\right)'\right]$, where $f(Y_{t-1})$ is now a linear function of the information set at time $t-1$. The "best" (in the sense of using the information embodied in the information set efficiently, according to the MMSE criterion), linear, predictor of the VAR(1) process $\boldsymbol{x}_t$, given information set Y_t, is the *linear projection*, defined as $\mathcal{P}[\boldsymbol{x}_t|Y_{t-1}] \equiv \boldsymbol{\hat{x}_{t|t-1}}$.

However, it is clear that the linear projection must be:

$$\boldsymbol{\hat{x}_{t|t-1}} = \boldsymbol{F}\boldsymbol{\epsilon}_{t-1} + \boldsymbol{F}^2\boldsymbol{\epsilon}_{t-2} + \dots \tag{67}$$

since we have that:

$$E[(\boldsymbol{y}_t - \hat{\boldsymbol{x}}_{\boldsymbol{t|t-1}})\boldsymbol{x}_{t-1}^{*'}] = \mathbf{0} \tag{68}$$

and from the discussion of the previous section we know that this value of $\hat{\boldsymbol{x}}_{t-1} = f(Y_{t-1})$ must be the argmin of the expected mean squared forecast error.

Note that we can also explicity solve for the value of $\boldsymbol{a}$ just as was done before, and this provides a moment condition that can be used to also *estimate* the system matrix $\boldsymbol{F}$ given sample data by employing the sample moment counterpart, $\hat{\boldsymbol{a}}$.

Rewrite the model in (66b) as:

$$\boldsymbol{x}_t = \boldsymbol{\epsilon}_t + \begin{bmatrix} \boldsymbol{F} & \boldsymbol{F}^2 & \dots \boldsymbol{F}^Q \end{bmatrix} \begin{bmatrix} \boldsymbol{\epsilon}_{t-1} \\ \boldsymbol{\epsilon}_{t-2} \\ \vdots \\ \boldsymbol{\epsilon}_{t-Q} \end{bmatrix} \tag{69a}$$

$$= \boldsymbol{\epsilon}_t + \boldsymbol{A}'\boldsymbol{\epsilon}_{t-1}^*, \tag{69b}$$

where we have truncated the lags at Q.

We can therefore solve for $\boldsymbol{A}'$ as:

$$\begin{aligned} \boldsymbol{x}_t &= \boldsymbol{\epsilon}_t + \boldsymbol{A}'\boldsymbol{\epsilon}_{t-1}^* \\ \Leftrightarrow \boldsymbol{x}_t\boldsymbol{\epsilon}_{t-1}^{*'} &= \boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t-1}^{*'} + \boldsymbol{A}'\boldsymbol{\epsilon}_{t-1}^*\boldsymbol{\epsilon}_{t-1}^{*'} \\ \Leftrightarrow E[\boldsymbol{x}_t\boldsymbol{\epsilon}_{t-1}^{*'}] &= E[\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_{t-1}^{*'}] + \boldsymbol{A}'E[\boldsymbol{\epsilon}_{t-1}^*\boldsymbol{\epsilon}_{t-1}^{*'}] \\ \Leftrightarrow \boldsymbol{A}' &= E[\boldsymbol{x}_t\boldsymbol{\epsilon}_{t-1}^{*'}]E[\boldsymbol{\epsilon}_{t-1}^*\boldsymbol{\epsilon}_{t-1}^{*'}]^{-1}. \end{aligned} \tag{70a}$$

Therefore, since the stochastic process is stationary and ergotic by assumption, we can esti-

mate the population moment condition above via the sample averages:

$$\frac{1}{T}\sum_{\tau=1}^{'} \boldsymbol{x}_\tau \boldsymbol{\epsilon}_{\tau-1}^{*'} \rightarrow^P E[\boldsymbol{x}_t \boldsymbol{\epsilon}_{t-1}^{*'}], \tag{71a}$$

$$\text{and} \quad \frac{1}{T}\sum_{\tau=1}^{'} \boldsymbol{\epsilon}_{\tau-1}^{*} \boldsymbol{\epsilon}_{\tau-1}^{*'} \rightarrow^P E[\boldsymbol{\epsilon}_{t-1}^{*} \boldsymbol{\epsilon}_{t-1}^{*'}]. \tag{71b}$$

Finally, consider the case where the state process $\boldsymbol{x}_t$ in (1) is in fact *unobserved*. [7] Now we must also predict $\boldsymbol{x}_t$ conditional on the information set Y_τ (given different possible values for τ). This is where the concepts of smoothing and filtering become important – we leave the discussion of prediction given unobserved state processes to Section 7 and the estimation of the system matrices under these circumstances to Section 8 on estimation.

5.2.2 Prediction of static state-space model (strong form)

As another example, consider the case where $\boldsymbol{x}_t$, the state process in the strong form state-space model (3), is *unobserved* and that the system matrices, $\boldsymbol{F}$ and $\boldsymbol{H}$, and covariance matrices, $\boldsymbol{\Sigma_\epsilon}$ and $\boldsymbol{\Sigma_\eta}$ are all time invariant. Moreover, in this case, the state process, $\boldsymbol{x}_t$, is i.i.d. Gaussian and does not exhibit serial correlation. We call this model the *static* state-space model and it is a special case of the strong form model in (3) – see formula (28) above.

From the model in (28) we know that:

$$\begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{y_t} \end{bmatrix} = MVN \left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \quad , \quad \begin{pmatrix} \boldsymbol{\Sigma_\epsilon} & \boldsymbol{\Sigma_\epsilon H'} \\ \boldsymbol{H\Sigma_\epsilon} & \boldsymbol{H\Sigma_\epsilon H' + \Sigma_\eta} \end{pmatrix} \right]. \tag{72}$$

Moreover, from the discussion above, we know that given the linear Gaussian setting, the linear projection is equivalent to conditional expectation. Therefore, the best linear predictors of

[7] For example if $\boldsymbol{\Sigma_\eta} >> \mathbf{0}$.

both observed and unobserved processes are given as:

$$E[\boldsymbol{y}_t|Y_{t-1}] = \boldsymbol{H}E[\boldsymbol{x}_t|Y_{t-1}] + \mathbf{0} = \mathbf{0}, \tag{73a}$$

$$\text{and} \quad E[\boldsymbol{x}_t|Y_{t-1}] = \mathbf{0}. \tag{73b}$$

However consider that the expectation of the state variable, conditional on current time t information, is not trivial:

$$E[\boldsymbol{x}_t|Y_t] = E[\boldsymbol{x}_t\boldsymbol{y}_t']E[\boldsymbol{y}_t\boldsymbol{y}_t']^{-1}\boldsymbol{y}_t \tag{74a}$$

$$= E[\boldsymbol{x}_t\left(\boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t\right)']\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right]^{-1}\boldsymbol{y}_t \tag{74b}$$

$$= \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{H}'\left[\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right]^{-1}\boldsymbol{y}_t. \tag{74c}$$

Note the difference between this example and that given in Section 5.2.1. In the previous case the state process $\boldsymbol{x}_t$ was *observable*, and so our prediction formula was really a prediction of the observation process. However, in this case it is not.

Therefore we are confronted with two problems: a) predict future values of the observed process $\boldsymbol{y}_t$ given information set Y_{t-1} and also b) predict the state process $\boldsymbol{x}_t$ given the information set Y_τ where τ can range across different values. If $\tau = t$, we call this the filtering problem and if $\tau > t$ we call this the smoothing problem. Under the static model framework, the solution is simple. It is simply the linear projection $\mathcal{P}[\boldsymbol{x}_t|Y_\tau]$ above. However, when the state process is not static, the problem becomes much more involved and involves updating the linear projection given new information. This is the essence of the Kalman filter and fixed point smoother algorithms discussed in Section 7. In fact, the Kalman filter solutions below collapse to the expressions above when the state process is static. For another example of a static state-space model, see the static Factor model of Section 9.

5.2.3 Bayesian interpretation of the state-space model

Within the Bayesian context we can always interpret the state transition equation (1a) as defining the dynamics of a prior distribution on the stochastic parameter $\boldsymbol{x}_t$. For example, from the linear state-space model in Section 2.1.2, if we view the "signal" $\boldsymbol{x}_t$ to be a *stochastic parameter*, then the model is more akin to a linear regression of $\boldsymbol{H}_t$ regressed on $\boldsymbol{y}_t$ but where the slope coefficient $\boldsymbol{x}_t$ is allowed to follow an unobserved stochastic process.

Given this interpretation we can work with the probability density functions implied by the respective model, to derive the *posterior* density $p(\boldsymbol{x}_t|Y_t)$, where $Y_t = \{\boldsymbol{y_t}, \boldsymbol{y_{t-1}}, \ldots, \boldsymbol{y_1}\}$ (the posterior density is therefore the filtered density of the latent state, $\boldsymbol{x}_t$). This will be done by appealing to Bayes theorem, and in the process we will make use of the prior density $p(\boldsymbol{x}_t|Y_{t-1})$ and the likelihood $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$, given observed values $\boldsymbol{y}_t$ and stochastic parameter $\boldsymbol{x}_t$.

First, using Bayes theorem it can be shown that the posterior distribution $p(\boldsymbol{x}_t|Y_t)$ is derived as:

$$
\begin{aligned}
p(\boldsymbol{x}_t|Y_t) &= \frac{p(\boldsymbol{x}_t, Y_t)}{p(Y_t)} = \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_t, Y_{t-1})}{p(\boldsymbol{y}_t, Y_{t-1})} \\
\text{where } \; p(\boldsymbol{x}_t, \boldsymbol{y}_t, Y_{t-1}) &= p(\boldsymbol{y}_t|\boldsymbol{x}_t, Y_{t-1}) p(\boldsymbol{x}_t|Y_{t-1}) p(Y_{t-1}) \\
&= p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|Y_{t-1}) p(Y_{t-1}) \\
\\
\Leftrightarrow p(\boldsymbol{x}_t|Y_t) &= \frac{p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|Y_{t-1})}{p(\boldsymbol{y}_t|Y_{t-1})} \\
&\propto p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|Y_{t-1}) \qquad (75a)
\end{aligned}
$$

Therefore, given the assumption of a linear model, from Section 2.1.2, with Gaussian inno-

vations, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$, we have that the prior density is defined as:

$$p(\boldsymbol{x}_t|Y_{t-1}) \sim MVN(\hat{\boldsymbol{x}}_{t|t-1}, \boldsymbol{P}_{t|t-1}) \tag{76a}$$

$$\text{where} \quad \hat{\boldsymbol{x}}_{t|t-1} = \boldsymbol{F}\hat{\boldsymbol{x}}_{t-1|t-1} \tag{76b}$$

$$\text{and} \quad \boldsymbol{P}_{t|t-1} = \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \tag{76c}$$

which follows directly from the linear projection of $\boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t$ into the space spanned by Y_{t-1} [see (89a) and (100a) below]. Therefore, $\hat{\boldsymbol{x}}_{t|t-1}$ is the forecast of the state variable $\boldsymbol{x}_t$, given information at time $t-1$, which is a linear function of $\hat{\boldsymbol{x}}_{t-1|t-1}$, the filtered state variable.

Moreover, the likelihood given observed values $\boldsymbol{y}_t$ is given as:

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) \sim MVN(\boldsymbol{H}\boldsymbol{x}_t, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}) \tag{77}$$

which follows immediately from the definition of the state-space model in (1). Therefore, given the linear model, we have that the forecast of $\boldsymbol{y}_t$ conditional on time $t-1$ information is therefore $\boldsymbol{H}\hat{\boldsymbol{x}}_{t|t-1}$, which again follows from the linear projection.

Furthermore, by standard results on the moments of the product of two normal densities, we have that the posterior distribution is proportional to:

$$p(\boldsymbol{x}_t|Y_t) \sim MVN(\hat{\boldsymbol{x}}_{t|t}, \boldsymbol{P}_{t|t}) \tag{78a}$$

$$\text{where} \quad \hat{\boldsymbol{x}}_{t|t} = \boldsymbol{P}_{t|t}\left(\boldsymbol{P}_{t|t-1}^{-1}\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{H}'{\boldsymbol{\Sigma}_{\boldsymbol{\eta}}}^{-1}\boldsymbol{y}_t\right) \tag{78b}$$

$$\text{and} \quad \boldsymbol{P}_{t|t} = \left[\boldsymbol{H}'{\boldsymbol{\Sigma}_{\boldsymbol{\eta}}}^{-1}\boldsymbol{H} + \boldsymbol{P}_{t|t-1}^{-1}\right]^{-1} \tag{78c}$$

where it can be shown that (78b) and (78c) are equal to (98b) and (100b), respectively. Therefore, the above expressions provide a recursive system of equations that together generate the filtered prediction of the state variable $\hat{\boldsymbol{x}}_{t|t}$ [to see this sub (76b) into (78b) and (76c) into (78c)]. This set of recursions is effectively the Kalman filter (1960) and so we can see that from the Bayesian framework the Kalman filter arises under the special case of the linear state-space model with

Gaussian innovations.

However, if the model is nonlinear we cannot expect the conditional expectation to be of a linear form. Generally, we must approximate the set of filter recursions by numerical methods such as Monte-Carlo. Essentially, we are able to generate the expectation of the filtered state density in (75a) by drawing from it and appealing to a law of large numbers.

Harrison and Stevens (1976) presents the general discussion of the Bayesian approach to state-space modeling.

6 Prediction in the frequency domain

Consider the engineering context, where the classical approach to linear time invariant system analysis involved spectral representations and "frequency response" to unit shocks passed through appropriate physical filters, the state, x_t, has kept its interpretation as the "signal" corrupted by noise. Since the signal is latent, we often wish to "extract" the signal from the noise and this defines the "signal extraction" problem:

$$\text{argmin}_{\{a_j\}_{j=0}^{\infty}} E[(x_t - \hat{x}_t)^2] \tag{79a}$$

$$\text{where} \quad \hat{x}_t = \sum_{j=0}^{\infty} a_j y_{t-j}, \tag{79b}$$

$$\text{and} \quad y_t = x_t + \eta_t, \tag{79c}$$

and where η_t is a weak white noise, and we only observe the noisey signal y_t. Moreover, x_t is uncorrelated with the noise η_t, and both processes are stationary. That is, we wish to "filter" out the noise, η_t, to recover the signal, x_t, when we only observe, y_t. Under the MMSE criterion, the optimal solution will minimize (79a), by choosing the sequence of filter coefficients $\{a_j\}_{j=0}^{\infty}$.

That is, given observations of the output $Y_t = \{y_t, y_{t-1}, y_{t-2}, \dots\}$, we wish to find the most efficient predictor of the latent variable x_t in the sense of the discussion provided above in Section 5. The most famous solution, in the frequency domain, is the Wiener filter developed during the

mid-part of the last century. This solution is interpreted in terms of an optimal *transfer function*, $A(e^{-i\omega})$, which is the Fourier transform of the impulse response function, $\{a_j\}_{j=0}^{\infty}$:

$$A(e^{-i\omega}) = \frac{h_{xy}(\omega)}{h_{yy}(\omega)} = \frac{h_{xx}(\omega)}{h_{xx}(\omega) + h_{\eta\eta}(\omega)} \tag{80a}$$

$$\text{where} \quad A(z) = \sum_{j=-\infty}^{\infty} a_j z^j, \tag{80b}$$

and where $h_{xy}(\omega)$ is the cross-spectral density between x_t and y_t, and $h_{xx}(\omega)$ is the spectral density of x_t. Therefore, we can see that the Wiener filter solution allows frequencies to pass through when the signal-to-noise ratio is high, and vice versa when low:

$$h_{\hat{x}\hat{x}} = |A(e^{-i\omega})|^2 h_{yy}(\omega). \tag{81}$$

See Priestley (1981), Chapter 10, for more details.

However, given the advent of digital computers, the more useful time domain counterpart is found in the work of Kalman (1960) and is discussed in more detail within Section 7.

Subsequently, the application of signal extraction methods to modern economics, and empirical time-series analysis more generally, was developed by Nerlove et al. (1979) and popularized by Harvey (1984,89) and the "Unobserved Components" framework which is discussed in Section 9.4 (see also example 2.2.10).

7 Prediction in the time domain

Let us turn now to the general solutions of the following prediction problems in the time domain, when the state-space representation is constrained to be linear:

1. Forecasting future values of $\boldsymbol{y}_{T+h}$ for some horizon h given observed values from $t = 1, \ldots, T$, or forecasting $\boldsymbol{x}_{T+h}$ given predictions of the state $\hat{\boldsymbol{x}}_{t|t}, t = 1, \ldots, T$.

2. Filtering $\boldsymbol{x}_\tau$ given observed values, $\boldsymbol{y}_t$, $t = 1, \ldots, \tau \leq T$.

3. Smoothing $\boldsymbol{x}_\tau$ given observed values, $\boldsymbol{y}_t$, $t = 1, \ldots, T$, where $\tau < T$.

The model in (1) implies the information set $Y_t \equiv \{\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1\}$. Therefore, the optimal forecast of either $\boldsymbol{y}_t$ or $\boldsymbol{x}_t$ employs the information set Y_{t-1}; the optimal filtering of $\boldsymbol{x}_t$ employs the set Y_t; and finally the optimal smoothing of $\boldsymbol{x}_t$ employs the set Y_T.

It is of historical interest that the first of the three problems was solved almost simultaneously by both Wiener and Kolmogorov around the middle of the 20th century. However, while Kolmogorov worked in the time domain, Wiener solved the equivalent problem in the frequency domain [see Chapter 10 in Priestley (1981)]. Interestingly, the filtering problem is a special case of the smoother which was solved by Wiener in terms of an optimal transfer function (and frequency response) but where we have access to a doubly infinite set of observations on $\boldsymbol{y}_t$ (i.e. our information set is $Y_\infty \equiv \{\ldots, \boldsymbol{y}_{t+1}, \boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots\}$) [see Section 6 above]. It was Kalman's celebrated work in (1960) that rephrased the three problems in terms of the state-space formulation in (1), allowing one to process the filtered state values through a recursive algorithm ideally suited to implementation on digital computers (as opposed to in an analog fashion using physically constructed electrical filters).

It can be shown that as $T \to \infty$ the steady-state solution to the Kalman state-space smoother is equivalent to the Wiener solution. However, the Kalman filter is superior in a number of ways: first, we need not work in the frequency domain; second, the Kalman optimal filter weights are easily updated at each time t, while the Wiener weights must be completely recomputed for each new time period; and finally, the Kalman filter allows us to easily extend the solution to the multivariate case, or to the linear time-variant system case.

The notions of forecasting, filtering, and smoothing are not as different as they first seem. All three involve determining the optimal prediction of some random variable, say Z_t, given a particular information set $\mathcal{F}_\tau$. Therefore, despite the misleading nomenclature, forecasting, filtering and smoothing are really just methods to predict stochastic processes (whether observable or not). In fact, it is only the information set available to us that determines which of the three problems we face. Therefore, all three are examples of the problem of statistical inference of

random variables.

7.1 Linear projections

Recall from Section 5 that if we adopt a quadratic loss function (e.g. minimum mean squared error (MMSE)) criterion, then in general, the argmin function $f(Y_t)$ is the conditional expectation, given the relevant information set [see Section 5.1 and Priestley (1981, pg. 76) for a formal proof]. However, within the linear context, where we adopt a Hilbert space framework, with inner product defined as the covariance between random variables, the best linear unbiased estimator is the *linear projection* (also called the "orthogonal projector").

Formally, for the first problem, given some $\mathcal{H}_T \subset \mathcal{H}_{T+h}$ where $\mathcal{H}$ is a Hilbert space, let $\boldsymbol{y}_{T+h} \in \mathcal{H}_{T+h}$. It can be shown that there exists a unique $\hat{\boldsymbol{y}}_{T+h} \in \mathcal{H}_T$ where $\|\boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h}\|^2$ is minimized, and where $\hat{\boldsymbol{y}}_{T+h}$ is uniquely determined by the property $(\boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h}) \perp \forall \ \boldsymbol{y} \in \mathcal{H}_T$. Therefore, $\hat{\boldsymbol{y}}_{T+h}$ represents the *orthogonal projector*.

For the second problem, given some $\mathcal{L}_t \subseteq \mathcal{L}_T$ where $\mathcal{L}$ is another Hilbert space (orthogonal to $\mathcal{H}$), let $\boldsymbol{x}_t \in \mathcal{L}_t$. It can be shown that there exists a unique $\hat{\boldsymbol{x}}_t \in \mathcal{H}_t$ where $\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|^2$ is minimized, and where $\hat{\boldsymbol{x}}_t$ is uniquely determined by the property $(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t) \perp \forall \ \boldsymbol{y} \in \mathcal{H}_t$.

Finally, the third problem suggests that given some $\mathcal{L}_t \subseteq \mathcal{L}_T$, let $\boldsymbol{x}_t \in \mathcal{L}_t$. It can be shown that there exists a unique $\hat{\boldsymbol{x}}_t \in \mathcal{H}_T$ where $\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|^2$ is minimized, and where $\hat{\boldsymbol{x}}_t$ is uniquely determined by the property $(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t) \perp \forall \ \boldsymbol{y} \in \mathcal{H}_T$.

7.2 The lag representation

The time-invariant version of the linear dynamic state-space model in (1) is obtained when the system matrices, $\boldsymbol{F}$ and $\boldsymbol{H}$, and the noise covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, do not depend on time. Therefore, we get:

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t, \tag{82a}$$

$$\text{and} \quad \boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{82b}$$

$$\text{where} \quad \begin{pmatrix} \boldsymbol{\epsilon}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \sim WWN \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \end{bmatrix} \right). \tag{83a}$$

This time-invariant system can be written in terms of the lag operator (L). The operator is defined on the set of second-order processes.[8] It transforms a given process ($\boldsymbol{x}_t$) into the process ($L\boldsymbol{x}_t$), where $(L\boldsymbol{x}_t) = (\boldsymbol{x}_{t-1})$; thus its components are deduced by drifting time by one lag. The operator is especially important in characterizing the second-order stationary process. Indeed, the process ($\boldsymbol{x}_t$) is second-order stationary if and only if the first and second-order moments of ($\boldsymbol{x}_t$) and ($L\boldsymbol{x}_t$) are the same for all t.

The system in (82) can be written as:

$$\boldsymbol{x}_t = \boldsymbol{F}L\boldsymbol{x}_t + \boldsymbol{\epsilon}_t \quad \text{and} \quad \boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t, \tag{84a}$$

$$\Leftrightarrow (\boldsymbol{I} - \boldsymbol{F}L)\,\boldsymbol{x}_t = \boldsymbol{\epsilon}_t \quad \text{and} \quad \boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t. \tag{84b}$$

If the operator $\boldsymbol{I} - \boldsymbol{F}L$ is invertible (i.e. the eigenvalues of $\boldsymbol{F}$ are strictly less than 1 in absolute value), then we can write (84b) as:

$$\boldsymbol{x}_t = (\boldsymbol{I} - \boldsymbol{F}L)^{-1} \boldsymbol{\epsilon}_t. \tag{85}$$

Therefore, from both (84b) and (82b) the model can be written purely in terms of the observable variable $\boldsymbol{y}_t$ and the error terms as:

$$\boldsymbol{y}_t = \boldsymbol{H}\,(\boldsymbol{I} - \boldsymbol{F}L)^{-1} \boldsymbol{\epsilon}_t + \boldsymbol{\eta}_t. \tag{86}$$

That is, if the model is stationary, we can always rewrite the state-space representation in terms of an *orthogonal basis*.

[8] That is, the lag operator, L, represents a mapping $L : \mathbb{R}^{Pt} \to \mathbb{R}^{P(t-1)}$ or $L^{-1} : \mathbb{R}^{Pt} \to \mathbb{R}^{P(t+1)}$ where $t \in \mathbb{Z}$. Moreover, the composition of lag operators implies that $LL^{-1} = 1$, since $L \circ L^{-1} \equiv LL^{-1} : \mathbb{R}^{Pt} \to \mathbb{R}^{P(t+1)} \to \mathbb{R}^{Pt}$. More generally of course, $L^k : \mathbb{R}^{Pt} \to \mathbb{R}^{P(t-k)}$.

7.3 Forecasting

Let us consider the model in (1), but with time invariant system and noise covariance matrices. That is, we write the system and observation matrices, $\boldsymbol{F}$ and $\boldsymbol{H}$, without their time subscripts and the covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$, also do not depend on time – see (82). From (1b) we have that the optimal linear forecast of future values of $\boldsymbol{y}_{T+h}$ for some horizon h given observed values from $t = 1, \ldots, T$ is derived as:

$$\begin{aligned}
\hat{\boldsymbol{y}}_{T+h} = \mathcal{P}[\boldsymbol{y}_{T+h}|Y_T] &= \boldsymbol{H}\mathcal{P}[\boldsymbol{x}_{T+h}|Y_T] + \mathbf{0} && \text{(87a)} \\
&= \boldsymbol{H}\left(\boldsymbol{F}\mathcal{P}[\boldsymbol{x}_{T+(h-1)}|Y_T] + \mathbf{0}\right) \\
&\vdots \\
&= \boldsymbol{H}\boldsymbol{F}^h\mathcal{P}[\boldsymbol{x}_T|Y_T] && \text{(87b)}
\end{aligned}$$

where $\mathcal{P}[\cdot]$ denotes linear projection, and where $\mathcal{P}[\boldsymbol{y}_{T+h}|Y_T]$ represents the optimal *orthogonal projector* since we have that for any $\boldsymbol{y}_T = \boldsymbol{H}\boldsymbol{x}_T + \boldsymbol{\eta}_T$, $(\boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h}) \perp \boldsymbol{y}_T$, according to the covariance norm.

Moreover, suppose we have already solved for the filtered values $\mathcal{P}[\boldsymbol{x}_T|\boldsymbol{y}_T, \boldsymbol{y}_{T-1}] \equiv \hat{\boldsymbol{x}}_{T|T}$, we have that the optimal forecast of $\boldsymbol{x}_{T+h}$ given estimates of the state $\boldsymbol{x}_t, t = 1, \ldots, T$, is derived as:

$$\begin{aligned}
\hat{\boldsymbol{x}}_{T+h} = \mathcal{P}[\boldsymbol{x}_{T+h}|Y_T] &= \boldsymbol{F}\mathcal{P}[\boldsymbol{x}_{T+(h-1)}|Y_T] + \mathbf{0} && \text{(88a)} \\
&= \boldsymbol{F}\left(\boldsymbol{F}\mathcal{P}[\boldsymbol{x}_{T+(h-2)}|Y_T] + \mathbf{0}\right) \\
&\vdots \\
&= \boldsymbol{F}^h\mathcal{P}[\boldsymbol{x}_T|Y_T] \\
&= \boldsymbol{F}^h\mathcal{P}[\boldsymbol{x}_T|\boldsymbol{y}_T, \boldsymbol{y}_{T-1}] \equiv \boldsymbol{F}^h\hat{\boldsymbol{x}}_{T|T}. && \text{(88b)}
\end{aligned}$$

7.4 Filtering

The second prediction problem above is solved by establishing an expression for $\mathcal{P}[\boldsymbol{x}_t|Y_t]$, for any $t \in \{1, \ldots, T\}$. That is, we wish to filter $\boldsymbol{x}_\tau$ given observed values, $\boldsymbol{y}_t$, $t = 1, \ldots, \tau \leq T$. For simplicity of notation, let $\mathcal{P}[\boldsymbol{x}_t|Y_t] \equiv \hat{\boldsymbol{x}}_{t|t}$. The Kalman filter provides a direct way to compute $\hat{\boldsymbol{x}}_{t|t}$ recursively given any starting values assumed for $\hat{\boldsymbol{x}}_{1|0}$ and $\boldsymbol{P}_{1|0}$ and given values for the system and noise covariance matrices, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.

First note that from (1a):

$$\hat{\boldsymbol{x}}_{t+1|t} = \mathcal{P}[\boldsymbol{F}\boldsymbol{x}_t + \boldsymbol{\epsilon}_{t+1}|Y_t] = \boldsymbol{F}\hat{\boldsymbol{x}}_{t|t}, \tag{89a}$$

$$\text{and} \quad \hat{\boldsymbol{y}}_{t+1|t} = \mathcal{P}[\boldsymbol{H}\boldsymbol{x}_{t+1}|Y_t] = \boldsymbol{H}\hat{\boldsymbol{x}}_{t+1|t}. \tag{89b}$$

Let:

$$\boldsymbol{f}_{t+1} = \boldsymbol{x}_{t+1} - \hat{\boldsymbol{x}}_{t+1|t}, \tag{90a}$$

$$\text{and} \quad \boldsymbol{e}_{t+1} = \boldsymbol{y}_{t+1} - \hat{\boldsymbol{y}}_{t+1|t}, \tag{90b}$$

be the one-step ahead state and observation forecast errors, respectively. We can now define $Var[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] = E[\boldsymbol{f}_t\boldsymbol{f}_t'] \equiv \boldsymbol{P}_{t|t-1}$.

Our goal then is to update the linear projection $\mathcal{P}[\boldsymbol{x}_t|Y_{t-1}]$ with new information, as it arises at time t. It is shown in Section 7.4.1 below, that an expression for updating a linear projection $\mathcal{P}[\boldsymbol{x}_t|Y_{t-1}]$ with time t information is given as:

$$\mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}] = \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] + E[\boldsymbol{f}_t\boldsymbol{e}_t']E[\boldsymbol{e}_t\boldsymbol{e}_t']^{-1}\boldsymbol{e}_t \tag{98a}$$

$$\equiv \hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{e}_t \tag{98b}$$

where $\boldsymbol{K}_t$ is given from (99b), below, as:

$$\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{H}'\left[\boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right]^{-1} \tag{99b}$$

and where from (99c), below, we have that:

$$Var[\boldsymbol{y}_t|\boldsymbol{y}_{t-1}] = E[\boldsymbol{e}_t\boldsymbol{e}_t'] = \boldsymbol{\Sigma}_{\boldsymbol{e}t} = \boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}. \tag{99c}$$

Finally from plugging (100b) into (100a) we get an expression for the conditional variance of $\boldsymbol{x}_t$ in terms of a matrix difference equation. Equations (100b) and (100a), are introduced later but repeated here as

$$\begin{aligned} Var[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] = E[\boldsymbol{f}_t\boldsymbol{f}_t'] &= \boldsymbol{P}_{t|t-1} \\ &= \boldsymbol{F}Var[\boldsymbol{x}_{t-1}|\boldsymbol{y}_{t-1}]\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \qquad (100a) \\ \text{and} \quad \boldsymbol{P}_{t|t} &= \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{\Sigma}_{\boldsymbol{e}t}\boldsymbol{K}_t'. \qquad (100b) \end{aligned}$$

Together they imply that:

$$\begin{aligned} \boldsymbol{P}_{t|t-1} &= \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \\ &= \boldsymbol{F}\left[\boldsymbol{P}_{t-1|t-2} - \boldsymbol{K}_{t-1}\boldsymbol{\Sigma}_{\boldsymbol{e}t-1}\boldsymbol{K}_{t-1}'\right]\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \\ &= \boldsymbol{F}\left[\boldsymbol{P}_{t-1|t-2} - \boldsymbol{K}_{t-1}\left(\boldsymbol{H}\boldsymbol{P}_{t-1|t-2}\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right)\boldsymbol{K}_{t-1}'\right]\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \\ &= \boldsymbol{F}\left[\boldsymbol{P}_{t-1|t-2} - \boldsymbol{P}_{t-1|t-2}\boldsymbol{H}'\boldsymbol{K}_{t-1}'\right]\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \\ &= \boldsymbol{F}\boldsymbol{P}_{t-1|t-2}\left(\boldsymbol{F}\left(\boldsymbol{I} - \boldsymbol{K}_{t-1}\boldsymbol{H}\right)\right)' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \\ &= \boldsymbol{F}\boldsymbol{P}_{t-1|t-2}\boldsymbol{L}_{t-1}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} \qquad (93a) \end{aligned}$$

which is known as the discrete time algebraic Ricatti equation (ARE). Given sufficient conditions the ARE can be solved for a steady-state value which is often useful in embedded systems where computational memory is limited [see Simon (2006, pg.194-199)]. Note that a steady-state value for $\boldsymbol{P}_\infty$ implies a steady-state value for $\boldsymbol{K}_\infty$.

Therefore, to summarize we have the following expressions that together suggest a recursive algorithm for computing the linear filtered solution $\mathcal{P}[\boldsymbol{x}_t|Y_t] \equiv \hat{\boldsymbol{x}}_{t|t}$ for each $t = 1, \ldots, T$:

$$\hat{\boldsymbol{x}}_{t+1|t} = \boldsymbol{F}\hat{\boldsymbol{x}}_{t|t}, \tag{89a}$$

$$\hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{e}_t, \tag{98a}$$

$$\hat{\boldsymbol{y}}_{t+1|t} = \boldsymbol{H}\hat{\boldsymbol{x}}_{t+1|t}, \tag{89b}$$

$$\boldsymbol{f}_{t+1} = \boldsymbol{x}_{t+1} - \hat{\boldsymbol{x}}_{t+1|t}, \tag{90a}$$

$$\boldsymbol{e}_{t+1} = \boldsymbol{y}_{t+1} - \hat{\boldsymbol{y}}_{t+1|t}, \tag{90b}$$

$$\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1}\boldsymbol{H}'\boldsymbol{\Sigma}_{\boldsymbol{e}t}^{-1}, \tag{99b}$$

$$\boldsymbol{\Sigma}_{\boldsymbol{e}t} = \boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}, \tag{99c}$$

$$\boldsymbol{P}_{t|t-1} = \boldsymbol{F}\boldsymbol{P}_{t-1|t-2}\boldsymbol{L}'_{t-1} + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \tag{93a}$$

$$\text{and} \quad \boldsymbol{L}_{t-1} = \boldsymbol{F}\left(\boldsymbol{I} - \boldsymbol{K}_{t-1}\boldsymbol{H}\right). \tag{94a}$$

Kalman filter recursive algorithm:

1. Starting with $\boldsymbol{P}_{1|0}$, compute equation (99c) and (99b) to get $\boldsymbol{K}_1$.
2. Plug $\boldsymbol{K}_1$ into (93a) to obtain $\boldsymbol{P}_{2|1}$.
3. Plug $\hat{\boldsymbol{x}}_{1|0}$ into (89b) to obtain $\hat{\boldsymbol{y}}_{1|0}$.
4. Plug $\hat{\boldsymbol{y}}_{1|0}$ into (90b) to obtain $\boldsymbol{e}_1$.
5. Finally, plug $\boldsymbol{K}_1$ and $\boldsymbol{e}_1$ into (98a) to obtain the filtered value $\hat{\boldsymbol{x}}_{1|1}$.
6. Plugging $\hat{\boldsymbol{x}}_{1|1}$ into (89a) then yields $\hat{\boldsymbol{x}}_{2|1}$. Since we already have $\boldsymbol{P}_{2|1}$ from step (2) above, we can now continue from step (1), ultimately repeating all the steps until we solve for $\hat{\boldsymbol{x}}_{t|t}$ for some desired $t \leq T$.

7.4.1 Kalman filter in terms of orthogonal basis

Note that an intuitive way to approach the solution to $\hat{\boldsymbol{x}}_{t|t}$ is to write it in terms of an orthogonal basis, where the basis vectors are the forecast errors $\boldsymbol{e}_\tau = \boldsymbol{y}_\tau - \hat{\boldsymbol{y}}_{\tau|\tau-1}, \;\; \tau = 1, \ldots, t$. This

allows us to appreciate how the Kalman filter is really the time domain analog to the frequency domain solution which attenuates frequencies where the signal-to-noise ratio is low.

In order to construct the appropriate orthogonal basis, we start with the requirement that the innovations $\boldsymbol{u}_t = \boldsymbol{x}_t - \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}]$, $\boldsymbol{e}_t = \boldsymbol{y}_t - \mathcal{P}[\boldsymbol{y}_t|\boldsymbol{y}_{t-1}]$, and finally $\boldsymbol{y}_{t-1}$ itself should all be uncorrelated with each other. Since their joint covariance matrix is therefore diagonal this implies that:

$$E[\boldsymbol{b}\boldsymbol{b}'] = \boldsymbol{D}, \quad \text{where} \quad \boldsymbol{b}' = \begin{bmatrix} \boldsymbol{y}_{t-1}' & \boldsymbol{e}_t' & \boldsymbol{u}_t' \end{bmatrix}. \tag{95a}$$

$$\text{Let} \quad \boldsymbol{b}^* \equiv \boldsymbol{A}\boldsymbol{b}, \quad \text{where} \quad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{A}_{21} & \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{A}_{31} & \boldsymbol{A}_{32} & \boldsymbol{I} \end{bmatrix} \tag{95b}$$

$$\Leftrightarrow E[\boldsymbol{b}^*\boldsymbol{b}^{*'}] = \boldsymbol{A}\boldsymbol{D}\boldsymbol{A}' = \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} & \boldsymbol{\Omega}_{13} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} & \boldsymbol{\Omega}_{23} \\ \boldsymbol{\Omega}_{31} & \boldsymbol{\Omega}_{32} & \boldsymbol{\Omega}_{33} \end{bmatrix}. \tag{95c}$$

It can be shown that:

$$\boldsymbol{\Omega}' = \boldsymbol{\Omega}, \tag{96a}$$

$$\boldsymbol{A}_{i1} = \boldsymbol{\Omega}_{i1}\boldsymbol{\Omega}_{11}^{-1}, \quad \forall i = \{2, 3\}, \tag{96b}$$

$$\text{and} \quad \boldsymbol{A}_{32} = \left(\boldsymbol{\Omega}_{32} - \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\right)\left(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\right)^{-1}. \tag{96c}$$

The first row of $\boldsymbol{A}\boldsymbol{b}$ is $\boldsymbol{y}_{t-1} = \boldsymbol{y}_{t-1}$; however, the second row is $\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{y}_{t-1} + \boldsymbol{e}_t = \boldsymbol{y}_t$, since $\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{y}_{t-1} = \mathcal{P}[\boldsymbol{y}_t|\boldsymbol{y}_{t-1}]$ and $E[\boldsymbol{e}_t\boldsymbol{y}_{t-1}'] = 0$ from (95a) implies it (alternatively, note that it must be true by the definition of $\boldsymbol{e}_t$ given in (90b)). However, it is the third row of $\boldsymbol{A}\boldsymbol{b}$ that is of the most interest since it provides us with a way to "update" a linear projection given new

information. The third row is given as:

$$\boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{y}_{t-1} + \left(\boldsymbol{\Omega}_{32} - \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\right)\left(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\right)^{-1}\boldsymbol{e}_t + \boldsymbol{u}_t = \boldsymbol{x}_t \tag{97a}$$

$$\Leftrightarrow \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{y}_{t-1} + \left(\boldsymbol{\Omega}_{32} - \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\right)\left(\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}\right)^{-1}\boldsymbol{e}_t = \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}], \tag{97b}$$

which follows from the definition of $\boldsymbol{u}_t$.

Notice, however, that $E[\boldsymbol{e}_t\boldsymbol{e}_t'] = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} = \boldsymbol{D}_{22}$ is simply the variance of the observed prediction error $\boldsymbol{e}_t$.[9] Similarly, $E[\boldsymbol{f}_t\boldsymbol{e}_t'] = \boldsymbol{\Omega}_{32} - \boldsymbol{\Omega}_{31}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}$ where $\boldsymbol{f}_t = \boldsymbol{x}_t - \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] = \boldsymbol{x}_t - \hat{\boldsymbol{x}}_{t|t-1}$ represents the state forecast error. Furthermore, $E[\boldsymbol{f}_{t|t}\boldsymbol{f}_{t|t}'] = \boldsymbol{\Upsilon}_{33} - \boldsymbol{\Upsilon}_{32}\boldsymbol{\Upsilon}_{22}^{-1}\boldsymbol{\Upsilon}_{23} = \boldsymbol{D}_{33}$, given $\boldsymbol{\Upsilon}_{ij} = \boldsymbol{\Omega}_{ij} - \boldsymbol{\Omega}_{i1}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{1j}$, represents the variance of the state filtered error $\boldsymbol{f}_{t|t} = \boldsymbol{x}_t - \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}] = \boldsymbol{x}_t - \hat{\boldsymbol{x}}_{t|t}$. Therefore, what we have done in essence, is to derive the Kalman filter by means of orthogonal projectors and this is what (97) represents:

$$\mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_t, \boldsymbol{y}_{t-1}] = \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] + E[\boldsymbol{f}_t\boldsymbol{e}_t']E[\boldsymbol{e}_t\boldsymbol{e}_t']^{-1}\boldsymbol{e}_t \tag{98a}$$

$$\equiv \hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{e}_t \tag{98b}$$

To derive $\boldsymbol{K}_t$ in terms of the system matrices, we have:

$$\boldsymbol{K}_t = E[\boldsymbol{f}_t\boldsymbol{e}_t']E[\boldsymbol{e}_t\boldsymbol{e}_t']^{-1} \tag{99a}$$

$$= E[\boldsymbol{f}_t\left(\boldsymbol{H}\boldsymbol{f}_t + \boldsymbol{\eta}_t\right)']E[\left(\boldsymbol{H}\boldsymbol{f}_t + \boldsymbol{\eta}_t\right)\left(\boldsymbol{H}\boldsymbol{f}_t + \boldsymbol{\eta}_t\right)']^{-1}$$

$$= E[\boldsymbol{f}_t\boldsymbol{f}_t'\boldsymbol{H}' + \boldsymbol{f}_t\boldsymbol{\eta}_t']E[\boldsymbol{H}\boldsymbol{f}_t\boldsymbol{f}_t'\boldsymbol{H}' + \boldsymbol{H}\boldsymbol{f}_t\boldsymbol{\eta}_t' + \boldsymbol{\eta}_t\boldsymbol{f}_t'\boldsymbol{H}' + \boldsymbol{\eta}_t\boldsymbol{\eta}_t']^{-1}$$

$$\equiv \boldsymbol{P}_{t|t-1}\boldsymbol{H}'\left[\boldsymbol{H}\boldsymbol{P}_{t|t-1}\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right]^{-1} \tag{99b}$$

$$= \boldsymbol{P}_{t|t-1}\boldsymbol{H}'\boldsymbol{\Sigma}_{\boldsymbol{e},t}^{-1} \tag{99c}$$

so $\boldsymbol{P}_{t|t-1}$ is the state forecast error covariance and $\boldsymbol{\Sigma}_{\boldsymbol{e},t} = E[\boldsymbol{e}_t\boldsymbol{e}_t']$.

Note that the difference equation (98) represents the filtered value of the first-moment of the

[9] Or equivalently, the MSE of the projection.

state given information up to time t. Similarly, the filtered second-moment can be derived as:

$$Var[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] = \boldsymbol{P}_{t|t-1} = \boldsymbol{F}Var[\boldsymbol{x}_{t-1}|\boldsymbol{y}_{t-1}]\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \boldsymbol{F}\boldsymbol{P}_{t-1|t-1}\boldsymbol{F}' + \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \tag{100a}$$

$$\text{and} \quad \boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{\Sigma}_{\boldsymbol{e}t}\boldsymbol{K}_t' = \boldsymbol{D}_{33}, \tag{100b}$$

which when combined together represent a matrix difference equation (the ARE or algebraic Riccati equation) [see equation (93a)) which can be solved for a "steady-state" value of $\boldsymbol{P}_\infty$ given sufficient stability conditions; see also Simon (2006, pg.194-199)].

Note that (100b) follows from (98b):

$$\boldsymbol{x}_t - \hat{\boldsymbol{x}}_{t|t} = \boldsymbol{x}_t - \left(\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{e}_t\right) \tag{101a}$$

$$\begin{aligned}
\Leftrightarrow MSE[\hat{\boldsymbol{x}}_{t|t}] &= \boldsymbol{P}_{t|t} \\
&= E[\boldsymbol{f}_t\boldsymbol{f}_t'] - E[\boldsymbol{K}_t\boldsymbol{e}_t\boldsymbol{e}_t'\boldsymbol{K}_t'] \\
&= \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t\boldsymbol{\Sigma}_{\boldsymbol{e}t}\boldsymbol{K}_t'.
\end{aligned}$$

The usefulness of the orthogonal projections type of approach to deriving the Kalman filter is that we can immediately decompose the optimal filtered state $\hat{\boldsymbol{x}}_{t|t}$ according to its orthogonal basis vectors, $\boldsymbol{e}_\tau, \forall\tau = t, t-1, t-2, \ldots$. First, consider starting from time t and recursively substituting $\hat{\boldsymbol{x}}_{t|t-1} = \boldsymbol{F}\hat{\boldsymbol{x}}_{t-1|t-1}$ into (98):

$$\hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{e}_t \tag{102a}$$

$$\begin{aligned}
&= \boldsymbol{F}\left(\hat{\boldsymbol{x}}_{t-1|t-2} + \boldsymbol{K}_{t-1}\boldsymbol{e}_{t-1}\right) + \boldsymbol{K}_t\boldsymbol{e}_t \\
&= \boldsymbol{F}\boldsymbol{F}\left(\hat{\boldsymbol{x}}_{t-2|t-3} + \boldsymbol{K}_{t-2}\boldsymbol{e}_{t-2}\right) + \boldsymbol{F}\boldsymbol{K}_{t-1}\boldsymbol{e}_{t-1} + \boldsymbol{K}_t\boldsymbol{e}_t \\
&\vdots \\
&= \boldsymbol{F}^j\left(\hat{\boldsymbol{x}}_{t-j|t-j-1} + \boldsymbol{K}_{t-j}\boldsymbol{e}_{t-j}\right) + \sum_{u=0}^{j-1}\boldsymbol{F}^u\boldsymbol{K}_{t-u}\boldsymbol{e}_{t-u}
\end{aligned} \tag{102b}$$

so as $j \to \infty$ in (102b), if the eigenvalues of $\boldsymbol{F}$ are less than 1 in modulus, we have that:

$$\hat{\boldsymbol{x}}_{t|t} = \sum_{u=0}^{\infty} \boldsymbol{F}^u \boldsymbol{K}_{t-u} \boldsymbol{e}_{t-u}. \tag{103}$$

But this is nothing more than a exponentially weighted moving average (EWMA)! In other words, the Kalman filtered state is simply weak white noise, $\boldsymbol{e}_t$, passed through a low-pass filter (see Section 7.4.2 below for further discussion on this point).

Moreover, we can also represent $\hat{\boldsymbol{x}}_{t|t}$ in terms of the basis Y_t, which is useful for interpreting the observed values as the "input" of the filter. From (98b) we have:

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t|t} &= \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t \left(\boldsymbol{y}_t - \boldsymbol{H} \hat{\boldsymbol{x}}_{t|t-1} \right) \\
&= \left(\boldsymbol{I} - \boldsymbol{K}_t \boldsymbol{H} \right) \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t \boldsymbol{y}_t, \quad \text{so let} \quad \hat{\boldsymbol{L}}_t = \boldsymbol{I} - \boldsymbol{K}_t \boldsymbol{H}. \qquad (104a) \\
\Leftrightarrow \quad \hat{\boldsymbol{x}}_{t|t} &= \hat{\boldsymbol{L}}_t \boldsymbol{F} \hat{\boldsymbol{x}}_{t-1|t-1} + \boldsymbol{K}_t \boldsymbol{y}_t \\
&= \hat{\boldsymbol{L}}_t \boldsymbol{F} \left(\hat{\boldsymbol{L}}_{t-1} \hat{\boldsymbol{x}}_{t-1|t-2} + \boldsymbol{K}_{t-1} \boldsymbol{y}_{t-1} \right) + \boldsymbol{K}_t \boldsymbol{y}_t \\
&= \hat{\boldsymbol{L}}_t \boldsymbol{L}_{t-1} \boldsymbol{F} \left(\hat{\boldsymbol{L}}_{t-2} \hat{\boldsymbol{x}}_{t-2|t-3} + \boldsymbol{K}_{t-2} \boldsymbol{y}_{t-2} \right) + \hat{\boldsymbol{L}}_t \boldsymbol{F} \boldsymbol{K}_{t-1} \boldsymbol{y}_{t-1} + \boldsymbol{K}_t \boldsymbol{y}_t \\
&= \hat{\boldsymbol{L}}_t \boldsymbol{L}_{t-1} \boldsymbol{L}_{t-2} \hat{\boldsymbol{x}}_{t-2|t-3} + \hat{\boldsymbol{L}}_t \boldsymbol{L}_{t-1} \boldsymbol{F} \boldsymbol{K}_{t-2} \boldsymbol{y}_{t-2} + \hat{\boldsymbol{L}}_t \boldsymbol{F} \boldsymbol{K}_{t-1} \boldsymbol{y}_{t-1} + \boldsymbol{K}_t \boldsymbol{y}_t \\
&\vdots \\
&= \hat{\boldsymbol{L}}_t \left(\prod_{k=1}^{j} \boldsymbol{L}_{t-k} \right) \hat{\boldsymbol{x}}_{t-j|t-(j+1)} \\
&\quad + \hat{\boldsymbol{L}}_t \sum_{u=2}^{j} \left(\prod_{k=1}^{u-1} \boldsymbol{L}_{t-k} \right) \boldsymbol{F} \boldsymbol{K}_{t-u} \boldsymbol{y}_{t-u} + \hat{\boldsymbol{L}}_t \boldsymbol{F} \boldsymbol{K}_{t-1} \boldsymbol{y}_{t-1} + \boldsymbol{K}_t \boldsymbol{y}_t. \qquad (104b)
\end{aligned}$$

where the third equality after the implication holds since $\boldsymbol{L}_{t-1} = \boldsymbol{F} \left(\boldsymbol{I} - \boldsymbol{K}_{t-1} \boldsymbol{H} \right) = \boldsymbol{F} \hat{\boldsymbol{L}}_{t-1}$.

So, if as $j \to \infty$ we have that $\boldsymbol{P}_{t|t-1} \to \boldsymbol{P}_\infty$, and the eigenvalues of $\boldsymbol{L}_\infty$ are less than 1 in modulus, then we can write:

$$\hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{L}} \sum_{u=0}^{\infty} \boldsymbol{L}^u \boldsymbol{F} \boldsymbol{K} \boldsymbol{y}_{t-1-u} + \boldsymbol{K} \boldsymbol{y}_t \tag{105}$$

which implies directly from (89a) that:

$$\hat{\boldsymbol{x}}_{t+1|t} = \sum_{u=0}^{\infty} \boldsymbol{L}^u \boldsymbol{F} \boldsymbol{K} \boldsymbol{y}_{t-u}. \tag{106}$$

And so we can see that the steady-state Kalman solution represents a linear time-invariant filter, with the observations $\{\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \dots\}$ as the filter input.

Similarly, we can solve for the response to a shock to the input process vector, $\boldsymbol{\epsilon}_t$, by deriving the impulse response matrices:

$$\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{\eta}_t && \text{(107a)} \\
&= \boldsymbol{H}\left(\boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t\right) + \boldsymbol{\eta}_t \\
&= \boldsymbol{H}\left(\boldsymbol{F}\left(\boldsymbol{F}\boldsymbol{x}_{t-2} + \boldsymbol{\epsilon}_{t-1}\right) + \boldsymbol{\epsilon}_t\right) + \boldsymbol{\eta}_t \\
&\vdots \\
&= \boldsymbol{H}\boldsymbol{F}^j\left(\boldsymbol{F}\boldsymbol{x}_{t-(j+1)} + \boldsymbol{\epsilon}_{t-j}\right) + \sum_{u=0}^{j-1} \boldsymbol{H}\boldsymbol{F}^u \boldsymbol{\epsilon}_{t-u} + \boldsymbol{\eta}_t && \text{(107b)}
\end{aligned}$$

so as $j \to \infty$ in (107b), if the eigenvalues of $\boldsymbol{F}$ are less than 1 in modulus, we have that:

$$\boldsymbol{y}_t = \sum_{u=0}^{\infty} \boldsymbol{H}\boldsymbol{F}^u \boldsymbol{\epsilon}_{t-u} + \boldsymbol{\eta}_t \tag{108}$$

therefore, $\boldsymbol{H}\boldsymbol{F}^u$ represents the impulse response to a unit shock at time t, u periods later. Note that this result is equivalent to (86) above, and so we can see that $\boldsymbol{y}_t$ can be interpreted as the result of filtering the weak white noise, $\boldsymbol{\epsilon}_t$, which drives the state, plus a current observation noise $\boldsymbol{\eta}_t$.

7.4.2 Spectral properties of Kalman filter

Equation (108) provides an expression for the coefficient matrices making up the impulse response, $\boldsymbol{A}(L)$. The spectral density $\boldsymbol{h}_{yy}(\omega)$ and power spectrum $|\boldsymbol{A}(e^{-i\omega})|^2$ are given respectively

from the associated z-transform with the same coefficients, at $z = e^{-i\omega}$, as: [10]

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{A}(L)\boldsymbol{\epsilon}_t + \boldsymbol{\eta}_t \\ &= \left(\boldsymbol{H}\left[\boldsymbol{I} - \boldsymbol{F}L\right]^{-1}\right)\boldsymbol{\epsilon}_t + \boldsymbol{\eta}_t \qquad (109a) \\ \Leftrightarrow \boldsymbol{h}_{yy}(\omega) &= \frac{1}{2\pi}\left(\boldsymbol{A}(e^{-i\omega})\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{A}(e^{+i\omega})' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}\right) \qquad (109b) \\ \Leftrightarrow |\boldsymbol{A}(e^{-i\omega})|^2 &= \left[\boldsymbol{H}\left[\boldsymbol{I} - \boldsymbol{F}e^{-i\omega}\right]^{-1}\right]\left[\boldsymbol{H}\left[\boldsymbol{I} - \boldsymbol{F}e^{+i\omega}\right]^{-1}\right]' \qquad (109c) \end{aligned}$$

Furthermore, to be more specific about the nature of the Kalman filter's frequency response let us consider taking the Fourier transform of a z-transform with the same coefficients as the impulse response function from (103):

$$\begin{aligned} \hat{\boldsymbol{x}}_{t|t} &= \boldsymbol{A}_t(L)\boldsymbol{e}_t \\ &= \left[\boldsymbol{I} - \boldsymbol{F}L\right]^{-1}\boldsymbol{K}_t\boldsymbol{e}_t \qquad (110a) \\ \Leftrightarrow \quad |\boldsymbol{A}_t(e^{-i\omega})|^2 &= \left[\boldsymbol{I} - \boldsymbol{F}e^{-i\omega}\right]^{-1}\boldsymbol{K}_t\boldsymbol{K}_t'\left[\boldsymbol{I} - \boldsymbol{F}e^{+i\omega}\right]^{-1'} \qquad (110b) \end{aligned}$$

The Kalman filter attenuates the high frequencies and passes through the lower ones in predicting the state from the observation errors, given information Y_t. The shape of the power spectrum will depend on both the eigenvalues of $\boldsymbol{K}_t$ and $\boldsymbol{F}$. Interestingly, the optimal filter weights (and thus frequency response) are computed easily for each t, since they depend only on $\boldsymbol{K}_t$ in the linear time invariant case. This contrasts to the Wiener solution where the transfer function needs to be recomputed independently for every change in t.

In terms of the Y_t basis we can interpret the $\boldsymbol{y}_t$'s as "inputs" passing through the filter. From

[10] As in the example given in Section 2.2.4, we have that since the z-transform $\boldsymbol{A}(z) = \sum_{u=0}^{\infty}\boldsymbol{a}_u z^u$, where $z \in \mathbb{C}$, admits the same coefficients as $\boldsymbol{A}(L)$, we can solve for the transfer function as $\boldsymbol{A}(z)$ where $z = e^{-i\omega}$.

(105), and assuming the steady-state filter, we have:

$$\begin{aligned}\hat{\boldsymbol{x}}_{t|t} &= \boldsymbol{A}(L)\boldsymbol{y}_t \\ &= \left(\hat{\boldsymbol{L}}\left[\boldsymbol{I}-\boldsymbol{L}L\right]^{-1}\boldsymbol{F}\boldsymbol{K}L+\boldsymbol{K}\right)\boldsymbol{y}_t \qquad (111a)\\ \Leftrightarrow\quad |\boldsymbol{A}(e^{-i\omega})|^2 &= \left(\hat{\boldsymbol{L}}\left[\boldsymbol{I}-\boldsymbol{L}e^{-i\omega}\right]^{-1}\boldsymbol{F}\boldsymbol{K}e^{-i\omega}+\boldsymbol{K}\right) \\ &\qquad \left(\hat{\boldsymbol{L}}\left[\boldsymbol{I}-\boldsymbol{L}e^{+i\omega}\right]^{-1}\boldsymbol{F}\boldsymbol{K}e^{+i\omega}+\boldsymbol{K}\right)'. \qquad (111b)\end{aligned}$$

All the transfer functions and power spectrums above are periodic matrix functions, with period 2π in their argument ω.

7.5 The MA and AR representations of the state-space model

Another useful representation of (1) can be obtained by rewriting the observation and state transition equations in terms of the observed forecast errors, $\boldsymbol{e}_t$. First, add $\boldsymbol{y}_{t+1}$ to both sides of (89b) to get:

$$\boldsymbol{y}_{t+1} = \boldsymbol{H}\hat{\boldsymbol{x}}_{t+1|t} + \boldsymbol{e}_{t+1}. \qquad (112)$$

From (104a) we have that:

$$\begin{aligned}\hat{\boldsymbol{x}}_{t+1|t} &= \boldsymbol{L}_t\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{F}\boldsymbol{K}_t\boldsymbol{y}_t \\ &= \boldsymbol{F}\hat{\boldsymbol{x}}_{t|t-1} - \boldsymbol{F}\boldsymbol{K}_t\boldsymbol{H}\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{F}\boldsymbol{K}_t\boldsymbol{y}_t \\ &= \boldsymbol{F}\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{F}\boldsymbol{K}_t\boldsymbol{e}_t. \qquad (113a)\end{aligned}$$

Therefore, together, (112) and (113a) represent the state-space model in (1) where both the observation and state transition noise have been replaced by the one-step ahead observation forecast errors.

The form of the state-space model in (112) and (113a) suggests both an MA and AR representation, where the $\boldsymbol{e}_t$'s represent weak white noise innovations. The MA representation is

derived by substituting (103) into (89a) and then plugging this into (112):

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{H}\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{e}_t \\ &= \boldsymbol{H}\left(\sum_{u=0}^{\infty} \boldsymbol{F}^{u+1}\boldsymbol{K}_{t-1-u}\boldsymbol{e}_{t-1-u}\right) + \boldsymbol{e}_t \end{aligned} \tag{112}$$

and again assuming that $\boldsymbol{P}_{t|t-1} \to \boldsymbol{P}_{\infty}$ exists, we can write:

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{H}\left(\boldsymbol{I} - \boldsymbol{F}L\right)^{-1}\boldsymbol{F}\boldsymbol{K}L\boldsymbol{e}_t + \boldsymbol{e}_t \\ \Leftrightarrow \boldsymbol{y}_t &= \boldsymbol{\Phi}(L)\boldsymbol{e}_t \qquad (115a) \\ \text{where} \quad \boldsymbol{\Phi}(L) &= \boldsymbol{I} + \boldsymbol{H}\left(\boldsymbol{I} - \boldsymbol{F}L\right)^{-1}\boldsymbol{F}\boldsymbol{K}L. \end{aligned}$$

Furthermore, the AR representation is derived from subbing (106) into (112):

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{H}\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{e}_t \qquad (112) \\ &= \boldsymbol{H}\left(\sum_{u=0}^{\infty} \boldsymbol{L}^{u}\boldsymbol{F}\boldsymbol{K}\boldsymbol{y}_{t-1-u}\right) + \boldsymbol{e}_t \end{aligned}$$

so we can write

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{H}\left(\boldsymbol{I} - \boldsymbol{L}L\right)^{-1}\boldsymbol{F}\boldsymbol{K}L\boldsymbol{y}_t + \boldsymbol{e}_t \\ \Leftrightarrow \boldsymbol{\Theta}(L)\boldsymbol{y}_t &= \boldsymbol{e}_t \qquad (117a) \\ \text{where} \quad \boldsymbol{\Theta}(L) &= \boldsymbol{I} - \boldsymbol{H}\left(\boldsymbol{I} - \boldsymbol{L}L\right)^{-1}\boldsymbol{F}\boldsymbol{K}L. \end{aligned}$$

Interestingly, if we assume that the MA representation is invertible, then we must have that:

$$\boldsymbol{\Phi}^{-1}(L) = \boldsymbol{\Theta}(L) \tag{118a}$$

$$\Leftrightarrow \left(\boldsymbol{I} + \boldsymbol{H}\left(\boldsymbol{I} - \boldsymbol{F}L\right)^{-1}\boldsymbol{F}\boldsymbol{K}L\right)^{-1} = \boldsymbol{I} - \boldsymbol{H}\left(\boldsymbol{I} - \boldsymbol{L}L\right)^{-1}\boldsymbol{F}\boldsymbol{K}L. \tag{118b}$$

7.6 Smoothing

The Kalman smoother is a generalization of the Kalman filter, where instead of computing $\mathcal{P}[\boldsymbol{x}_t|Y_t]$, the best linear predictor given information up to time t, we wish rather to use all the more recent information available to us up to some time $T > t$. Therefore, we wish to compute the best linear predictor of $\boldsymbol{x}_\tau$ given observed values, $\boldsymbol{y}_t$, $t = 1, \ldots, T$, where $\tau < T$; that is, we wish to compute $\mathcal{P}[\boldsymbol{x}_t|Y_T]$. Again, for simplicity of notation, let $\mathcal{P}[\boldsymbol{x}_t|Y_T] \equiv \hat{\boldsymbol{x}}_{t|T}$. The smoother is derived by a simple extension of (98a):

$$\mathcal{P}[\boldsymbol{x}_t|Y_t] = \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] + E[\boldsymbol{f}_t\boldsymbol{e}_t']E[\boldsymbol{e}_t\boldsymbol{e}_t']^{-1}\boldsymbol{e}_t \tag{98a}$$

$$\equiv \hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{e}_t \tag{98b}$$

which becomes:

$$\mathcal{P}[\boldsymbol{x}_t|Y_T] = \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_T, \ldots, \boldsymbol{y}_t, Y_{t-1}] = \mathcal{P}[\boldsymbol{x}_t|\boldsymbol{y}_{t-1}] + \sum_{k=t}^{T} E[\boldsymbol{f}_t\boldsymbol{e}_k']E[\boldsymbol{e}_k\boldsymbol{e}_k']^{-1}\boldsymbol{e}_k \tag{120a}$$

$$\equiv \hat{\boldsymbol{x}}_{t|T} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t\boldsymbol{e}_t + \sum_{k=t+1}^{T} \{\boldsymbol{P}_{t|t-1}\left(\prod_{j=t}^{k-1}\boldsymbol{L}_j'\right)\boldsymbol{H}'\}\boldsymbol{\Sigma}_{\boldsymbol{e}_k}^{-1}\boldsymbol{e}_k. \tag{120b}$$

Equation (120a) is therefore a generalization of (98a), where we wish to update the best linear forecast of the state variable with not just information from the current time period t, but also with all future information, $\tau = t+1, \ldots, T$. Alternatively, we can see that we have replaced $\boldsymbol{e}_t$ in (95a) with the stacked vector $\boldsymbol{e}_t^{*T} = \begin{bmatrix}\boldsymbol{e}_t' & \boldsymbol{e}_{t+1}' & \ldots & \boldsymbol{e}_{T-1}' & \boldsymbol{e}_T'\end{bmatrix}'$, which suggests a generalized

form of (97b). Moreover, (120b) follows from:

$$\begin{aligned} E[\boldsymbol{f}_t \boldsymbol{e}'_{t+1}] &= E[\boldsymbol{f}_t \left(\boldsymbol{H}\boldsymbol{f}_{t+1} + \boldsymbol{\eta}_{t+1}\right)'] \\ &= E[\boldsymbol{f}_t \boldsymbol{f}'_{t+1}\boldsymbol{H}' + \boldsymbol{f}_t \boldsymbol{\eta}'_{t+1}] \\ &= E[\boldsymbol{f}_t \boldsymbol{f}'_{t+1}]\boldsymbol{H}' \\ &= \boldsymbol{P}_{t|t-1}\boldsymbol{L}'_t \boldsymbol{H}' \end{aligned} \tag{121a}$$

$$\begin{aligned} \Leftrightarrow E[\boldsymbol{f}_t \boldsymbol{e}'_{t+2}] &= \boldsymbol{P}_{t|t-1}\boldsymbol{L}'_t \boldsymbol{L}'_{t+1}\boldsymbol{H}' \\ &\vdots \\ E[\boldsymbol{f}_t \boldsymbol{e}'_{T}] &= \boldsymbol{P}_{t|t-1}\boldsymbol{L}'_t \boldsymbol{L}'_{t+1} \dots \boldsymbol{L}'_{T-2}\boldsymbol{L}'_{T-1}\boldsymbol{H}', \end{aligned} \tag{121b}$$

where equation (121a) follows from:

$$\begin{aligned} \boldsymbol{f}_{t+1} &= \boldsymbol{x}_{t+1} - \hat{\boldsymbol{x}}_{t+1|t} \\ &= \boldsymbol{x}_{t+1} - \boldsymbol{F}\hat{\boldsymbol{x}}_{t|t} \\ &= \boldsymbol{x}_{t+1} - \boldsymbol{F}\left(\hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t \boldsymbol{e}_t\right) + \boldsymbol{F}\boldsymbol{x}_t - \boldsymbol{F}\boldsymbol{x}_t \\ &= \boldsymbol{F}\left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_{t|t-1}\right) + \boldsymbol{\epsilon}_{t+1} - \boldsymbol{F}\boldsymbol{K}_t \boldsymbol{e}_t \\ &= \boldsymbol{F}\boldsymbol{f}_t + \boldsymbol{\epsilon}_{t+1} - \boldsymbol{F}\boldsymbol{K}_t\left(\boldsymbol{H}\boldsymbol{f}_t + \boldsymbol{\eta}_t\right) \\ &= \left(\boldsymbol{F}\left(\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H}\right)\right)\boldsymbol{f}_t + \boldsymbol{\epsilon}_{t+1} - \boldsymbol{F}\boldsymbol{K}_t\boldsymbol{\eta}_t \\ &= \boldsymbol{L}_t\boldsymbol{f}_t + \boldsymbol{\epsilon}_{t+1} - \boldsymbol{F}\boldsymbol{K}_t\boldsymbol{\eta}_t. \end{aligned} \tag{122a}$$

Note that from De Jong (1989) we can re-express (120b) in terms of the following recursion:

$$\boldsymbol{q}_{t-1} = \boldsymbol{L}'_t\boldsymbol{q}_t + \boldsymbol{H}'{\boldsymbol{\Sigma}_{e_t}}^{-1}\boldsymbol{e}_t, \quad \text{where} \quad t = T, \dots, 1. \tag{123}$$

That is, we can write (120b) as:

$$\hat{\boldsymbol{x}}_{t|T} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{P}_{t|t-1}\boldsymbol{q}_{t-1}. \tag{124}$$

Moreover, from (124) (or alternatively from the same method employed in (101a)) we have that:

$$\boldsymbol{P}_{t|T} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} Var[\boldsymbol{q}_{t-1}] \boldsymbol{P}_{t|t-1} \tag{125a}$$

$$\text{where} \quad Var[\boldsymbol{q}_{t-1}] \equiv \boldsymbol{M}_{t-1} = \boldsymbol{H}' {\boldsymbol{\Sigma}_{\boldsymbol{e}_t}}^{-1} \boldsymbol{H} + \boldsymbol{L}_t' \boldsymbol{M}_t \boldsymbol{L}_t \tag{125b}$$

noting that $\boldsymbol{P}_{t|T} = \boldsymbol{P}'_{t|T}$.

Therefore, (125) implies that:

$$\begin{aligned}
\boldsymbol{P}_{t|T} &= \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} \left(\boldsymbol{H}' {\boldsymbol{\Sigma}_{\boldsymbol{e}_t}}^{-1} \boldsymbol{H} + \boldsymbol{L}_t' \boldsymbol{M}_t \boldsymbol{L}_t \right) \boldsymbol{P}_{t|t-1} \\
&= \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} \boldsymbol{H}' {\boldsymbol{\Sigma}_{\boldsymbol{e}_t}}^{-1} \boldsymbol{H} \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} \boldsymbol{L}_t' \boldsymbol{M}_t \boldsymbol{L}_t \boldsymbol{P}_{t|t-1} \\
&= \boldsymbol{P}_{t|t-1} - \boldsymbol{K}_t \boldsymbol{\Sigma}_{\boldsymbol{e}_t} \boldsymbol{K}_t' - \boldsymbol{P}_{t|t-1} \boldsymbol{L}_t' \boldsymbol{M}_t \boldsymbol{L}_t \boldsymbol{P}_{t|t-1} \\
&= \boldsymbol{P}_{t|t} - \boldsymbol{P}_{t|t-1} \boldsymbol{L}_t' \boldsymbol{M}_t \boldsymbol{L}_t \boldsymbol{P}_{t|t-1}.
\end{aligned} \tag{126a}$$

So $\boldsymbol{P}_{t|t-1} \boldsymbol{L}_t' \boldsymbol{M}_t \boldsymbol{L}_t \boldsymbol{P}_{t|t-1}$ represents the reduction in MSE the smoother represents over the filter, since it uses the extra information embodied in $\{\boldsymbol{y}_T, \boldsymbol{y}_{T-1}, \ldots, \boldsymbol{y}_{t+1}\}$.

What follows is known as the "fixed interval smoother" algorithm. This algorithm covers the case where T is fixed, and we desire the smoothed $\hat{\boldsymbol{x}}_{t|T}$ for any $t \leq T$. This is different from the "fixed-point smoother," where t remains fixed and T increases, or the "fixed-lag smoother" where both t and T vary but their difference $T - t$ remains fixed.

Fixed Interval Smoother algorithm

1. Run the "Kalman filter algorithm" described in Section 7.4 to obtain $\hat{\boldsymbol{x}}_{t|t} = \hat{\boldsymbol{x}}_{t|t-1} + \boldsymbol{K}_t \boldsymbol{e}_t$.

2. Since we already have $\boldsymbol{P}_{t+1|t}$ from step (1), use equation (99c) to compute $\boldsymbol{\Sigma}_{\boldsymbol{e}_{t+1}}^{-1}$ and (99b) to get $\boldsymbol{K}_{t+1}$.

3. Plug $\boldsymbol{K}_{t+1}$ into (93a) to obtain $\boldsymbol{P}_{t+2|t+1}$. $\boldsymbol{K}_{t+1}$ can then be used to compute $\boldsymbol{L}_{t+1}$.

4. Plug $\hat{\boldsymbol{x}}_{t+1|t}$ [which we already have from step (1) and (89a)] into (89b) to obtain $\hat{\boldsymbol{y}}_{t+1|t}$.

5. Plug $\hat{\boldsymbol{y}}_{t+1|t}$ into (90b) to obtain $\boldsymbol{e}_{t+1}$.

6. Since we have $\boldsymbol{P}_{t+2|t+1}$ from step (3) above, we can now start again from step (2), ultimately repeating all the steps until we solve for $\boldsymbol{P}_{T|T-1}$. Once we've computed all the values of $\boldsymbol{e}_\tau$, $\boldsymbol{L}_{\tau-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{e}_\tau}^{-1}$ for all $\tau = t+1, t+2, \ldots, T$ we can then directly compute the sum in (120b).

7.6.1 Smoother in terms of orthogonal basis

Just as was done in Section 7.4.1, the fixed interval smoother can be represented in terms of an orthogonal basis. In fact, we have already done so–all that remains is to represent it here again in its entirety. This is useful since it allows us to appreciate the connection between the Kalman state-space approach to the prediction problem and that of the Wiener solution in the frequency domain.

The representation in terms of orthogonal basis is given from (89a) combined with (102b), and (120b):

$$\hat{\boldsymbol{x}}_{t|T} = \boldsymbol{F}^{j+1}\left(\hat{\boldsymbol{x}}_{t-1-j|t-2-j} + \boldsymbol{K}_{t-1-j}\boldsymbol{e}_{t-1-j}\right) + \sum_{u=0}^{j-1} \boldsymbol{F}^{u+1}\boldsymbol{K}_{t-1-u}\boldsymbol{e}_{t-1-u} + \boldsymbol{K}_t\boldsymbol{e}_t + \sum_{k=t+1}^{T} \boldsymbol{P}_{t|t-1}\left(\prod_{j=t}^{k-1}\boldsymbol{L}_j'\right)\boldsymbol{H}'\boldsymbol{\Sigma}_{\boldsymbol{e}_k}^{-1}\boldsymbol{e}_k \quad (127)$$

so the first and second components in the sum represent past information, the third represents current information (at time t), and the last component represents future information.

Therefore, provided the eigenvalues of $\boldsymbol{F}$ are less than 1 in modulus and $\boldsymbol{P}_{t|t-1} \to \boldsymbol{P}_\infty$,

allowing both j and T to approach ∞ results in the following expression of the impulse response function:

$$\begin{aligned}
\hat{\boldsymbol{x}}_{t|\infty} &= \sum_{u=0}^{\infty} \boldsymbol{F}^u \boldsymbol{F}\boldsymbol{K}\boldsymbol{e}_{t-1-u} + \boldsymbol{K}\boldsymbol{e}_t + \sum_{k=0}^{\infty} \boldsymbol{P}\left(\boldsymbol{L}'\right)^{k+1} \boldsymbol{H}'{\boldsymbol{\Sigma}_{\boldsymbol{e}}}^{-1}\boldsymbol{e}_{t+1+k} \\
&= \sum_{u=0}^{\infty} \boldsymbol{F}^u \boldsymbol{F}\boldsymbol{K}\boldsymbol{e}_{t-1-u} + \sum_{k=0}^{\infty} \boldsymbol{P}\left(\boldsymbol{L}'\right)^{k} \boldsymbol{H}'{\boldsymbol{\Sigma}_{\boldsymbol{e}}}^{-1}\boldsymbol{e}_{t+k} \\
&= \left[\boldsymbol{I} - \boldsymbol{F}L\right]^{-1} \boldsymbol{F}\boldsymbol{K}L\boldsymbol{e}_t + \boldsymbol{P}\left[\boldsymbol{I} - \boldsymbol{L}'L^{-1}\right]^{-1} \boldsymbol{H}'{\boldsymbol{\Sigma}_{\boldsymbol{e}}}^{-1}\boldsymbol{e}_t \\
&= \left(\left[\boldsymbol{I} - \boldsymbol{F}L\right]^{-1} \boldsymbol{F}\boldsymbol{K}L + \boldsymbol{P}\left[\boldsymbol{I} - \boldsymbol{L}'L^{-1}\right]^{-1} \boldsymbol{H}'{\boldsymbol{\Sigma}_{\boldsymbol{e}}}^{-1}\right)\boldsymbol{e}_t \qquad (128a) \\
&= \boldsymbol{C}(L)\boldsymbol{e}_t.
\end{aligned}$$

7.6.2 Spectral properties of the fixed interval smoother

The spectral properties of the filter $\boldsymbol{C}(L)$ in (128a) are available using the same methods as in 7.4.2 [see for example in equation (110a)] so they do not bear repeating.

However, what is of interest is the equivalence between the Wiener solution to the problem of signal extraction and that of the Kalman state-space smoother, as $T \to \infty$. I will not reproduce the entire equivalence proof here: interested readers can refer to Harvey & Proietti (2005). However, what can be said briefly is as follows.

First, take (128a) and replace $\boldsymbol{e}_t$ with its AR representation $\boldsymbol{\Theta}(\boldsymbol{L})\boldsymbol{y}_t$:

$$\hat{\boldsymbol{x}}_{t|\infty} = \boldsymbol{C}(L)\boldsymbol{\Theta}(L)\boldsymbol{y}_t$$

Now, using the Woodbury matrix inversion lemma, the fact that the autocovariance generating matrix of $\boldsymbol{y}_t$ is $\boldsymbol{\Gamma}_{\boldsymbol{yy}}(L) = \boldsymbol{\Phi}(L)\boldsymbol{\Sigma}_{\boldsymbol{e}}\boldsymbol{\Phi}(L^{-1})'$ (from Section 7.5), the assumption of $\boldsymbol{P}_{t|t-1} \to \boldsymbol{P}_{\infty}$

and the solution to the discrete Riccatti ARE, (93a), we have that:

$$\begin{aligned}\hat{\boldsymbol{x}}_{t|\infty} &= \boldsymbol{C}(L)\boldsymbol{\Theta}(L)\boldsymbol{y}_t \\ &= \boldsymbol{\Gamma}_{\boldsymbol{xx}}(L)\boldsymbol{H}'{\boldsymbol{\Gamma}_{\boldsymbol{yy}}}^{-1}(L)\boldsymbol{y}_t \\ &= \boldsymbol{\Gamma}_{\boldsymbol{xy}}(L){\boldsymbol{\Gamma}_{\boldsymbol{yy}}}^{-1}(L)\boldsymbol{y}_t \end{aligned} \tag{130a}$$

$$\text{where} \quad \boldsymbol{\Gamma}_{\boldsymbol{xx}}(L) = \frac{1}{2\pi}\left[\boldsymbol{I} - \boldsymbol{F}L\right]^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\left(\left[\boldsymbol{I} - \boldsymbol{F}L^{-1}\right]^{-1}\right)' \tag{130b}$$

$$\text{and} \quad \boldsymbol{\Gamma}_{\boldsymbol{yy}}(L) = \frac{1}{2\pi}\boldsymbol{H}\boldsymbol{\Gamma}_{\boldsymbol{xx}}(L)\boldsymbol{H}' + \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \tag{130c}$$

and where (130a) represents the Wiener solution in the time domain.

7.7 Conclusion

If we impose the stronger condition that the noise in (1), $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$, is Gaussian (and not simply weak white noise), then we can replace all of the linear projections discussed above, $\mathcal{P}[\cdot]$, with expectations, since in the linear Gaussian case they coincide. Moreover, if the noise is Gaussian and uncorrelated, then it is also independent, and so these models are strong form.

In contrast, consider the case where the noise in (1) is weak white but the higher order moments rule out independence (e.g. $E[\epsilon^k\eta^j] \neq 0$ for some $k, j > 1$). Even in this case, the Kalman filter is still the best unbiased predictor amongst the class of *linear* predictions – see Simon (2006) pg.130. However, it is clear that in this case a *nonlinear* predictor will be more efficient.

Finally, if the noise has finite variance, but is coloured (and we are correct in assuming that all the relevant higher order moments are zero), then we can always modify the model in (1) to take account of this – see Simon (2006) sections 7.1 and 7.2 for more details.

8 Estimation in the time domain

The most straightforward way to estimate the parameters of the model in (1) is to employ Maximum Likelihood in the time domain. Given that the model is linear, and assuming the observation forecast errors, $\boldsymbol{e}_t$, are Gaussian, their joint density can be factorized as:

$$p(\boldsymbol{e}_T, \boldsymbol{e}_{T-1}, \ldots, \boldsymbol{e}_1|\boldsymbol{\Theta}) = p(\boldsymbol{y}_T, \boldsymbol{y}_{T-1}, \ldots, \boldsymbol{y}_1|\boldsymbol{\Theta}) \tag{131a}$$

$$= p(\boldsymbol{y}_T|\boldsymbol{y}_{T-1}, \boldsymbol{\Theta})p(\boldsymbol{y}_{T-1}|\boldsymbol{y}_{T-2}, \boldsymbol{\Theta}) \ldots p(\boldsymbol{y}_2|\boldsymbol{y}_1, \boldsymbol{\Theta})p(\boldsymbol{y}_1|\boldsymbol{\Theta}) \tag{131b}$$

where $\boldsymbol{\Theta}$ is the parameter set. The first equality holds since $\boldsymbol{e}_t = \boldsymbol{y}_t - \hat{\boldsymbol{y}}_{t|t-1}$ implies that the change of variables has unit Jacobian determinant (this can be proved by simply writing out $\hat{\boldsymbol{y}}_{t|t-1}$ recursively all the way back to $\hat{\boldsymbol{x}}_{1|0}$ – any $\boldsymbol{e}_t$ only depends on *present and past* values of $\boldsymbol{y}_t$ and thus the Jacobian matrix will be block lower triangular with identity matrices on the diagonal).

Therefore, (131) implies that the log-likelihood of the model can be derived as:

$$p(\boldsymbol{e}_T, \boldsymbol{e}_{T-1}, \ldots, \boldsymbol{e}_1|\boldsymbol{\Theta}) = \prod_{t=1}^{T} f(\boldsymbol{e}_t|\boldsymbol{\Theta}) \tag{132a}$$

$$\propto \prod_{t=1}^{T} \frac{1}{det\left(\boldsymbol{P}_{t|t-1}\right)^{1/2}} exp\{-\frac{1}{2}\boldsymbol{e}_t'\boldsymbol{P}_{t|t-1}^{-1}\boldsymbol{e}_t\} \tag{132b}$$

$$\Leftrightarrow \mathcal{L}(\boldsymbol{\Theta}) = \sum_{t=1}^{T} -\frac{1}{2} ln\left(det\left(\boldsymbol{P}_{t|t-1}\right)\right) - \frac{1}{2}\boldsymbol{e}_t'\boldsymbol{P}_{t|t-1}^{-1}\boldsymbol{e}_t \tag{132c}$$

where the first equality holds since the predicted observation errors are Gaussian and thus their uncorrelatedness implies independence.

Therefore, given (132) and some initial starting value for $\boldsymbol{x}_{1|0}$, we can optimize the log-likelihood with respect to $\boldsymbol{\Theta}$ given the recursive state estimates $\hat{\boldsymbol{x}}_{t|t-1}, t = 1, \ldots, T$, the Kalman gain $\boldsymbol{K}_t$ and the estimated predicted state MSE, $\boldsymbol{P}_{t|t-1}$, for each choice of $\boldsymbol{\Theta}$.

If the predicted observation errors are not truly Gaussian, then this reduces to a Quasi-Maximum likelihood estimator which is still consistent asymptotically.

9 Factor models and common trends

Historically, the use of Factor models in Economics has largely involved their application to financial data. However, what may not be obvious to most practitioners, is that the Factor model is really just a special case of the general state-space model. This Section will briefly touch on the role of Factor models in finance and discuss both the traditional static Factor model and the newer dynamic Factor model representations. Moreover, we will show that the dynamic Factor model representation is intimately connected to the notion of co-integrated processes. In fact, the dynamic Factor model is equivalent to the co-integrated Vector Autoregressive Moving Average process (VARMA) or its equivalent Vector Error Correction (VECM) process representation.

9.1 Factor models in finance

The impetus for the popularity of Factor models in finance can probably be attributed to the seminal paper on Arbitrage Pricing Theory by Ross (1976). The CAPM model of Sharpe (1970) generated a linear security market line representation for excess returns, based on Markowitz mean-variance optimization and the implication of an efficient "market" portfolio. The "Sharpe ratio" implied that under equilibrium the beta of an asset, β, entirely determined its price proportional to the hypothetical market excess return. Concerned over the plausibility of the assumptions imposed by the CAPM and Markowitz portfolio theory, Ross (1976) aimed to reproduce the same linear representation for excess returns but where the model made no assumption of equilibrium behaviour. Since the model assumed only arbitrage as the force driving prices, the linear relationship could exist even under disequilibrium in the asset markets. In fact, it required no assumption of a "market" portfolio at all.

In order to formulate this linear model, Ross (1976) chose the Factor representation, where excess returns are explained as a linear function of some common factor and where an appropriately chosen arbitrage portfolio could still induce full diversification and elimination of the idiosyncratic risk component. This not only freed us from the need to model excess returns as a function of the hypothetical market return, but it allowed us to test for the predictive power of

other potential factors. Moreover, it presented us with a tractible way to explain excess returns using a limited number of factors, as opposed to calculating the direct covariance matrix between a multitude of assets as under Markowitz theory.

For example, Chen, Roll, and Ross (1986) introduced a multifactor model for stock returns, where the factors consist of unexpected, yet observable, changes in macroeconomic variables. Another famous example is the so-called 3-Factor model by Fama and French (1992), where they showed that beta only explained part of excess returns, and that book-to-market price ratios and market capitalization were also important explanatory factors.

In the previous examples, the factors chosen were observed (either as proxies or directly). However, factors can either be observed or unobserved (sometimes referred to as latent). If the factors chosen are unobserved then we must employ statistical methods to reconstitute the factors from the observable data. For the simple i.i.d. factor model, the traditional method of Principle Component Analysis can be used. For a dynamic factor model, the methods of state-space prediction discussed above in Section 7 are available.

9.2 The static Factor model

Traditionally, Factor models had treated the factors as static i.i.d. processes. The static Factor model for the observed variables $\boldsymbol{y}_t$ is given as:

$$\boldsymbol{y}_t = \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{\eta}_t \tag{133a}$$

$$\text{and} \quad \boldsymbol{f}_t = \boldsymbol{\epsilon}_t, \tag{133b}$$

$$\begin{aligned} \text{where} \quad \boldsymbol{\eta}_t &\sim IIN(0, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}), \\ \boldsymbol{\epsilon}_t &\sim IIN(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \\ \text{and} \quad 0 &= E[\boldsymbol{\eta}_{t-j}\boldsymbol{\epsilon}'_{t-s}], \ \ \forall j, s \in \mathbb{Z}, \end{aligned}$$

which is the same as the (strong form) model in (28) but where $\boldsymbol{H} \equiv \boldsymbol{B}$ and $\boldsymbol{x_t} \equiv \boldsymbol{f}_t$ and where we assume that:

1. The observed data has $E[\boldsymbol{y}_t] = 0$ and $\boldsymbol{y}_t$ is $n \times 1$ and weakly stationary with no serial correlation.

2. $E[\boldsymbol{f}_t] = 0$ and $E[\boldsymbol{f}_t \boldsymbol{f}_t'] = \boldsymbol{\Omega}$, such that $\boldsymbol{\Omega}$ is diagonal and $\boldsymbol{f}_t$ is a $k \times 1$ vector of common factors and $k < n$.

3. $\boldsymbol{B}$ is $n \times k$ and represents the factor loadings.

4. $\boldsymbol{\eta}_t$ is such that $E[\boldsymbol{\eta}_t] = 0$ and $E[\boldsymbol{\eta}_t \boldsymbol{\eta}_t'] = \boldsymbol{\Lambda_\eta}$ which is a diagonal matrix, $\boldsymbol{\Lambda_\eta} = diag\{\sigma_1^2, \ldots, \sigma_n^2\}$. Moreover, $\boldsymbol{\eta}_t$ is also $n \times 1$ and represents the specific error (often call the "idiosyncratic error").

5. Finally, $\boldsymbol{f}_t$ and $\boldsymbol{\eta}_t$ are independent.

Therefore, given the discussion in Section 5 we know that the optimal predictor of the unknown factor $\boldsymbol{f}_t$ is given as:

$$E[\boldsymbol{f}_t|\boldsymbol{y}_t] = E[\boldsymbol{f}_t \boldsymbol{y}_t'] E[\boldsymbol{y}_t \boldsymbol{y}_t']^{-1} \boldsymbol{y}_t \tag{134a}$$

$$= E[\boldsymbol{f}_t \left(\boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{\eta}_t\right)'] \left[\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}' + \boldsymbol{\Lambda_\eta}\right]^{-1} \boldsymbol{y}_t \tag{134b}$$

$$= \boldsymbol{\Omega}\boldsymbol{B}' \left[\boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}' + \boldsymbol{\Lambda_\eta}\right]^{-1} \boldsymbol{y}_t \tag{134c}$$

which is equivalent to the static state-space model prediction, (74c), provided in the Section 5 examples, or analogously, the steady-state Kalman gain solution in (99b) where the state covariance matrix is time invariant, $\boldsymbol{\Omega} \equiv \boldsymbol{P}_\infty$.

9.2.1 Estimation in the time domain: the static Factor model and PCA

Unobserved static factors can be statistically reconstituted by employing the method of principle components analysis (PCA). The aim of PCA is to identify, from the observed data, statistically

orthogonal factors that can account for most of the variation in the covariance matrix of $\boldsymbol{y}_t$. To do so, first let us write the covariance matrix of $\boldsymbol{y}_t$:

$$E[\boldsymbol{y}_t \boldsymbol{y}_t'] \equiv \boldsymbol{\Sigma}_{\boldsymbol{y}} = \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}' + \boldsymbol{\Lambda}_{\boldsymbol{\eta}}. \tag{135}$$

Since $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ is symmetric, real, and positive-definite, it admits a spectral decomposition $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}'$ where $\boldsymbol{L}$ has as its columns the orthonormal eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{y}}$ and $\boldsymbol{D}$ is diagonal with $\boldsymbol{D} = diag\{\lambda_1, \ldots, \lambda_n\}$, the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{y}}$, arranged in decreasing order from largest to smallest. This implies that $E[\boldsymbol{L}'\boldsymbol{y}_t\boldsymbol{y}_t'\boldsymbol{L}] = \boldsymbol{D}$ is diagonal since $\boldsymbol{L}'\boldsymbol{L} = \boldsymbol{I}$.

Thus, $\boldsymbol{L}'\boldsymbol{y}_t$ makes a perfect candidate for the unobserved factor $\boldsymbol{f}_t$ since it satisfies points (2) and (5) above in Section 9.2, and the i'th row of $\boldsymbol{L}'\boldsymbol{y}_t$ represents a linear combination of the observed data which accounts for some proportion $\frac{\lambda_i}{\sum_{j=1}^n \lambda_j}$ of the variance in $\boldsymbol{y}_t$, since $tr(\boldsymbol{\Sigma}_{\boldsymbol{y}}) = \sum_{j=1}^n \lambda_j$.

Now, suppose that we choose to only employ some k factors, such that $k < n$. What is the optimal choice of $\boldsymbol{B}$ (i.e. what value of $\boldsymbol{B}$ satisfies the optimal predictor of $\boldsymbol{y}_t$ in the sense of Section 5) given that $\boldsymbol{f}_t = \boldsymbol{L}^{*'}\boldsymbol{y}_t$ (where $\boldsymbol{L}^*$ denotes the matrix composed of only the first k eigenvectors of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$, the estimated unconditional covariance of $\boldsymbol{y}_t$) included as its columns? Under the constraint that $\boldsymbol{L}^{*'}\boldsymbol{L}^* = \boldsymbol{I}$ and the MMSE criterion, we have that the optimal value of $\boldsymbol{B}$ is the linear projection of $\boldsymbol{y}_t$ onto the space spanned by the columns of $\boldsymbol{f}_t = \boldsymbol{L}^{*'}\boldsymbol{y}_t$:

$$E[\boldsymbol{y_t}(\boldsymbol{y}_t'\boldsymbol{L}^*)]E[(\boldsymbol{L}^{*'}\boldsymbol{y}_t)(\boldsymbol{y}_t'\boldsymbol{L}^*)]^{-1} = \boldsymbol{B}, \tag{136a}$$

which can be estimated by sample moment counterpart. Moreover, since we have that $\boldsymbol{\eta}_t$ must lie in the space orthogonal to $\boldsymbol{B}$, we have that $\boldsymbol{\eta}_t = \boldsymbol{y}_t - E[\boldsymbol{y}_t\boldsymbol{f}_t']E[\boldsymbol{f}_t\boldsymbol{f}_t']^{-1}\boldsymbol{f}_t$. Therefore the estimate for the residual error is $\hat{\boldsymbol{\Lambda}}_{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{y}} - \boldsymbol{B}\hat{\boldsymbol{\Omega}}\boldsymbol{B}'$.

However, since simple PCA relies on the assumption that the dependent variable process $\boldsymbol{y}_t$ is weakly stationary with no serial correlation, the method's popularity amongst cross-sectional researchers has not been matched within the time series community.

Attempts at reconciling PCA with dynamic factor models can be made by employing a two-step procedure where first some static factors are estimated by PCA as above, and, in the second step, a VAR model is specified as the factor process – see Breitung & Eickmeier (2006). Another method involves estimating the dynamic principal components within the frequency domain by eigenvalue decomposition of the spectral density of $\boldsymbol{y}_t$, then an estimate of the common components is obtained through its inversion – see Brillinger (1981, ch.9).

The nice feature of the state-space approach is that it allows us to easily predict latent, dynamic, factors, by means of the Kalman filter. Moreover, the dynamic Factor model for "common trends," implied by the Engle-Granger representation theorem (1987) discussed below, manifests in a state-space representation. This makes the state-space framework useful for modeling unobserved stochastic trends that may drive the dynamics of co-integrated multivariate series.

9.3 The dynamic Factor model for common trends

The modern *dynamic Factor model* fits naturally within the state-space framework since it explicity defines a parametric model for the dynamics of the unobserved factors. Moreover, within the multivariate context, the existence of co-integrating relationships amongst dependent variables leads directly to the "common trends" representation.

The *dynamic Factor model* for common trends is given in (137). Pre-multiplying the stochastic trend component $\boldsymbol{\mu}_t$, is the $N \times K$ matrix $\boldsymbol{\Theta}$ of *factor loadings*, where $\boldsymbol{\mu}_t$ is now $K \times 1$ and represents the unobserved *common factors* where $0 \leq K \leq N$ and:

$$\boldsymbol{y}_t = \boldsymbol{\Theta}\boldsymbol{\mu}_t + \boldsymbol{\mu}_0 + \boldsymbol{\eta}_t, \quad \text{for} \quad t = 1, \ldots, T \tag{137a}$$

$$\text{and} \quad \boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \boldsymbol{\beta} + \boldsymbol{\epsilon}_t. \tag{137b}$$

Additionally, $\boldsymbol{\mu}_0$ is included as a constant on the last $N - K$ series. That is, the first N elements of $\boldsymbol{\mu}_0$ are zero, and the last $N - K$ are $\tilde{\mu}$.

As it stands the model in (137) is unidentifiable since for any singular $N \times N$ matrix $\boldsymbol{H}$, we

could redefine $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}\boldsymbol{H}^{-1}$ and $\boldsymbol{\mu}_t^* = \boldsymbol{H}\boldsymbol{\mu}_t$ so that $\boldsymbol{y}_t = \boldsymbol{\Theta}^*\boldsymbol{\mu}_t^* + \boldsymbol{\mu}_0 + \boldsymbol{\eta}_t$ and $\boldsymbol{\mu}_t^* = \boldsymbol{\mu}_{t-1}^* + \boldsymbol{\beta}^* + \boldsymbol{\epsilon}_t^*$ are observationally equivalent to (137), given $\boldsymbol{\epsilon}_t^* = \boldsymbol{H}\boldsymbol{\epsilon}_t$ and $\boldsymbol{\beta}_t^* = \boldsymbol{H}\boldsymbol{\beta}_t$.

Therefore, restrictions must be placed on the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and the factor loadings $\boldsymbol{\Theta}$. One way to solve the problem is to set $\boldsymbol{\Theta}_{ij} = 0$ for $j > i$ and $\boldsymbol{\Theta}_{ii} = 1$ where $i = 1, \ldots, K$, while $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is set as a diagonal matrix. Interestingly, if $K = N$, then the model is equivalent to (142) below, the simplest form of the so-called "structural" models popularized by Harvey (1984). Rewriting $\boldsymbol{\mu}_t^* = \boldsymbol{\Theta}\boldsymbol{\mu}_t$, we have that the variance of $\boldsymbol{\epsilon}_t$ is the positive semi-definite matrix $\boldsymbol{\Theta}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}\boldsymbol{\Theta}'$.

As it stands the identification restriction placed above on $\boldsymbol{\Theta}$ is a rather arbitrary one and does not suggest any reasonable economic intuition. However, this problem is not insurmountable since once the model in (137) is estimated under the suggested restriction, we can then subsequently introduce a *factor rotation* by including an orthogonal matrix $\boldsymbol{H}$, as above, in between $\boldsymbol{\Theta}$ and $\boldsymbol{\mu}_t$. The new common trends are then $\boldsymbol{\mu}_t^*$. This approach may yield useful interpretations since we would be able to associate certain subsets of the elements our vector of observed variables, $\boldsymbol{y}_t$, with certain subsets of the common factors. Moreover, it may be possible to assign the observed series into various *clusters*, allowing the factor loadings matrix $\boldsymbol{\Theta}$ to be block-diagonal.

9.3.1 Co-integration of levels process $\boldsymbol{y}_t$

The model in (137) also has the interesting feature that $N - K$ linear combinations of $\boldsymbol{y}_t$ are stationary, despite the fact that all the elements of $\boldsymbol{y}_t$ individually are nonstationary, $I(1)$, processes. Consider a matrix $\boldsymbol{C}$ which is $(N - K) \times N$ and which has the property that $\boldsymbol{C}\boldsymbol{\Theta} = 0$. We have from (137) that $\boldsymbol{C}\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{\mu}_0 + \boldsymbol{C}\boldsymbol{\eta}_t$, and thus the elements of $\boldsymbol{C}\boldsymbol{y}_t$ now represent $(N - K)$ stationary series, where $\boldsymbol{C}$ has as its rows the $(N - K)$ *co-integrating relations*.

It turns out that this result is one implication of the Engle-Granger representation theorem (1987) which states that for any multivariate series which are co-integrated, there exist three equivalent representations: the VARMA, the VECM (vector error correction model), and the "common trends" Factor model representation.

To illustrate the connection suppose that:

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{y}_{t-1} + \boldsymbol{\eta}_t \tag{138}$$

so that $\boldsymbol{y}_t$, $N \times 1$, follows a $VAR(1)$ process (and so $\boldsymbol{A}$ is not generally symmetric).

Now, suppose that $\boldsymbol{A}$ has a spectral decomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$, where $\boldsymbol{P}$ contains the eigenvectors of $\boldsymbol{A}$ as its columns at that $\boldsymbol{P}$ has full column rank (i.e. there exists N distinct linearly independent eigenvectors). Moreover, $\boldsymbol{D}$ is diagonal and contains the eigenvalues of $\boldsymbol{A}$.

The conditional forecast for $\boldsymbol{y}_{t+h}$, given some horizon $h \in \mathbb{N}^+$ is from (138):

$$\begin{aligned} E[\boldsymbol{y}_{T+h}|\boldsymbol{y}_T] &= \boldsymbol{A}^h\boldsymbol{y}_T \\ &= [\boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}]^h\boldsymbol{y}_T \\ &= [\boldsymbol{P}\boldsymbol{D}^h\boldsymbol{P}^{-1}]\boldsymbol{y}_T \end{aligned} \tag{139a}$$

Without any loss of generality, let the first K eigenvalues in $\boldsymbol{D}$ be equal to 1 and the other $N-K$ eigenvalues have modulus strictly less than 1. Furthermore, let $\boldsymbol{P}^{-1}\boldsymbol{y}_t = \boldsymbol{z}_t$. Equation (139a) implies that in the limit, as $h \to \infty$, the first K elements of $E[\boldsymbol{z}_{t+h}|\boldsymbol{z}_T]$ will approach ∞ while the last $N-K$ elements will approach 0. That is, the last $N-K$ elements of $\boldsymbol{z_t}$ are mean reverting [11] while for the first K elements no unconditional mean exists. Therefore, there exist linear combinations of $\boldsymbol{y}_t$ which are $I(0)$ (i.e. the last $N-K$ elements of $\boldsymbol{z}_t$) despite the fact that all the elements $\boldsymbol{y}_t$ are $I(1)$ (by virtue of the K unit eigenvalues of $\boldsymbol{A}$). We say that there exist $N-K$ co-integrating relations, or equivalently, that there exist K common stochastic trends driving $\boldsymbol{y}_t$.

Given that the last $N-K$ diagonal elements of $\boldsymbol{D}^h$ approach 0 in the limit, we can always

[11] That is, the conditional mean reverts back to the unconditional mean of 0 as $h \to \infty$. Or in other words, $E[\boldsymbol{z}_{t+h}|\boldsymbol{z}_T] \to E[\boldsymbol{z}_t] = 0$ as $h \to \infty$ for those last $N-K$ stable elements of $\boldsymbol{z}_t$.

rewrite (139a) as:

$$\begin{aligned}\lim_{h\to\infty} E[\boldsymbol{y}_{T+h}|\boldsymbol{y}_T] &= \boldsymbol{A}^{\infty}\boldsymbol{y}_T \\ &= \boldsymbol{P}\begin{bmatrix}\bar{\boldsymbol{z}}_T \\ \mathbf{0}\end{bmatrix} \\ &= \bar{\boldsymbol{P}}\bar{\boldsymbol{z}}_T \end{aligned} \tag{140a}$$

where $\bar{\boldsymbol{z}}_T$ is is the first K rows of $\boldsymbol{z}_T$, and $\bar{\boldsymbol{P}}$, $N \times K$, is the first K columns of $\boldsymbol{P}$.

Therefore, we have that the "common trends" representation of the co-integrated system $\boldsymbol{y}_t$ is given as:

$$\begin{aligned}\boldsymbol{y}_t &= \lim_{h\to\infty} E[\boldsymbol{y}_t|\boldsymbol{y}_{t-h}] + \boldsymbol{\eta}_t \\ &= \bar{\boldsymbol{P}}\bar{\boldsymbol{z}}_t + \boldsymbol{\eta}_t \end{aligned} \tag{141a}$$

so if we allow $\boldsymbol{C} \equiv \overline{\boldsymbol{P}}^{-1}$ to be the $(N-K)\times N$ matrix with the co-integrating relations (i.e. the last $N-K$ rows of $\boldsymbol{P}^{-1}$) as its rows, then we have the result that $\boldsymbol{C}\boldsymbol{y}_t = \boldsymbol{C}\boldsymbol{\eta}_t$ is $I(0)$, since $\overline{\boldsymbol{P}}^{-1}\bar{\boldsymbol{P}} = \mathbf{0}$. Note the similarity to (137) where $\bar{\boldsymbol{z}}_t \equiv \boldsymbol{\mu}_t$, $\boldsymbol{\Theta} \equiv \bar{\boldsymbol{P}}$, and of course, $\bar{\boldsymbol{z}}_t$ is $I(1)$.

9.4 Unobserved Components

The *dynamic Factor model* for common trends in (137) also represents the simplest form of the structural (or "Unobserved Components") framework popularized by Harvey (1984), wherein it is referred to as the *Local level random walk plus noise* model:[12]

$$\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\eta}_t, \quad \text{for} \quad t = 1,\ldots,T, \tag{142a}$$

$$\text{and} \quad \boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \boldsymbol{\epsilon}_t, \tag{142b}$$

[12] See Harvey (1989) pg. 429 and Nerlove et al. (1979) for a review of the unobserved components models more generally.

where $\boldsymbol{\mu}_t$ is an $N \times 1$ vector of local level components, $\boldsymbol{\eta}_t$ and $\boldsymbol{\epsilon}_t$ are $N \times 1$ vectors of multivariate white noise with zero mean and covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ respectively, and $\boldsymbol{y}_t$ represents an $N \times 1$ vector of series which may be contemporaneously related (e.g. cross-sectional data).

The common trends framework, with $I(1)$ trends, models the local level component, $\boldsymbol{\mu}_t$, as a random walk (i.e. stochastic trend), although it is free to take on any number of different structural representations [see (148a) below or example 2.2.10 for the general representation]. Not only do the $\boldsymbol{\eta}_t$'s exhibit contemporaneous correlation, but the innovations of the trend components themselves, $\boldsymbol{\epsilon}_t$, do as well. The N series, $\boldsymbol{y}_t$, are thus linked via the off-diagonals of the two innovation covariance matrices, each of which include $N(N+1)/2$ parameters to be estimated. It is worth mentioning that (142) is similar to the time-varying SUR model of Zellner (1962), with the exception that we replace the regressors $\boldsymbol{X}_t\boldsymbol{\beta}$ with the time-varying intercept vector $\boldsymbol{\mu}_t$.

9.4.1 Alternative representations

The model in (142) can also be rewritten in stationary single equation form as:

$$(\boldsymbol{I} - \boldsymbol{I}L)\boldsymbol{y}_t = \boldsymbol{\epsilon}_t + (\boldsymbol{I} - \boldsymbol{I}L)\boldsymbol{\eta}_t, \quad \text{where} \quad t = 1, \ldots, T, \tag{143}$$

so that $E[\triangle \boldsymbol{y}_t] = 0, Var[\triangle \boldsymbol{y}_t] = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + 2\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$, and $Covar[\triangle \boldsymbol{y}_t, \triangle \boldsymbol{y}_{t-1}] = -\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$.[13] Therefore, the reduced form stationary model of (143) is of the VARIMA(0,1,1) type since it matches its autocorrelation representation. That is, if $\boldsymbol{x}_t = \boldsymbol{u}_t + \boldsymbol{\Phi}\boldsymbol{u}_{t-1}$ we have that $\boldsymbol{\Gamma}_{\boldsymbol{x}}(0) = E[\boldsymbol{x}_t\boldsymbol{x}_{t-1}'] = \boldsymbol{\Sigma}_{\boldsymbol{u}} + \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{u}}\boldsymbol{\Phi}'$ and $\boldsymbol{\Gamma}_{\boldsymbol{x}}(1) = E[\boldsymbol{x}_t\boldsymbol{x}_{t-1}'] = \boldsymbol{\Phi}\boldsymbol{\Sigma}_{\boldsymbol{u}}$ which is of the same form if $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = 0$ and $\boldsymbol{\Phi} = -\boldsymbol{I}$.

The structural form of the model in (142) implies restrictions on the parameter space of the reduced form VARIMA(0,1,1) representation of (143).[14] Take for simplicity the example of the

[13] All other autocovariances are zero.

[14] Note, this is similar to translating between the state-space form and the transfer-function form of a system within the engineering context. While the state-space form implies a unique transfer function form, the transfer function form has any number of respective state-space represention. So in this context, the specified state-space form implies restrictions on the transfer function form of the model. By transfer function form, we mean the model written purely in terms of impulse reponse to exogenous shocks. In this sense, we can think of the multiplicity of state-space forms within the context of Akaike (1975) and the "minimal" representation.

univariate analog to (143), the ARIMA(0,1,1) model:

$$x_t = u_t + \phi u_{t-1}, \quad \text{where} \quad u_t \sim N(0, \sigma_u^2). \tag{144}$$

Matching autocorrelations of (144) with the univariate counterpart to (143) results in:

$$\rho(1) = \frac{-\sigma_\eta^2}{\sigma_\epsilon^2 + 2\sigma_\eta^2} = \frac{\phi\sigma_u^2}{(1+\phi^2)\sigma_u^2} \tag{145a}$$

$$\Leftrightarrow \quad 0 = \phi^2\sigma_\eta^2 + \phi(\sigma_\epsilon^2 + 2\sigma_\eta^2) + \sigma_\eta^2 \tag{145b}$$

which is a second-order polynominal in ϕ. One of the roots is dropped since it implies non-invertibility. Solving (145b) we find that:

$$\phi = -(q+2)/2 + \sqrt{q^2/4 + q}, \quad \text{where} \quad q = \frac{\sigma_\epsilon^2}{\sigma_\eta^2} \tag{146a}$$

$$\Rightarrow \quad \text{if} \quad 0 \leq q \leq \infty \quad \text{we have that} \quad -1 \leq \phi \leq 0. \tag{146b}$$

And thus the standard stationary region of the parameter space is cut in half. Note that the order of the polynominal is equal to one plus the order of the MA process. Therefore solving for the restrictions of an $MA(q)$ process would generally require solving a Yule-Walker type system of equations, which include, at most, a $q+1$ order polynomial. Given this, Nerlove et al. (1979, pp. 70-78) provides a more general algorithm applicable to complicated univariate structural forms. As for the multivariate case, the problem is even more complex. In the univariate case, the structural form implies restrictions on the parameter space of the reduced form, but in the multivariate case it also implies a reduction in the dimension of the model. See Harvey (1989, pp. 432) for more details.

Interestingly, the model in (142) can also be viewed as the multivariate *UCARIMA* type (Engle,1978). The UCARIMA representation decomposes the process $\boldsymbol{y}_t = T_t + S_t + C_t + I_t$ into individual ARIMA processes (where T_t is the trend component, S_t is some seasonal component, C_t a non-seasonal cyclical component, and I_t is some residual irregular component.) For example

in (142) we have only defined the trend component, $T_t = \boldsymbol{\mu}_t$ and the irregular $I_t = \boldsymbol{\eta}_t$, so that the second and third components are effectively zero. However, this need not be the case–we could choose to define other components for S_t and C_t. For example, we could have chosen stochastic harmonic processes.

Either way, the first trend component can manifest as an ARIMA process by representing (142) as an ARIMA(0,1,1) process, with its parameter, $\boldsymbol{\Phi}$, characterized in the discussion above in determining the reduced form of (143):

$$\boldsymbol{y}_t = \triangle^{-1}\boldsymbol{\epsilon}_t + \boldsymbol{\eta}_t \tag{147a}$$

$$\equiv \triangle^{-1}(\boldsymbol{I} + \boldsymbol{\Phi}L)\boldsymbol{u}_t, \tag{147b}$$

where $\triangle = (\boldsymbol{I} - \boldsymbol{I}L)$. Therefore, $\boldsymbol{y}_t$ is composed of a singular ARIMA(0,1,1) process (where we have subsumed the irregular component into the trend for clarity).

More generally, we can write the multivariate *UCARIMA* form under any particular set of components $\{T_t, S_t, C_t, \ldots, I_t\}$ as:

$$\boldsymbol{y}_t = \sum_{m=0}^{M} \triangle_m^{-1}(L)\boldsymbol{\Theta}_m^{-1}(L)\boldsymbol{\Phi}_m(L), \tag{148a}$$

$$\text{where} \quad \triangle(L) = \prod_{m=0}^{M} \triangle_m(L), \tag{148b}$$

so that pre-multiplying (148a) by (148b) makes the process stationary. For reference, consider the example in 2.2.10 where we presented the *Basic Structural Model*; it also has a UCARIMA representation – see Harvey (1989, pg.74).

10 References

H. Akaike (1974) "Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes," *Ann. Inst. Statist. Math.,* 26, 363-387

———— (1975) "Markovian representation of stochastic processes by canonical variables," *SIAM Journal on Control* 13, 162-173

G. Box and G. Jenkins (1970) *Time series analysis: Forecasting and control*, San Francisco: Holden-Day

J. Breitung, S. Eickmeier (2006) "Dynamic factor models," in *Modern Econometric Analysis: Surveys on Recent Developments*, eds. O. Hubler, J. Frohn, Springer

D.R. Brillinger (1981) *Time Series: Data Analysis and Theory: Expanded edition*, Holden-Day, San Francisco

N.F. Chen, R. Roll, S.A. Ross (1986) "Economic forces and the stock market," *The Journal of Business*, 59, 383-404

De Jong (1989) "Smoothing and interpolation with the state space model," *Journal of the American Statistical Association*, 84, 1085-1088

R. Engle (1978) "Estimating structural models of seasonality," in *Seasonal Analysis of Economic Time Series*, A. Zellner ed., NBER

———— (1982) "Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation," *Econometrica*, 50(4), 987-1007

———— (1987) "Multivariate ARCH with factor structures – cointegration in variance," University of California Press, San Diego

R.F. Engle, C.W.J. Granger (1987) "Co-integration and error correction: representation, estimation, and testing," *Econometrica*, 55, 251-276

E.F. Fama, K.R. French (1992) "The cross-section of expected stock returns," *Journal of Finance*, 47(2), 427-465

J.D. Hamilton (1994) *Time Series Analysis*, Princeton University Press, New Jersey

E.J. Hannan, R.D. Terrell, N. Tuckwell (1970) "The seasonal adjustment of economic time series," *International Economic Review*, 11, 24-52

P.J. Harrison, C.F. Stevens (1976) "Bayesian forecasting," *Journal of the Royal Statistical Society Series B*, 38(3), 205-247

A.C. Harvey (1984) "A unified view of statistical forecasting procedures," *Journal of Forecasting*, 3(3), 245-275

———— (1989) *Forecasting, structural time series models, and the Kalman filter*, Cambridge University Press

A.E. Kalman (1960) "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, 82(1), 35-45

Nerlove, Grether, and Carvalho (1979) *Analysis of economic time series: A synthesis*, New York: Academic Press

M.B. Priestley (1981) *Spectral Analysis and Time Series, 2 volumes*, London: Elsevier Academic Press

S. Ross (1976) "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, 13(3), 341-360

W.F. Sharpe (1966) "Mutual fund performance," *The Journal of Business*, 39(1), 119-138

———— (1970) *Portfolio theory and capital markets*, McGraw-Hill, New York

D. Simon (2006) *Optimal State Estimation*, New Jersey: Wiley

P.C. Young (2011) *Recursive Estimation and Time-Series Analysis,* 2nd ed., Berlin: Springer-Verlag

Zellner (1962) “An efficient method of estimating seemingly unrelated regression equations and tests of aggregation bias,” *Journal of the American Statistical Association*, 57, 500-509