Contents

4 State space models

We consider here a class of models that represent the data as measurements derived from a latent stochastic process, itself assumed to be a Markov chain. We first describe Markov chains, move on to general state space models, before focusing on the case of linear Gaussian models. Some people use "state space models" and "hidden Markov models" interchangeably.

4.1 Markov chains

We start with a brief introduction to the idea of "Markov chains". There are many reasons to care about Markov chains, and entire textbooks are devoted to their study, e.g. *Markov chains*, by James Norris for chains on discrete spaces, and the recent book *Markov chains*, by Randal Douc, Eric Moulines, Pierre Priouret & Philippe Soulier, on chains on continuous spaces. Here we introduce the minimal ingredients required to the understanding of state space models and Kalman filters, described below.

We consider a stochastic process $(X_t)_{t\geq 0}$ where the time index t is discrete: $t = 0, 1, 2, \ldots$. We say that the process is "Markov", or "Markovian", or that it is a Markov chain, or that it satisfies the Markov property, if the distribution of X_t given $X_0, \ldots, X_{t-1}$ only depends on X_{t-1}. In other words, given X_{t-1}, the present state X_t is independent from the the past $X_0, \ldots, X_{t-2}$.

For Markov chains in discrete state spaces, we can decompose the probability of any trajectory of the chain as

$$\mathbb{P}(X_0 = x_0, \ldots, X_t = x_t) = \mathbb{P}(X_0 = x_0) \prod_{s=1}^{t} \mathbb{P}(X_s = x_s | X_{s-1} = x_{s-1}). \tag{4.1}$$

We see that the distribution of trajectories of (X_t) is entirely determined by the initial distribution of X_0, denoted by π_0, i.e. $\pi_0(x_0) = \mathbb{P}(X_0 = x_0)$ for all x_0, and the transition probabilities $\mathbb{P}(X_s = j|X_{s-1} = i)$ for all s, i, j, which we will denote by P_{ij} later on.

Remark 4.1. In principle, the transition probabilities $\mathbb{P}(X_s = j|X_{s-1} = i)$ could change over time. We would then call the Markov chain "inhomogeneous". Hereafter we focus on the homogeneous case, so that $\mathbb{P}(X_s = j|X_{s-1} = i)$ is the same for all s, i.e. is the same as $\mathbb{P}(X_1 = j|X_0 = i)$.

We could consider Markov chains "of order m", with the property that X_t given $X_{t-1}, \ldots, X_{t-m}$ is independent of $X_0, \ldots, X_{t-m-1}$. The first definition then corresponds to $m = 1$. Note that if (X_t) is a Markov chain of order $m \geq 2$, we can define a process $(Y_t)_{t\geq 0}$ as $Y_t = (X_{t-m+1}, \ldots, X_{t-1}, X_t)$ for $t \geq m$ such that $(Y_t)_{t\geq m}$ is a (vector-valued) Markov chain of order 1. So we are not losing much generality by considering only Markov chains of order 1.

If the Markov chain takes values in a continuous state space, e.g. $\mathbb{R}$, then the Markov property can be defined similarly. We write π_0 for the initial distribution of X_0 and we write $P(x_t|x_{t-1})$ for the density of the transition probability from $X_{t-1} = x_{t-1}$, evaluated at x_t. With this notation, the joint density of a trajectory of the chain can be written

$$p(x_0, \ldots, x_t) = \pi_0(x_0) \prod_{s=1}^{t} P(x_s|x_{s-1}). \tag{4.2}$$

The conditional distribution denoted by P is referred to as the "Markov transition kernel", and it describes the distribution of X_t given that X_{t-1} is equal to x_{t-1}. Alternative notation is common, for example $P(x_{t-1}, x_t)$ or $P(x_{t-1} \to x_t)$.

Example 4.1. (AR(p) processes as Markov chains). Consider first an AR(1) model, with $Y_t = \varphi_1 Y_{t-1} + W_t$ where W_t are i.i.d. $\mathcal{N}(0, \sigma_W^2)$. It is a Markov chain, with transition kernel $P(y_t|y_{t-1}) = \mathcal{N}(y_t; \varphi_1 y_{t-1}, \sigma_W^2)$, where $\mathcal{N}(x; \mu, \sigma_W^2)$ refers to the probability density function of $\mathcal{N}(\mu, \sigma_W^2)$ evaluated at x. If we consider an AR(p) process $Y_t = \sum_{j=1}^{p} \varphi_j Y_{t-j} + W_t$, we can see it either as a Markov chain of order p, or we can introduce the vector $Z_t = (Y_{t-p}, \ldots, Y_t)^T$, and note that

$$Z_t = \begin{pmatrix} Y_{t-p+1} \\ Y_{t-p+2} \\ \vdots \\ Y_{t-1} \\ Y_t \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \varphi_p & \varphi_{p-1} & \varphi_{p-2} & \ldots & \varphi_1 \end{pmatrix} \begin{pmatrix} Y_{t-p} \\ Y_{t-p+1} \\ \vdots \\ Y_{t-2} \\ Y_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t = \Phi Z_{t-1} + \tilde{W}_t,$$

upon defining the matrix Φ and vector $\tilde{W}_t$ appropriately. In the latter form we can convince

ourselves that (Z_t) is a Markov chain of order 1, with transition probabilities given by

$$P(z_t|z_{t-1}) = \left\{ \mathcal{N}(z_t(p); \sum_{j=1}^{p} \varphi_j z_{t-1}(p+1-j), \sigma_W^2) \right\} \prod_{j=1}^{p-1} \mathbb{1}(z_t(j) = z_{t-1}(j+1)),$$

where $z_t(j)$ refers to the j-th element of the vector $z_t \in \mathbb{R}^p$.

Example 4.2. (Nonlinear autoregressive process). Consider a process of the form $X_t = f(X_{t-1}) + W_t$, where $f : \mathbb{X} \to \mathbb{X}$ is an arbitrary (well, at least measurable) function of the past term X_{t-1}, and the noise terms (W_t) are i.i.d. $\mathcal{N}(0, \sigma_W^2)$ variables. Combined with some initial distribution, e.g. $X_0 \sim \pi_0$, this defines a Markov chain. The transition density reads $P(x_t|x_{t-1}) = \mathcal{N}(x_t; f(x_{t-1}), \sigma_W^2)$. We can then apply (4.2) to write the joint density of the chain. This is a nonlinear extension of the AR(1) model.

4.2 State space models on continuous spaces

The idea of a state space model is to view the observations $(Y_t)_{t\geq 1}$ as noisy measurements of an underlying process $(X_t)_{t\geq 0}$. That process $(X_t)_{t\geq 0}$ is assumed to be a Markov chain: the distribution of X_t given $X_0, \ldots, X_{t-1}$ depends only on X_{t-1}. The *initial distribution* of X_0 is denoted by μ_θ: $X_0 \sim \mu_\theta$. The distribution of X_t given X_{t-1} is denoted by f_θ: $X_t \sim f_\theta(\cdot|X_{t-1})$; it is called the *transition distribution*, or *state equation*. Finally, the distribution of Y_t given X_t is denoted by g_θ: $Y_t|X_t \sim g_\theta(\cdot|X_t)$; it is called the *emission* or *measurement distribution*, or *observation equation*. The parameter θ parametrizes these (conditional) distributions; it can be considered known or unknown. Note that θ refers to a generic parameter here, not exclusively to parameters of MA processes as in the previous chapter. It is common to summarize the model by writing

$$X_0 \sim \mu_\theta, \quad \forall t \geq 1 \quad X_t|X_{t-1} \sim f_\theta(\cdot|X_{t-1}), \quad \text{and} \quad Y_t|X_t \sim g_\theta(\cdot|X_t). \tag{4.3}$$

The conditional dependencies of the random variables in a state space model are shown in a graph in Figure 4.1.

Example 4.3. (Target tracking). Consider the problem of tracking an object with position given by (x_t, z_t) in the plane at time t. Denote the horizontal and vertical velocities by $\dot{x}_t$ and $\dot{z}_t$. A

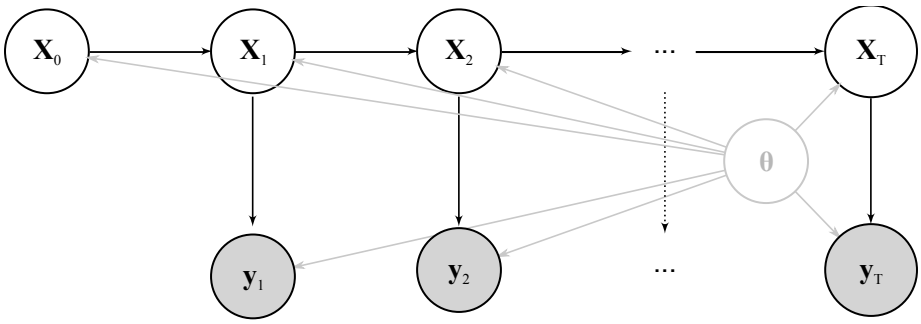

Figure 4.1: Variables involved in a state space model. The graph helps visualizing dependencies: for instance Y_t is connected only to X_t (and to the parameter θ), so that Y_t given X_t (and θ) is independent of everything else.

simple model for target tracking assumes that the process $\alpha_t = (x_t, \dot{x}_t, z_t, \dot{z}_t)$ follows the equation

$$\alpha_t = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \alpha_{t-1} + W_t,$$

where W_t follows a multivariate Normal distribution. The initial position of the object might be specified by some probability distribution μ_θ. We might not directly observe the position (x_t, z_t) and the velocities of the object, but we might observe an angle with respect to a reference line, as with radar measurements. The observations Y_t are e.g. assumed to be equal to $\tan^{-1}(z_t/x_t)$, plus some noise ϵ_t; see left of Figure 4.2. The parameters might include the covariance matrix of W_t and the variance of ϵ_t. This is the type of problem ("target tracking") that was studied in the field of electrical engineering in the 1950s, leading to the development of the Kalman filter.

Example 4.4. (Stochastic volatility). Consider the problem of estimating the volatility of a financial asset. Denote the daily price of an asset by p_t, at time t. Define the observations to be the log-returns: $Y_t = \log(p_t/p_{t-1})$. A stochastic volatility model assumes that the log-returns are normally distributed with mean 0 and variance $\exp(X_t)$, where X_t follows some stochastic process, for example an AR process, $X_t = \varphi X_{t-1} + W_t$. The parameters include φ and the variance of W_t. See right of Figure 4.2, and e.g. the book of Tsay, *Analysis of financial time series*, 2005.

Example 4.5. (Disease outbreaks). Consider a model of disease outbreak, where people in a population are "susceptible" (S) to be infected, "infected" (I), or "recovered" (R). Suppose that there are P persons in the population (at all times), and denote by $s(t)$, $i(t)$ and $r(t)$ the proportions of susceptible, infected and recovered people, respectively, at time t. The basic SIR model, introduced

by Kermack and McKendrick in 1927, describes the continuous evolution of the proportions as follows:

$$\begin{aligned} \frac{ds}{dt} &= -\beta s(t) i(t), \\ \frac{di}{dt} &= \beta s(t) i(t) - \gamma i(t), \\ \frac{dr}{dt} &= \gamma i(t). \end{aligned} \tag{4.4}$$

The first equation says that the proportion of susceptible people decreases proportionally to $\beta s(t) i(t)$, where β parametrizes the contact rate between susceptible and infected people. The second equation says that the proportion of infected people increases with $\beta s(t) i(t)$, but decreases with $\gamma i(t)$; γ parametrizes the rate of transfer from I to R, i.e. the rate of recovery. The last equation ensures that $s(t)$, $i(t)$ and $r(t)$ sum to one, provided that $s(0) + i(0) + r(0) = 1$ at time 0. From this deterministic (non-stochastic) continuous-time model, one can introduce randomness and discretize time, to obtain, for instance, the following model,

$$\begin{aligned} S_t &= S_{t-\Delta t} - \Delta S_t, \text{ with } \Delta S_t \sim \text{Binomial}(S_{t-1}, 1 - \exp(-\beta \frac{I_{t-1}}{P} \Delta t)), \\ I_t &= I_{t-\Delta t} + \Delta S_t - \Delta R_t, \text{ with } \Delta R_t \sim \text{Binomial}(I_{t-1}, 1 - \exp(-\gamma \Delta t)), \\ R_t &= R_{t-\Delta t} + \Delta R_t, \end{aligned} \tag{4.5}$$

where S_t, I_t, R_t denote numbers of people in categories S,I,R at time t, and thus sum to the population size P, and Δt is the time step. The observations Y_t might be related to S_t, I_t, R_t in different ways. For instance, we might observe $Y_t \sim \text{Binomial}(\Delta S_t, \rho)$, i.e. we observe each new case with probability ρ. From there, we can express the model as a state space model, where the latent process is (S_t, I_t, R_t).

Natural questions arise: we want to estimate the parameters using data, to predict future observations, to compare models, etc. Furthermore, some tasks are specific to state space models, and concern the estimation of the latent process (X_t) given the observations (Y_t).

- The task of *filtering* refers to the problem of calculating the distribution of X_t given $Y_1, \ldots, Y_t$, for all $t \geq 1$. In the target tracking example, this corresponds to the task of tracking the object "on-line", given the currently available data at time t.
- The task of *smoothing* refers to the problem of calculating the distribution of X_t given $Y_1, \ldots, Y_n$, for all $t \in \{1, \ldots, n\}$. In the stochastic volatility example, this corresponds to the task of estimating the volatility X_t at all times t, given the data available both before and after t.

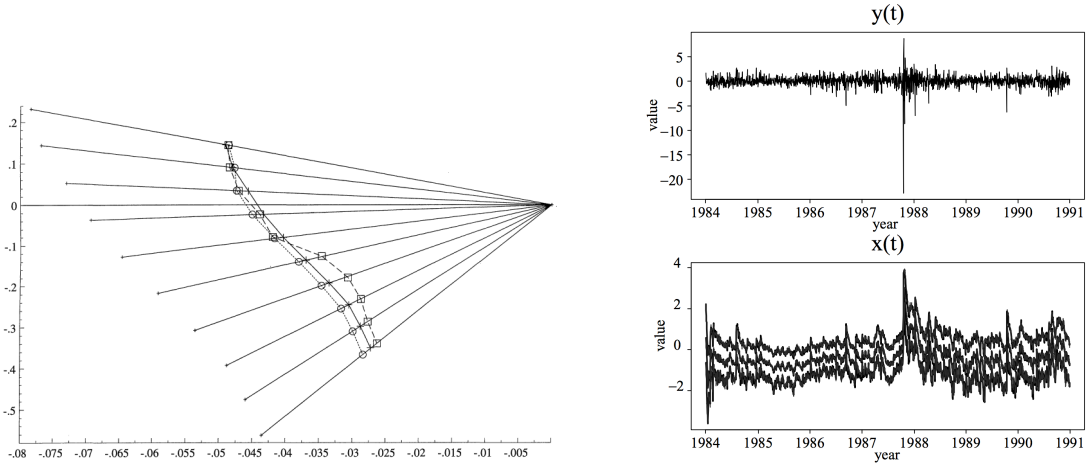

Figure 4.2: Left: angles measured to track an object in the plane. Right: log-returns of a financial asset (top) and estimated stochastic volatility (bottom) over time.

Filtering and smoothing are challenging problems. They can be solved exactly in some specific contexts. One such context is the case of linear Gaussian state space models (or linear Gaussian models, for brevity; they are also sometimes called Dynamic Linear Models). In that case, we will see that the Kalman filter and smoother give the distribution of X_t given $Y_1, \dots, Y_t$ and of X_t given $Y_1, \dots, Y_n$ respectively.

4.3 Linear Gaussian state space models

In linear Gaussian models, the state equation and the observation equation are both assumed to be linear and Gaussian, i.e.

$$\begin{aligned} X_t &= \Phi X_{t-1} + W_t, \\ Y_t &= A X_t + V_t, \end{aligned} \tag{4.6}$$

where W_t, V_t are independent multivariate Normal variables with mean zero and covariance matrices Σ_W and Σ_V respectively. The coefficients Φ and A are matrices. To be more precise, if X_t is of dimension d, Φ has to be $d \times d$, and W_t has to be d-dimensional. If Y_t is of dimension p, then A has to be of dimension $p \times d$ and V_t has to be of dimension p. The initial distribution of X_0 is given by a Normal distribution $\mathcal{N}(m_0, C_0)$, with mean m_0 and covariance matrix C_0. The parameters of the model are: $m_0, C_0, A, \Phi, \Sigma_W, \Sigma_V$.

The examples of target tracking and stochastic volatility models, provided above, are not linear

Gaussian models. Still, many interesting models belong to the class of linear Gaussian models. We have already seen that ARMA models can be seen as linear Gaussian models, and we recall this point below. We will also discuss structural time series models.

Linear Gaussian models can be generalized a little bit from (4.6): A and Φ can vary at every time t (they would then be denoted by A_t and Φ_t). Furthermore, we can add constant terms in the equations, for instance:

$$\begin{aligned} X_t &= \Phi X_{t-1} + \Gamma U_t + W_t, \\ Y_t &= A X_t + \Lambda U_t + V_t, \end{aligned}$$

where U_t is some observed time series or covariates, and Γ, Λ are matrices of appropriate dimensions. We can also extend the model by allowing W_t and V_t to be jointly correlated instead of independent. With these extensions, linear Gaussian models cover many more important models, such as linear regression with ARMA noise.

ARMA models. We recall that an ARMA(p,q) model is defined as $Y_t = \varphi_1 Y_{t-1} + \ldots + \varphi_p Y_{t-p} + W_t + \theta_1 W_{t-1} + \ldots + \theta_q W_{t-q}$. If we define $r = \max(p, q+1)$, and pad the vectors φ and θ with zeros to make their length equal to r, then we can define X_t with r elements $(X_{t,1}, \ldots, X_{t,r})$ as

$$\begin{pmatrix} X_{t,1} \\ X_{t,2} \\ \vdots \\ X_{t,r-1} \\ X_{t,r} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ \varphi_r & \varphi_{r-1} & \varphi_{r-2} & \ldots & \varphi_1 \end{pmatrix} \begin{pmatrix} X_{t-1,1} \\ X_{t-1,2} \\ \vdots \\ X_{t-1,r-1} \\ X_{t-1,r} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} W_t.$$

We can define the observation equation

$$Y_t = \begin{pmatrix} \theta_{r-1} & \theta_{r-2} & \ldots & \theta_1 & 1 \end{pmatrix} X_t + V_t,$$

where V_t is zero for all times t. Then Y_t can be seen to follow an ARMA(p,q) model.

Structural time series models. Structural models refers to models that represent the measurements using trends and seasonality components, just like classical decompositions. The simplest of

such models is called the "local level" or "random walk with noise" model,

$$X_t = X_{t-1} + W_t \tag{4.7}$$

$$Y_t = X_t + V_t. \tag{4.8}$$

We can check that $\nabla Y_t = Y_t - Y_{t-1}$ is equivalent to an MA(1) process, so that (Y_t) is an ARIMA(0,1,1) model. Recall from the previous lecture notes (Section 3.7) that this is the model underpinning simple exponential smoothing.

We can add a stochastic trend by re-defining $X_t = X_{t-1} + B_{t-1} + W_t$ where $B_t = B_{t-1} + W_t^b$, where W_t^b is another uncorrelated noise process with variance σ_b^2. This new "local linear trend" model can be written in state space form as

$$\bar{X}_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \bar{X}_{t-1} + \bar{W}_t \tag{4.9}$$

$$Y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \bar{X}_t + V_t, \tag{4.10}$$

where $\bar{X}_t = (X_t, B_t)$, $\bar{W}_t = (W_t, W_t^b)$.

Finally we can add a seasonal component S_t, with period s. The stochastic seasonal component can be defined as:

$$S_t = -\left(\sum_{j=1}^{s-1} S_{t-j}\right) + W_t^s,$$

where (W_t^s) is another white noise process, with variance σ_s^2. This way, the sum of s successive terms is W_t^s, a variable with expectation zero, and the variance $\sigma_s^2 > 0$ allows the seasonal component to vary over time. The state space form of the model with linear trend and seasonality involves the observation equation

$$Y_t = \begin{pmatrix} 1 & 0 & 1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} X_t \\ B_t \\ S_t \\ S_{t-1} \\ \vdots \\ S_{t-s+1} \end{pmatrix} + V_t.$$

The state equation describes the evolution of the $(s+2)$-dimensional latent process,

$$\begin{pmatrix} X_t \\ B_t \\ S_t \\ S_{t-1} \\ \vdots \\ S_{t-s+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & -1 & \dots & -1 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ B_{t-1} \\ S_{t-1} \\ S_{t-2} \\ \vdots \\ S_{t-s} \end{pmatrix} + \begin{pmatrix} W_t \\ W_t^b \\ W_t^s \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

These models are closely related to the exponential smoothing framework, while being fully probabilistic, thus allowing prediction intervals and not only deterministic forecasts.

References on this are the book of Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter* 1989, and more recently *Forecasting with Exponential Smoothing: The State Space Approach* by Rob J. Hyndman, Anne B. Koehler, J. Keith Ord and Ralph D. Snyder, 2008.

4.4 Kalman filter

We describe the Kalman filter, which is an algorithm to perform filtering in linear Gaussian models. We consider the model of (4.6). By filtering, we mean the calculation of the mean and the covariance matrix of X_t given the observations $Y_1, \dots, Y_t$ and *fixed parameter values*. We denote these means by $m_{t|t}$ and these covariance matrices by $C_{t|t}$. More generally, we denote by $m_{t|s}$ and $C_{t|s}$ the mean and covariance matrix of X_t given $Y_1, \dots, Y_s$, for all s, t. We will see that, as a fortunate by-product, the Kalman filter can be used to evaluate the likelihood function, and thus enables statistical inference by maximum likelihood in linear Gaussian models.

Remark 4.2. Some reminders about multivariate Normal distributions will be handy. If X is a vector of random variables with elements denoted in two blocks X_1 and X_2, then we can write that X follows a multivariate Normal distribution as follows

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_2 \end{pmatrix} \right).$$

Both blocks X_1 and X_2 can be univariate or multivariate. In the covariance matrix, $\Sigma_1 = \mathbb{V}[X_1]$, $\Sigma_2 = \mathbb{V}[X_2]$ and $\Sigma_{12} = \mathbb{C}\text{ov}(X_1, X_2) = \Sigma_{21}'$; Σ_1, Σ_2 are always square matrices and Σ_{12},Σ_{21} can be rectangular matrices. The "prime" denotes the transpose of a matrix. Then the conditional distribution of X_1 given $X_2 = x_2$ is a multivariate Normal distribution given by

$$X_1 | X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_2^{-1}(x_2 - \mu_2), \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{21}\right). \tag{4.11}$$

If X is multivariate Normal with mean μ and covariance matrix Σ, then for any rectangular matrix A of size $p \times n$ and vector β of size p,

$$AX + \beta \sim \mathcal{N}\left(A\mu + \beta, A\Sigma A'\right). \tag{4.12}$$

The justification of the Kalman filter relies on repeated applications of (4.11) and (4.12), we will now see.

By definition, $X_0 \sim \mathcal{N}(m_0, C_0)$. Since $X_1 = \Phi X_0 + W_1$, we obtain $X_1 \sim \mathcal{N}(\Phi m_0, \Phi C_0 \Phi' + \Sigma_W)$ using (4.12). We thus define $m_{1|0} = \Phi m_0$ and $C_{1|0} = \Phi C_0 \Phi' + \Sigma_W$. Next assume that we have managed to find $m_{t|t-1}, C_{t|t-1}$ at some time $t \geq 1$. We describe how to obtain $m_{t|t}$, $C_{t|t}$ and $m_{t+1|t}$, $C_{t+1|t}$.

We start by writing the joint distribution of X_t and Y_t given $Y_{1:t-1} = y_{1:t-1}$ (or "given nothing" if $t = 1$). Marginally, X_t given $Y_{1:t-1} = y_{1:t-1}$ follows $\mathcal{N}\left(m_{t|t-1}, C_{t|t-1}\right)$. Since $Y_t = AX_t + V_t$, we obtain marginally $Y_t \sim \mathcal{N}\left(Am_{t|t-1}, AC_{t|t-1}A' + \Sigma_V\right)$ using (4.12). The covariance between X_t and Y_t, given $Y_{1:t-1} = y_{1:t-1}$, is given by $\mathbb{C}\text{ov}\left(X_t, Y_t\right) = \mathbb{C}\text{ov}\left(X_t, AX_t + V_t\right) = \mathbb{V}[X_t]A' = C_{t|t-1}A'$, since V_t is independent of X_t (note: everything is conditional upon $Y_{1:t-1} = y_{1:t-1}$ at this point, so by $\mathbb{C}\text{ov}(X_t, Y_t)$ we mean $\mathbb{C}\text{ov}(X_t, Y_t | Y_{1:t-1} = y_{1:t-1})$). Thus, we can write the joint distribution of (X_t, Y_t) given $Y_{1:t-1} = y_{1:t-1}$ as

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} | Y_{1:t-1} = y_{1:t-1} \quad \sim \quad \mathcal{N}\left(\begin{pmatrix} m_{t|t-1} \\ Am_{t|t-1} \end{pmatrix}, \begin{pmatrix} C_{t|t-1} & C_{t|t-1}A' \\ AC_{t|t-1} & AC_{t|t-1}A' + \Sigma_V \end{pmatrix}\right).$$

We apply (4.11) to obtain the distribution of X_t given $Y_{1:t} = y_{1:t}$:

$$\begin{aligned} m_{t|t} &= m_{t|t-1} + C_{t|t-1}A'(AC_{t|t-1}A' + \Sigma_V)^{-1}\left(y_t - Am_{t|t-1}\right), \\ C_{t|t} &= C_{t|t-1} - C_{t|t-1}A'(AC_{t|t-1}A' + \Sigma_V)^{-1}AC_{t|t-1}. \end{aligned}$$

This gives the filtering mean and covariance matrix at time t. It is called the *update step*, since we update the predictive distribution $\mathcal{N}\left(m_{t|t-1}, C_{t|t-1}\right)$ in the light of a new observation y_t. To get the next predictive distribution, we apply (4.12) again: since $X_{t+1} = \Phi X_t + W_{t+1}$, we obtain $m_{t+1|t} = \Phi m_{t|t}$ and $C_{t+1|t} = \Phi C_{t|t}\Phi' + \Sigma_W$. This is called the *prediction step*. The Kalman filter alternates between prediction and update steps, assimilating observations one at a time.

The Kalman filter provides the filtering and predictive means and covariance matrices at each time t. It can also be used to calculate the likelihood. At time t we have computed the marginal distribution of Y_t given $Y_{1:t-1} = y_{1:t-1}$, namely $Y_t \sim \mathcal{N}\left(Am_{t|t-1}, AC_{t|t-1}A' + \Sigma_V\right)$. Thus the conditional likelihood $p(y_t | y_{1:t-1}, \text{parameters})$ is the density of that Normal distribution, evaluated

at the observed y_t. In the case of univariate observations, this gives

$$p(y_t|y_{1:t-1}, \text{parameters}) = \frac{1}{\sqrt{2\pi(AC_{t|t-1}A' + \Sigma_V)}} \exp\left(-\frac{1}{2(AC_{t|t-1}A' + \Sigma_V)}\left(y_t - Am_{t|t-1}\right)^2\right).$$

The full likelihood $p(y_1, \ldots, y_n|\text{parameters})$ is obtained via the product of these conditional likelihoods,

$$p(y_1, \ldots, y_n|\text{parameters}) = p(y_1|\text{parameters}) \prod_{t=2}^{n} p(y_t|y_{1:t-1}, \text{parameters}).$$

The Kalman filter requires one forward pass over the data, therefore its computational cost is of the order of n operations, where n is number of observations. We now finally know how to obtain the likelihood in ARMA(p,q) models for a cost of $\mathcal{O}(n)$ operations! This is useful to perform maximum likelihood estimation, in combination with a numerical optimizer. Note that the operations involved in the Kalman filter are differentiable with respect to the inputs $m_0, C_0, A, \Phi, \Sigma_W, \Sigma_V$, so that we can compute gradients with respect to these inputs for numerical optimization purposes.

4.5 Kalman smoother

The Kalman smoother is used to obtain $m_{t|n}, C_{t|n}$ for all $t = 1, \ldots, n$. It works by first running the Kalman filter and storing the filtering and predictive means and covariance matrices. At that point, we have $m_{n|n}$ and $C_{n|n}$, which defines the mean and variance of the terminal smoothing distribution. The Kalman smoother then works backward in time, for an index t going from n down to 1. We describe how to compute $m_{t-1|n}$ and $C_{t-1|n}$, provided that we have managed to compute $m_{t|n}$ and $C_{t|n}$.

We first note that the distribution of X_{t-1} given $X_t = x_t$ and $Y_{1:n} = y_{1:n}$ does not depend on $y_t, \ldots, y_n$. This can be seen directly from Figure 4.1: if we remove node X_t, no edge connects X_{t-1} to $Y_t, \ldots, Y_n$. In terms of densities, this means

$$\begin{aligned} p(x_{t-1}|x_t, y_{1:n}) &= p(x_{t-1}|x_t, y_{1:t-1}) \\ &= \frac{p(x_t|x_{t-1}, y_{1:t-1})p(x_{t-1}|y_{1:t-1})}{p(x_t|y_{1:t-1})}, \end{aligned} \tag{4.13}$$

the second line coming from Bayes' formula. Furthermore, $p(x_t|x_{t-1}, y_{1:t-1}) = p(x_t|x_{t-1})$, since X_t given X_{t-1} is independent of the past observations. From this reasoning, the Kalman smoothing procedure consists in calculating the distribution of X_{t-1} given $X_t = x_t$ and $Y_{1:n} = y_{1:n}$, which is the same as the distribution of X_{t-1} given $X_t = x_t$ and $Y_{1:t-1} = y_{1:t-1}$. Then we obtain the distribution of X_{t-1} given $Y_{1:n} = y_{1:n}$ by using the law of total expectation, also called the tower property.

Let's delve into the details. Given the past observations $y_1, \dots, y_{t-1}$, we have

$$\begin{pmatrix} X_{t-1} \\ X_t \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} m_{t-1|t-1} \\ m_{t|t-1} \end{pmatrix}, \begin{pmatrix} C_{t-1|t-1} & C_{t-1|t-1}\Phi' \\ \Phi C_{t-1|t-1} & C_{t|t-1} \end{pmatrix} \right),$$

thus, using (4.11), $X_{t-1}|X_t = x_t, Y_{1:t-1} = y_{1:t-1}$ follows a Normal distribution with

$$\text{mean} = m_{t-1|t-1} + C_{t-1|t-1}\Phi' C_{t|t-1}^{-1}(x_t - m_{t|t-1}), \tag{4.14}$$

$$\text{variance} = C_{t-1|t-1} - C_{t-1|t-1}\Phi' C_{t|t-1}^{-1}\Phi C_{t-1|t-1}. \tag{4.15}$$

To simplify notation, we introduce $L_{t-1} = C_{t-1|t-1}\Phi' C_{t|t-1}^{-1}$. Then the above mean is written $m_{t-1|t-1} + L_{t-1}(x_t - m_{t|t-1})$, and the above variance is $C_{t-1|t-1} - L_{t-1}C_{t|t-1}L'_{t-1}$.

We now use the law of total expectation to recover the distribution of X_{t-1} given $Y_{1:n} = y_{1:n}$. Recall that $X_t \sim \mathcal{N}(m_{t|n}, C_{t|n})$, given $Y_{1:n} = y_{1:n}$. Therefore, the expectation of X_{t-1} given $Y_{1:n} = y_{1:n}$ is given by

$$\begin{aligned}
\mathbb{E}[X_{t-1}|y_{1:n}] &= \mathbb{E}[\mathbb{E}[X_{t-1}|X_t, y_{1:n}]|y_{1:n}] \text{ by tower property} \\
&= \mathbb{E}[\mathbb{E}[X_{t-1}|X_t, y_{1:t-1}]|y_{1:n}] \text{ by Eq. (4.13)} \\
&= \mathbb{E}[m_{t-1|t-1} + L_{t-1}(X_t - m_{t|t-1})|y_{1:n}] \text{ by Eq. (4.14)} \\
&= m_{t-1|t-1} + L_{t-1}(m_{t|n} - m_{t|t-1}).
\end{aligned}$$

For the variance, we use the formula $\mathbb{V}[A] = \mathbb{E}[\mathbb{V}[A|B]] + \mathbb{V}[\mathbb{E}[A|B]]$,

$$\begin{aligned}
\mathbb{V}[X_{t-1}|y_{1:n}] = \mathbb{E}[\mathbb{V}[X_{t-1}|X_t, y_{1:t-1}]|y_{1:n}] \\
+ \mathbb{V}[\mathbb{E}[X_{t-1}|X_t, y_{1:t-1}]|y_{1:n}],
\end{aligned}$$

and compute the two terms separately. On the one hand

$$\begin{aligned}
&\mathbb{E}[\mathbb{V}[X_{t-1}|X_t, y_{1:t-1}]|y_{1:n}] = \mathbb{E}[C_{t-1|t-1} - L_{t-1}C_{t|t-1}L'_{t-1}|y_{1:n}] \text{ by Eq. (4.15)} \\
&= C_{t-1|t-1} - L_{t-1}C_{t|t-1}L'_{t-1}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{V}[\mathbb{E}[X_{t-1}|X_t, y_{1:t-1}]|y_{1:n}] = \mathbb{V}[m_{t-1|t-1} + L_{t-1}(X_t - m_{t|t-1})|y_{1:n}] \\
&= L_{t-1}\mathbb{V}[X_t|y_{1:n}]L'_{t-1} = L_{t-1}C_{t|n}L'_{t-1}.
\end{aligned}$$

So, in the end, we obtain the mean and the covariance of X_{t-1} given $Y_{1:t} = y_{1:t}$ as

$$m_{t-1|n} = m_{t-1|t-1} + L_{t-1}(m_{t|n} - m_{t|t-1}),$$
$$C_{t-1|n} = C_{t-1|t-1} + L_{t-1}\left(C_{t|n} - C_{t|t-1}\right) L'_{t-1}.$$

This completes one step backward in the Kalman smoother. When we reach time $t = 0$, we have computed all $m_{t|n}$ and $C_{t|n}$ for all $t \in \{0, \ldots, n\}$.

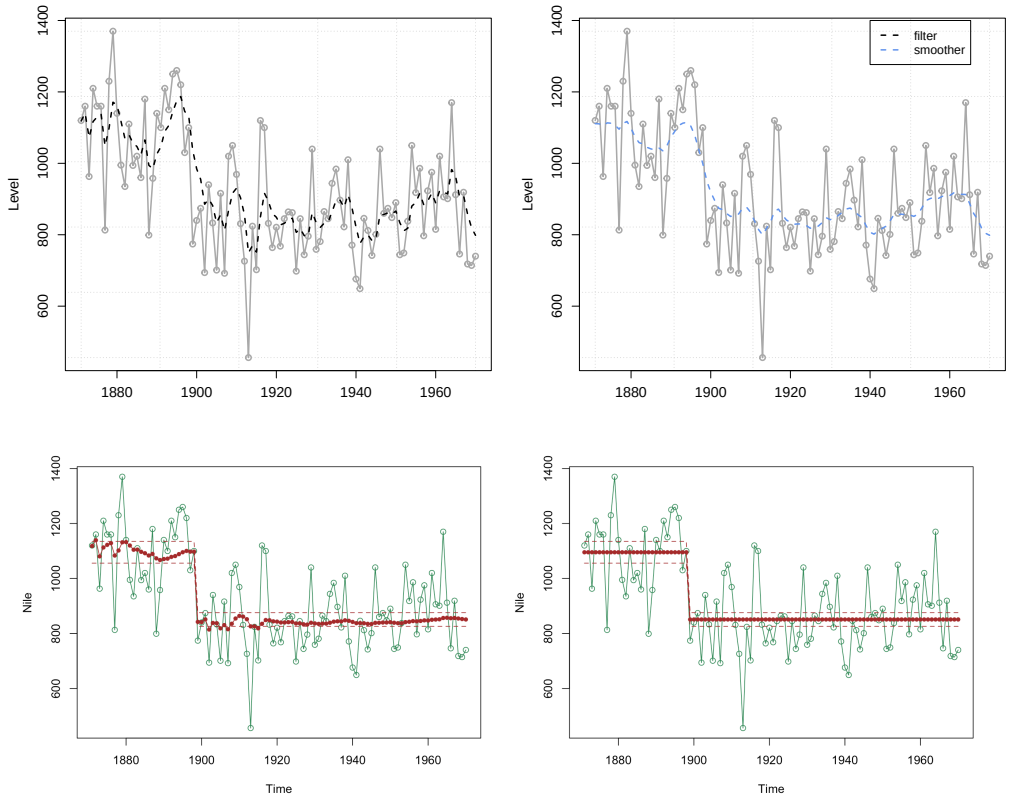

Figure 4.3: (Top panels) Measurements of the annual flow of the river Nile at Aswan 1871-1970, in 10^8 m^3, along with the means of the latent process obtained with Kalman filter (Top left) and smoother (Top right), using a local level model as in (4.8). (Bottom panels) The same data and model, with additional stochastic volatility in the (latent) state variable: the variance of the innovation is zero except in 1899. The panels report the filtered (bottom left) and smoothed (bottom right) state variables, together with confidence bands. The bottom model will be considered later in the lectures as a "change point" model.

Remark 4.3. The Kalman filter and smoother can be re-implemented but there are also many implementations already available. In `R`, the functions `Kfilter0`, `Ksmooth0` from the package

`astsa`, and the functions `dlmModPoly`, `dlm`, `dlmFilter`, `dlmSmooth`, from package `dlm` can be used to define linear Gaussian models and run these algorithms. The `dlm` package was used to obtain the filtering $m_{t|t}$ and smoothing $m_{t|n}$ means in a local level for the Nile river annual flows in the top graphs in Figure 4.3.

A similar local-level model was estimated in Figure 4.3 with an additional parameter that accounts for the large change in 1899.

4.6 Sampling paths

On top of obtaining the marginal mean and variance of the smoothing distributions, we might want to sample paths $x_{0:n}$ given the observations $y_{1:n}$ and parameters. Indeed, the Kalman smoother provides the mean and variance of distributions of X_t given $Y_{1:n} = y_{1:n}$, but we might be interested in quantities that are defined on the path space, and not on marginal distributions. For instance, we might wonder: what is the probability that the latent process exceeds a certain value η, at any time between $t = 0$ and $t = n$? This question can be formalized as the calculation of $\mathbb{P}((\max_{t=0,\dots,n}\ x_t) > \eta | Y_{1:n} = y_{1:n})$. This is not directly available given the marginal means $m_{t|n}$ and variances $C_{t|n}$. On the other hand, if we generate many paths $x_{0:n}$ following the distribution of $X_{0:n}$ given $Y_{1:n} = y_{1:n}$, then we could approximate the above probability with an empirical frequency computed with the paths, using the Monte Carlo principle.

To sample paths, we might start by running a Kalman filter, yielding $m_{n|n}$ and $C_{n|n}$ at the final step. We can sample the latest component of the path x_n from $\mathcal{N}\left(m_{n|n}, C_{n|n}\right)$. Then, we have already worked out above that the distribution of X_{t-1}, given $X_t = x_t$ and the observations, is given by a Normal distribution with mean and variance as in (4.14)-(4.15). Thus, we can sample X_{n-1} given $X_n = x_n$, and then X_{n-2} given $X_{n-1} = x_{n-1}$, etc, until we obtain a complete path $x_{0:n}$.

4.7 Missing observations

An appeal of state space models is that we can easily accommodate the common situation where some observations are missing. We can write the observations as $Y_{t_1}, \dots, Y_{t_n}$ where t_i are the observation times, which might not be consecutive times such as $1, 2, 3, \dots$. We can define the latent process (X_t) at all times, and not only the observation times. The Kalman filter/smoother can be adapted to this setting as follows.

For the filter, assume that t is such that Y_t is not observed. Assume that we have obtained $m_{t|t-1}, C_{t|t-1}$. We then "update" with $m_{t|t} = m_{t|t-1}, C_{t|t} = C_{t|t-1}$, as there is no new information at time t. We can directly proceed to the prediction step to obtain $m_{t+1|t}$,$C_{t+1|t}$ as previously described.

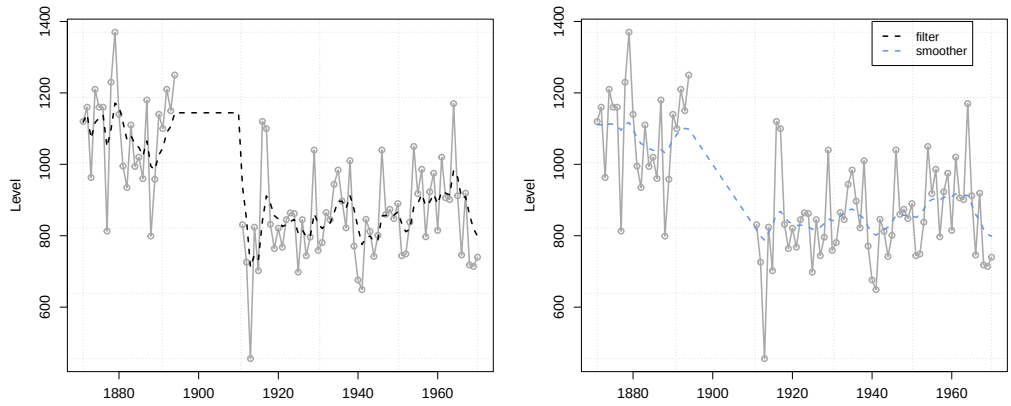

Figure 4.4: Means of the latent process obtained with Kalman filter (left) and smoother (right), using a local level model as in (4.8), for the Nile data where observations are missing from 1895 to 1910.

For the smoother, we observe that the backward updates do not explicitly involve the observation at time t, but only indirectly through the filtering mean and variance. Therefore the backward smoothing recursions can be employed without any modification.

Figure 4.4 shows the filtering and smoothing means in a local level model applied to the Nile river data, as in Figure 4.3, where we have removed observations from 1895 to 1910. We observe that the filtering means are constant in the period of missing data, while the smoothing means form a downward segment, anticipating the lower observations that arrive at future times.

4.8 Nonlinear and Markov switching models

We can wonder how Kalman calculations can extend to nonlinear and/or non-Gaussian models. Indeed Kalman filters have been modified to accommodate various models, but this often comes at a price: either some systematic bias must be accepted, as with "Extended Kalman filters", or some Monte Carlo error comes in, as with "particle filters" (or both, as in "Ensemble Kalman filters").[1] These techniques can be applied to the examples of target-tracking, stochastic volatility and disease outbreak mentioned early on in this chapter. Research on computational methods for generic state space models is still actively on-going.

A simple class of nonlinear state-space models in found in the literature on time-varying param-

[1] A particle filter is a simulation-based method that consists in generating artificial samples of the latent processes (the so-called particles). At time date t, the set of "particles" is assessed by a score to decide whether it is worth propagating them to the next date, using the conditional distribution $X_{t+1}|X_t = x_t$. In a sense this is a principle not unsimilar to that used by Deepmind to play the game of Go: the algorithm generated artificial moves and scored them by checking the frequency with which they would lead to a win.

eter (TVP) models. For instance, we could think of equations that involve the same variables (and lags thereof) but with parameters that change every period, such at

$$Y_t = a + b_t X_t + \epsilon_t$$

where Y_t and X_t are both observables (and ϵ_t may denote an ARMA process) but it is now the parameter b_t themselves which are latent, we may, e.g., assume they follow random walks:

$$b_t = b_{t-1} + W_t$$

This type of model is nonlinear requires that parameters change constantly, see for instance Stock and Watson (1996), *Evidence on Structural Instability in Macroeconomic Time Series Relations*, for an assessment of time variation in macroeconomic equations. This may be too much variation and we might want to restrict the amount of time variation to improve inference and interpretability.

Other models are related in spirit and involve similar computational questions. A prime example is "Markov switching" models. The idea is to allow the parameters of a model to abruptly change at some times, and to possibly come back to their previous values. This is useful to account for phenomena undergoing different dynamics, for example the economy "in a recession" or "in an expansion". For example, consider an AR(1) process with mean μ:

$$Y_t = \mu + \varphi(Y_{t-1} - \mu) + \sigma W_t,$$

where $W_t \overset{iid}{\sim} \mathcal{N}(0,1)$. If we suspect that the phenomenon under study can switch from one regime to another, we can consider the model

$$\begin{aligned} Y_t &= \mu^{(S_t)} + \varphi^{(S_t)}(Y_{t-1} - \mu^{(S_t)}) + \sigma^{(S_t)} W_t, \\ S_t &\sim P(\cdot|S_{t-1}), \quad \text{i.e.} \quad \mathbb{P}(S_t = s'|S_{t-1} = s) = P_{ss'} \end{aligned}$$

where (S_t) is a Markov chain taking two states, with transition matrix P, and $\mu^{(s)}, \varphi^{(s)}, \sigma^{(s)}$ can be different for different states s. Thus a Markov chain drives "switches" from one AR(1) model to another. This type of model is used to estimate the probability of the economy entering or leaving a state of recession, see James D. Hamilton, *Calling Recessions in Real Time*, 2010; see Figure 4.5.

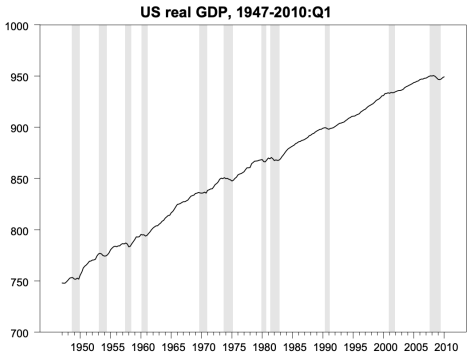

Figure 1. One hundred times the natural logarithm of U.S. real GDP, 1947:Q1-2010:Q1. Last shaded region covers 2007:Q4-2009:Q2; other shaded regions correspond to NBER recession dates.

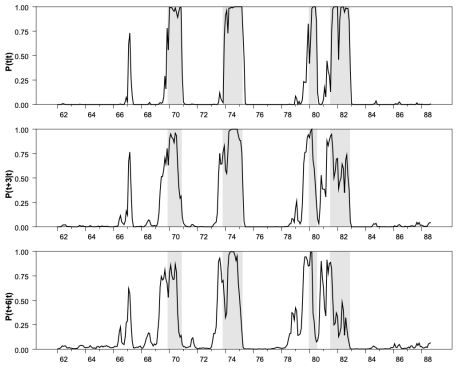

Figure 5. In-sample values for probability of recession from Stock-Watson business cycle model. Top panel: contemporaneous probability. Middle panel: 3-month-ahead forecast. Bottom panel: 6-month-ahead forecast. Adapted from Figure 2.1 in Stock and Watson (1993), using code provided by Mark Watson (http://www.princeton.edu/~mwatson/ddisk/pr.zip).

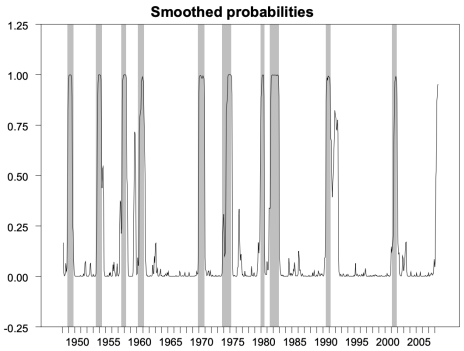

Figure 15. Inference from 3-phase model of unemployment as published on September 5, 2008. Graph shows $\Pr(S_t > 1 \mid y_T, y_{T-1}, \ldots, y_1; \hat{\theta}_T)$ for each t and is reproduced from www.econbrowser.com/archives/2008/09/rising_unemploy.html.

Figure 4.5: (Top left) US quarterly GDP, with periods of recession as indicated by the National Bureau of Economic Research (the NBER). A question is whether these periods can be determined from data in real time (it takes in general several years for the NBER to time the recessions in hindsight, but it is crucial for policy to be able to do so as early as possible). (Top right) Probabilities of recession obtained from a linear Gaussian state space model, where the latent process represents business cycles. (Bottom) Probabilities of recession obtained from a Markov switching AR model. Taken from James D. Hamilton, *Calling Recessions in Real Time*, 2010.