Chapter 2
Basic Models

The dynamic mixture estimation discussed in the book requires a reader to be familiar with the basic single models used as mixture components. The presented edition operates with the regression model, the categorical model, and the state-space model. This chapter recalls the Bayesian estimation algorithms existing for these models. This specific theoretical background helps a reader be effective and fast oriented in the subsequent text. However, a knowledge of the basics of statistical distributions and the Bayesian theory is assumed.

2.1 Regression Model

A well-known regression model is one of the most frequently used system descriptions. It is used for modeling a continuous output variable which is supposed to be linearly dependent on delayed output values and other present or also delayed variables. They can include optional inputs which control the output and external variables which influence the output but cannot be affected. An important element of the model can also be an absolute term of the model (constant) which represents a nonzero mean value of the modeled output.

The regression model has two parts: (i) deterministic, which is a difference equation on measured data and (ii) stochastic, which is represented by a noise term. Under the assumption of stationarity, it is a stochastic sequence whose elements are i.i.d. (independent and identically distributed) with zero expectations and constant variances (covariance matrices in a multivariate case).

The most frequently used distribution for noise is the normal one. Belonging to the exponential family, it can be easily estimated even in its multivariate form, using the conjugate Gauss-inverse-Wishart distribution (GiW) [30, 31, 34, 35]. However, other distributions belonging to the exponential family can be used as well and their estimation can be made feasible.

I. Nagy and E. Suzdaleva, *Algorithms and Programs of Dynamic Mixture Estimation*, SpringerBriefs in Statistics, DOI 10.1007/978-3-319-64671-8_2

The normal regression model is represented by the following probability density function (denoted by the pdf in the text)

$$f\left(y_t|\psi_t, \Theta\right) \tag{2.1}$$

that can be defined through the difference equation

$$y_t = \psi_t'\theta + e_t = b_0 u_t + \sum_{i=1}^{n}(a_i y_{t-i} + b_i u_{t-i}) + k + e_t, \tag{2.2}$$

where

- $t = 1, 2, \ldots$ denotes discrete time instants;
- y_t is the output variable;
- u_t is the control input variable;
- $\psi_t' = [u_t', y_{t-1}', u_{t-1}', \ldots, y_{t-n}', u_{t-n}', 1]$ is the regression vector;
- n is the memory length;
- $\Theta \equiv \{\theta, r\}$ are parameters, where
  - $\theta = [b_0, a_1, b_1, \ldots, a_n, b_n, k]$ is a collection of regression coefficients,
  - and r is the constant variance of the normal noise e_t with the zero expectation.

If y_t is a vector, it is the multivariate case and the parameters are matrices of appropriate dimensions. An example of a two-dimensional output and a three-dimensional input is given below. The model is

$$\begin{bmatrix} y_{1;t} \\ y_{2;t} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1;t-1} \\ y_{2;t-1} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{bmatrix} \begin{bmatrix} u_{1;t} \\ u_{2;t} \\ u_{3;t} \end{bmatrix} + \begin{bmatrix} k_1 \\ k_2 \end{bmatrix} + \begin{bmatrix} e_{1;t} \\ e_{2;t} \end{bmatrix}, \tag{2.3}$$

where $[y_{1;t},\ y_{2;t}]'$ are entries of the vector y_t. Corresponding matrices with regression coefficients $a_{11}, a_{12}, \ldots, b_{11}, b_{12}, \ldots$ and k_1, k_2 enter the collection θ. The noise has zero expectation $[0,\ 0]'$ and the constant (time invariant) covariance matrix

$$r = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}. \tag{2.4}$$

2.1.1 Estimation

The estimation of the model (2.1) obeys Bayes rule (7.25). According to [8, 30], the model (2.1) is rewritten as

$$f\left(y_t|\psi_t, \Theta\right) = (2\pi)^{-k_y/2}|r|^{-1/2} \exp\left\{-\frac{1}{2} tr\left(r^{-1}\begin{bmatrix} -I \\ \theta \end{bmatrix}' D_t \begin{bmatrix} -I \\ \theta \end{bmatrix}\right)\right\}, \tag{2.5}$$

where k_y denotes the dimension of the vector y_t; tr is a trace of the matrix; I is the unit matrix of the appropriate dimension and

$$D_t = \begin{bmatrix} y_t \\ \psi_t \end{bmatrix} \begin{bmatrix} y_t \\ \psi_t \end{bmatrix}' \tag{2.6}$$

is the so-called data matrix at time t. For a normal model (2.1) or in the rewritten form (2.5), the conjugate prior GiW pdf has the form

$$f(\Theta|d(t-1)) \propto |r|^{-0.5\kappa_{t-1}} \exp\left\{-\frac{1}{2} tr\left(r^{-1} \begin{bmatrix} -I \\ \theta \end{bmatrix}' V_{t-1} \begin{bmatrix} -I \\ \theta \end{bmatrix}\right)\right\} \tag{2.7}$$

with the recomputable statistics V_{t-1}, which is the information matrix, and κ_{t-1}, which is the counter of used data samples. Notation $d(t)$ corresponds to the data collection up to the time t, i.e., $d(t) \equiv \{d_0, d_1, \ldots, d_t\}$, where $d_t = \{y_t, u_t\}$, and d_0 is the prior information. It means that $d(t-1) \equiv \{d_0, d_1, \ldots, d_{t-1}\}$. After substituting (2.5) and (2.7) in (7.25), these statistics are recursively updated starting from chosen initial statistics V_0 and k_0 as follows:

$$V_t = V_{t-1} + \begin{bmatrix} y_t \\ \psi_t \end{bmatrix} \begin{bmatrix} y_t \\ \psi_t \end{bmatrix}' = V_{t-1} + \begin{bmatrix} \underbrace{y_t y_t'}_{D_y} & \underbrace{y_t \psi_t'}_{D_{y\psi}'} \\ \underbrace{\psi_t y_t'}_{D_{y\psi}} & \underbrace{\psi_t \psi_t'}_{D_\psi} \end{bmatrix}, \quad \kappa_t = \kappa_{t-1} + 1, \tag{2.8}$$

whereupon the updated matrix V_t is partitioned similarly to (2.8) keeping the appropriate dimensions, i.e.,

$$V_t = \begin{bmatrix} V_y & V_{y\psi}' \\ V_{y\psi} & V_\psi \end{bmatrix}. \tag{2.9}$$

2.1.2 Point Estimates

The updated statistics V_t is partitioned as it is shown in (2.8) with dimensions given by y_t and ψ_t. Then the point estimates of the regression coefficients θ and of the noise covariance matrix r are computed respectively

$$\hat{\theta}_t = V_\psi^{-1} V_{y\psi} \quad \text{and} \quad \hat{r}_t = \frac{V_y - V_{y\psi}' V_\psi^{-1} V_{y\psi}}{\kappa_t}. \tag{2.10}$$

Detailed information is available in [8, 30]. See also details in Appendix 7.9.1.

2.1.3 Prediction

The one-step output prediction with the regression model (2.1) is defined through the predictive pdf $f\left(y_t|u_t, d\left(t-1\right)\right)$. It can generally be computed in the following way:

$$f\left(y_t|u_t, d\left(t-1\right)\right) = \int_{\Theta^*} f\left(y_t, \Theta|u_t, d\left(t-1\right)\right) d\Theta =$$

$$= \int_{\Theta^*} f\left(y_t|\psi_t, \Theta\right) f\left(\Theta|d\left(t-1\right)\right) d\Theta, \tag{2.11}$$

where the first pdf inside the integral is the parametrized model, and the second one is the prior pdf (i.e., the posterior pdf from the last step of the estimation). The result of integration in this simple case is analytical.[1] However, integration is relatively complex.

Accepting the point rather than the full estimate gives us a considerable simplification. The point estimate $\hat{\Theta}_{t-1}$ of the parameter Θ can be introduced by substituting the Dirac delta function $\delta\left(\Theta, \hat{\Theta}_{t-1}\right)$, where $\hat{\Theta}_{t-1} = \{\hat{\theta}_{t-1}, \hat{r}_{t-1}\}$, instead of $f\left(\Theta|d\left(t-1\right)\right)$ into the above integral (see Appendix 7.9.3). Thus the simplified formula for prediction is obtained in the form

$$f\left(y_t|y\left(t-1\right)\right) = f\left(y_t|\psi_t, \hat{\Theta}_{t-1}\right) = N\left(\psi_t\hat{\theta}_{t-1}, \hat{r}_{t-1}\right), \tag{2.12}$$

where N ('expectation', 'variance') stands for the normal pdf.

This formula is not only very simple but it also has a very clear meaning: for prediction, take the model and substitute the point estimates instead of unknown parameters.

2.2 Categorical Model

A discrete model with the categorical distribution can be used if all involved variables are discrete or discretized. For practical reasons, this model is applicable only if the number of variables is not too high and also if the number of different values of individual variables is small. The discrete model assigns probabilities to configurations of variable values. If there are too many such configurations, the model has a very high dimension and becomes unfeasible.

The meaning of the discrete model is not the "proportionality" as it is in the regression model. It is rather in a "sorting"—"if the case is so and so, the result will belong to this category". It indicates that methods related to the discrete model are close to the classification. From this point of view, a value of the model output represents the label of the class to which the regression vector belongs.

[1] The result is the Student distribution [30].

The general form of the discrete model is described by the following probability function (also denoted by a pdf)

$$f\left(y_t = i|\psi_t = j, \Theta\right), \quad i \in y^*, \quad j \in \psi^*, \tag{2.13}$$

which is represented by the table of transition to the value $y_t = i$ under condition that the $\psi_t = j$, and where Θ is the parameter, and y^*, ψ^* are finite sets of integers. The regression vector ψ_t can generally include variables $[u_t, y_{t-1}, u_{t-1}, \ldots, y_{t-n}, u_{t-n}]$, where at each time instant the output y_t has m_y possible values and the input u_t has m_u possible values. However, a large number of variables in the regression vector increases the dimension of the transition table extremely. The dimension of the regression vector can be reduced to a scalar by coding its configurations $[1, 1] \to 1, [2, 1] \to 2, \ldots, [m_u, m_y] \to m_\psi$, where m_ψ is the whole number of configurations. The set of possible configurations of the regression vector ψ_t is denoted by $\psi^* \in \{1, \ldots, m_\psi\}$. Although the general form of the categorical model (2.13) is similar to the regression model pdf (2.1), here the parameter Θ is the matrix containing stationary transition probabilities $\Theta_{i|j}$ $\forall i \in y^*$ and $\forall j \in \psi^*$, and it holds

$$\Theta_{i|j} \geq 0, \sum_{i=1}^{m_y} \Theta_{i|j} = 1, \quad \forall i \in y^*, \forall j \in \psi^*, \tag{2.14}$$

which means that the sum of probabilities in each row is equal to 1. This statement is used for all discrete models throughout the text.

The regression vector ψ_t can also be kept as a vector. For instance, for $\psi_t = [u_t, y_{t-1}]$ the model (2.13) is the transition table with m_y columns and $m_y \times m_u$ rows, i.e., for $i, j \in y^*, k \in u^*$

$$f\left(y_t = i|u_t = k, y_{t-1} = j, \Theta\right) \equiv \tag{2.15}$$

	$y_t = 1$	$y_t = 2$	$\cdots$	$y_t = m_y$
$u_t = 1,\ y_{t-1} = 1$	$\Theta_{1\|11}$	$\Theta_{2\|11}$	$\cdots$	$\Theta_{m_y\|11}$
$u_t = 2,\ y_{t-1} = 1$	$\Theta_{1\|21}$	$\cdots$	$\cdots$	$\Theta_{m_y\|21}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$u_t = m_u, y_{t-1} = 1$	$\Theta_{1\|m_u 1}$	$\cdots$	$\cdots$	$\Theta_{m_y\|m_u 1}$
$u_t = 1,\ y_{t-1} = 2$	$\Theta_{1\|12}$	$\Theta_{2\|12}$	$\cdots$	$\Theta_{m_y\|12}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$u_t = m_u, y_{t-1} = 2$	$\Theta_{1\|m_u 2}$	$\cdots$	$\cdots$	$\Theta_{m_y\|m_u 2}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$u_t = 1,\ y_{t-1} = m_y$	$\Theta_{1\|1m_y}$	$\Theta_{2\|1m_y}$	$\cdots$	$\Theta_{m_y\|1m_y}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$u_t = m_u, y_{t-1} = m_y$	$\Theta_{1\|m_u m_y}$	$\cdots$	$\cdots$	$\Theta_{m_y\|m_u m_y}$

where each $\Theta_{i|kj}$ is the probability of the value $y_t = i$ conditioned by $u_t = k$ and $y_{t-1} = j$ and the statement (2.14) holds.

2.2.1 Estimation

According to [31], estimation of the discrete model (2.13) with the help of the Bayes rule (7.25) is based on using the conjugate prior Dirichlet distribution with the recursively updated statistics. The model (2.13) is written in the so-called product form [31]

$$f\left(y_t = i|\psi_t = j, \Theta\right) = \prod_{i\in y^*, j\in\psi^*} \left(\Theta_{i|j}\right)^{\delta(i,j;\, y_t,\psi_t)}, \quad i \in y^*, \quad j \in \psi^*, \tag{2.16}$$

where the Kronecker delta function $\delta\left(i, j;\ y_t, \psi_t\right) = 1$, when $i = y_t$ and $j = \psi_t$ and it is equal to 0 otherwise. The conjugate prior Dirichlet pdf for the model (2.13) has the form

$$f\left(\Theta|d(t-1)\right) \propto \prod_{i\in y^*, j\in\psi^*} \left(\Theta_{i|j}\right)^{(\nu_{i|j})_{t-1}-1}, \quad i \in y^*, \quad j \in \psi^*, \tag{2.17}$$

where the statistics ν_{t-1} is a matrix of the same dimension as (2.13) and $(\nu_{i|j})_{t-1}$ are its entries.

Substituting the model (2.16) and the prior pdf (2.17) into (7.25) gives updating the entries of the statistics ν_{t-1}

$$(\nu_{i|j})_t = (\nu_{i|j})_{t-1} + \delta\left(i, j;\ y_t, \psi_t\right), \quad \forall i \in y^*, \quad \forall j \in \psi^*, \tag{2.18}$$

with some chosen initial statistics ν_0. In practice, it means that the statistics counts occurrences of the combinations of values of y_t and ψ_t.

2.2.2 Point Estimates

The point estimate of the parameter Θ is obtained by normalizing the statistics ν_t

$$(\hat{\Theta}_{i|j})_t = \frac{(\nu_{i|j})_t}{\sum_{k=1}^{m_y} (\nu_{k|j})_t}, \quad i \in y^*, \quad j \in \psi^*. \tag{2.19}$$

The detailed information is available in [31]. See also Appendix 7.8.

2.2.3 Prediction

For the output prediction with the discrete model (2.13) it holds

$$f\left(y_t = i|u_t = k, d\left(t-1\right)\right) = \int_0^1 \underbrace{f\left(y_t = i|\psi_t = j, \Theta\right)}_{(2.13)} \underbrace{f\left(\Theta|d\left(t-1\right)\right) d\Theta}_{(2.17)} = (\hat{\Theta}_{i|j})_{t-1}, \tag{2.20}$$

which is the expectation of the Dirichlet distribution and the regression vector ψ_t is comprised from values of the input u_t and past data $d(t-1)$.

2.3 State-Space Model

A state is a variable whose statistical properties are fully determined by its last value and by the actual control input (if present). Its prediction depends only on its last value, not on the whole history of its evolution. Very often the state variable cannot be measured, i.e., it is modeled and estimated. For its description, the state-space model is suitable. It consists of two parts: the *state model* describes the state evolution in dependence on its last value and the control. The *output model* (sometimes called *the measurement model*) reflects the effect of the state and possibly of the control on the output variable.

The general form of the state-space model is presented in the following two pdfs:

$$f\left(x_t|x_{t-1}, u_t\right) \quad \text{state model,} \tag{2.21}$$

$$f\left(y_t|x_t, u_t\right) \quad \text{output model,} \tag{2.22}$$

where x_t denotes the unobservable state to be estimated. The linear normal state-space model described by these pdfs can be written with the help of the equations

$$x_t = Mx_{t-1} + Nu_t + F + \omega_t, \tag{2.23}$$

$$y_t = Ax_t + Bu_t + G + v_t, \tag{2.24}$$

where M, N, F, A, B and G are matrices of parameters of appropriate dimensions supposed to be known; ω_t and v_t are the process as well as the measurement Gaussian white noises with zero expectations and covariance matrices R_ω and R_v respectively, which are usually supposed to be known.

2.3.1 State Estimation

Bayesian state estimation (that can be found in various sources, e.g., [30, 36], etc.) operates with the prior state pdf $f\left(x_{t-1}|d(t-1)\right)$ and the state-space model (2.21)–(2.22) to obtain the posterior state pdf $f\left(x_t|d(t)\right)$ via the recursion

$$f\left(x_t|d\left(t-1\right)\right) = \int_{x^*} f\left(x_t|x_{t-1}, u_t\right) f\left(x_{t-1}|d\left(t-1\right)\right) dx_{t-1}, \tag{2.25}$$

$$f\left(x_t|d\left(t\right)\right) = \frac{f\left(y_t|x_t, u_t\right) f\left(x_t|d\left(t-1\right)\right)}{f\left(y_t|u_t, d\left(t-1\right)\right)}, \tag{2.26}$$

with

$$f\left(y_t|u_t, d\left(t-1\right)\right) = \int_{x^*} f\left(y_t|x_t, u_t\right) f\left(x_t|d\left(t-1\right)\right) dx_t. \tag{2.27}$$

The recursion starts with the chosen prior pdf $f\left(x_0|d(0)\right)$ representing prior knowledge about the state.

Substituting the linear form (2.23)–(2.24) into the above recursion gives us the famous Kalman filter [30, 37–39]. For linear normal models and the normal initial state given by the prior pdf

$$f\left(x_{t-1}|d(t-1)\right) = N(\hat{x}_{t-1|t-1}, R_{t-1|t-1}), \tag{2.28}$$

the state description during its evolution stays normal. With this prior pdf and the models (2.23) and (2.24), the Kalman filter can be presented as follows.

Prediction

$$\text{state prediction } \hat{x}_{t|t-1} = M\hat{x}_{t-1|t-1} + Nu_t + F, \tag{2.29}$$

$$\text{prediction of state covariance } R_{t|t-1} = R_\omega + MR_{t-1|t-1}M', \tag{2.30}$$

Filtration

$$\text{output prediction } \hat{y}_t = A\hat{x}_{t|t-1} + Bu_t + G, \tag{2.31}$$

$$\text{noise covariance update } R_y = R_v + AR_{t|t-1}A', \tag{2.32}$$

$$\text{update of the state covariance } R_{t|t} = R_{t|t-1} - R_{t|t-1}A'R_y^{-1}AR_{t|t-1}, \tag{2.33}$$

$$\text{Kalman gain } K_g = R_{t|t}A'R_v^{-1}, \tag{2.34}$$

$$\text{state correction } \hat{x}_{t|t} = \hat{x}_{t|t-1} + K_g\left(y_t - \hat{y}_t\right), \tag{2.35}$$

where $\hat{y}_t$ denotes the point prediction of the output obtained by substituting the current point estimates of the state. The point estimates of the state are given by the values of $\hat{x}_{t|t}$ after the state correction.

Remark In this way, this chapter summarizes the available knowledge on basic models which are used further for modeling the components and the pointer. Now, it is necessary to introduce the dynamic mixture model which is explained in the next chapter.

http://www.springer.com/978-3-319-64670-1

Algorithms and Programs of Dynamic Mixture Estimation
Unified Approach to Different Types of Components
Nagy, I.; Suzdaleva, E.
2017, XI, 113 p. 27 illus. in color., Softcover
ISBN: 978-3-319-64670-1