New Laboratory of Stochastic Systems

Šárka Jozová and Ivan Nagy

Laboratory of Applied Mathematics in Transport and Logistics (lambda.fd.cvut.cz/en)

Faculty of Transportation Sciences, CTU, Prague

Department of Applied Mathematics

Contents

1 Introduction

Stochastic systems deal with data analysis (modeling, prediction and classification) under uncertainty in the measured variables.

1.1 Variables and tasks

2 Models and their estimation

The Bayesian statistics is based on mathematical model of the investigated object (system). Generally, the model is conditional distribution of the target variable (output of the system) with condition containing values of explanatory variables (which enter the system and influence the target). It has the form

$$f\left(y_t|\psi_t,\Theta\right)$$

where y_t is the target variable at time t and ψ_t consists of values of those variables that influence y_t and Θ are model parameters.

The model structure is given by the choice of variables: target variable and its distribution and the values of variables in the regression vector. The final form of the model that really describes the monitored process (in the sense of close outputs for identical inputs) is achieved only in the process of estimation of the model parameters from the measured data.

2.1 Regression normal model

Most frequently used model for continuous data and standard (normally distributed) randomness.

2.2 Exponential model

Model for nonnegative continuous data with maximum frequency at zero.

2.3 Rayleigh model

Model for nonnegative continuous data with estimated maximum frequency.

2.4 Discrete categorical model

Basic model for discrete data. Prone to overparametrization.

2.5 Modified binomial model

Model for discrete data with fixed finite number of different values with estimated maximum frequency point.

2.6 Poisson model

Model for discrete data with infinite number of values. Suitable for modeling queue-theory tasks.

2.7 State-space model

Estimation of measured variable from measured input and output of the system. Filtration of noise.

2.8 Mixture model

Models multimodal or slightly nonlinear data. Performs classification.

3 Initialization of estimation

Bayesian approach to estimation has the great advantage of a possibility to set prior knowledge about the estimated system. Both, the information from prior data and that obtained from an expert can be used. As for the expert knowledge - the best way is if the expert expresses his knowledge through so called fictitious data, i.e. for example: "if the this road will be closed and the traffic in that one will be restricted, we can strongly expect that the density of the traffic in the monitored point will be very high", etc.

The basic idea for initialization is the following one: If we know (or guess) the parameter p value and the estimate is given as $\hat{p} = S/\kappa$ (S is sum and κ is count of data) then we set $\kappa = \kappa_0$ (corresponds with the number if data used) and $S_0 = \hat{p}\kappa_0$ which corresponds to the sum of κ_0 times added data leading to our estimate $\hat{p}$.

Example: We estimate expectation of data generated by the model

$$y_t = k + e_t$$

Statistics is $S_t = \sum_{i=1}^{t} y_i$; $\kappa_t = t$. *Estimate* $\hat{p} = \frac{S_t}{\kappa_t}$.

Now, we want to insert knowledge $\hat{p} = 2$.

- *Weak information set:* $\kappa_0 = 0.1$, $S_0 = \kappa_0\hat{p} = 0.2$. *(The estimate is* $\hat{p} = \frac{S_0}{\kappa_0} = \frac{0.2}{0.1} = 2$*).*
- *Strong information set:* $\kappa_0 = 100$, $S_0 = \kappa_0\hat{p} = 200$. *(The estimate is* $\hat{p} = \frac{S_0}{\kappa_0} = \frac{200}{100} = 2$*).*

However, in the first case the coming real data compete with 0.1 data records while in the second case with 100 data records.

3.1 Example to tossing a coin

Example to demonstrate the basic principle of initialization.

3.2 Initialization to single models

In some cases, when the information of the data is weak, the expert knowledge about the system is very important. In extreme situations, the model can be built primarily on the expert knowledge and the data only for some its correction.

3.3 Initialization to mixture models

With mixture estimation, the prior setting of components is crucial. If they are badly positioned, the estimation can be very long or rather fails. A strong advice is always use some real or expertly produced prior data for initial setting.

4 Prediction

What is prediction for (zero step, k-step, point prediction)

Prediction is estimation of the value of a future output. We speak about k-step prediction where k is number of steps for which we predict.

The agreement about timing is as follows: We are at time t, the output y_t has not been measured, yet, and the known data are $y(t-1) = \{y_0 y_1, y_2, \cdots, y_{t-1}\}$. If the regression vector contains other variables, they must be known ahead for the whole prediction interval. Thus, what we need to derive is the predictive pdf

$$f(y_{t+k}|y(t-1))$$

or point prediction which is the expectation

$$\hat{y}_{t+k} = E[y_{t+k}|y(t-1)]$$

There are two different cases

1. Zero-step prediction
$$f(y_t|y(t-1))$$
which predicts the value of the output that is just to be measured. After its measurement, the prediction error
$$e_t = y_t - \hat{y}_t$$
can be computed and its serves to evaluation of the quality of the model as predictor.

2. k-step prediction for $k > 0$ which really predicts the value of the future output.
Remark
This task requires dynamic model, i.e. dependence of y_t on its past value or values of several delayed outputs. This prediction is not possible with a static model.

For prediction (say zero-step) we need predictive pdf $f(y_t|y(t-1))$. It is similar to model. Only parameters are missing. How to get them inside?

4.1 Introduction to prediction

For a predictive model with known parameters (what is rather unrealistic from a practical viewpoint) the prediction can be made by recursively calling this model. If the parameters of the model are unknown, it is necessary to estimate them. A theoretical derivation of such prediction is indicated here.

4.2 Zero-step prediction with regression model

Zero-step prediction can be used for validation of the estimated model.

4.3 K-step prediction with regression model

K-step prediction can be used for estimation of the future behavior of the monitored system - e.g. a future level of service in a traffic system can be forecasted.

Remark

The **K-step prediction with state-space model** can be found in the following Section 3 Filtration in the paragraph **??**.

5 Filtration

Filtration means estimation of the values of a variable which is not measured and is influenced by input variables and influences the output one. Such variable is called the state x_t and we assume that it is of the first order (it depends only on its previous value and inputs).

The general model describing the state and its dependency on measured data is

$$f\left(x_t|x_{t-1},u_{t-1}\right), \quad f\left(y_t|x_t,u_t\right)$$

Its linear form is

$$x_t = Mx_{t-1} + Nu_{t-1} + w_t$$

$$y_t = Ax_t + Bu_t + v_t$$

where w_t and v_t are model noises.

5.1 State estimation

This is the basic task of the state estimation problem. The Kalman filter procedure is used to estimate the unmeasured state x_t form the input u_t and output y_t.

5.2 Nonlinear state estimation

The state estimation problem with nonlinear state-space model is solved by linearization of th model using the first two terms of Taylor expansion in the previous point estimate of the state.

5.3 Model with unknown parameters

The unknown parameters of the state-space model are included into the state variable. In this way we obtain a new model which, however, is nonlinear. Alter its linearization (see the above task) we can solve the task using Kalman filter.

5.4 State prediction

As the state-space model is dynamic it has sense to use it not only for estimation of values of an unknown variable but also for their prediction ahead in time. For this task, again, the Kalman filter procedure is used. The task solved can be e.g. prediction of the queue lengths on a crossroad.

6 Mixtures

Mixtures are represented by a set of single models - **components**, with any of the mentioned distribution

$$f_j\left(y_t|v_t,\Theta_j\right) \text{ or } f_j\left(v_t|\Theta_j\right), j=1,2,\cdots,n_c$$

where y_t is the target variable, $v_t=[v_1,v_2,\cdots,v_n]$ are explanatory variables and each component indexed by j has its own parameters Θ_j.

To be able to estimate mixture, we define a **pointer variable**. It is a discrete stochastic process c_t with random variables distributed uniformly or categorically and it is modeled either static or as a dynamic autoregression. The general model of the pointer is

$$f\left(c_t=j|\alpha\right)=\alpha_j$$

where $f\left(c_t=j|\cdot\right)=P\left(c_t=j|\cdot\right)$ is a probability of $c_t=j$.

The principle of estimation is: measure v_t, compute proximity of this data record to individual components, normalize the proximities to obtain weights, update all components statistics each with the corresponding weight, determine the point estimates of component parameters (for computation of the proximities in the next step).

As a result of the estimation we usually take the point estimates of the pointer $\hat{c}_t$ which classifies the coming data to the components. It also can create the overall mixture model in he form

$$f\left(y_t|v_t,\Theta\right)=\sum_{j=1}^{n_c}\alpha_j f_j\left(y_t|v_t,\Theta_j\right) \text{ or } f\left(v_t|\Theta\right)=\sum_{j=1}^{n_c}\alpha_j f_j\left(v_t|\Theta_j\right)$$

In any case, mixtures are used for modeling multimodal data - each mode is captured by one component.

6.1 Mixture estimation (basic example)

Here, the basic procedure of mixture estimation is introduced regardless on the concrete component model and uniform distribution of the pointer variable. A basic scheme for mixture estimation is presented.

6.2 Mixtures with descriptive components

The most frequently case. The model searches in n-dimensional space for density clusters. The result is clustering of the space and classification on newly measured record of values of the variables into the clusters.

6.3 Mixtures with explanatory components

The components in this case are characterized by similarity of the relation between the target variable and the explanatory ones. The goal is to model locally (within clusters) the dependency of the target on the explanatory variables. The model can also be used for zero-step prediction.

6.4 Mixtures with predictive components

The components here are dynamic models. The resulting mixture can be used for prediction of the target of a multimodal system.

7 MetaMixtures

They go behind the mixtures and introduce some special assumptions under which the dimensionality of the task can be reduced. The basic assumption is conditional independency of explanatory variables on condition of the pointer value. It means local independency in clusters. This assumption leads to scalar models of components, i.e. the dependency of the target variable always on a single explanatory variable. The final predictive pdf is composed from the local models using methodology of Naive Bayes.

7.1 Estimation of marginal mixtures

Here, the basis of theory of the marginal mixtures estimation is sketched. It is demonstrated for both method with common components and that with different components.

8 Special tasks with mixtures

The presented idea of mixtures with normal or categorical components has extensions in many directions. We can use it with other types of components, with the result of prediction or classification, we can substitute logistic regression or the famous iterative EM algorithm and some others. In this chapter we are going to show some of them.

8.1 Mixtures with uniform components

Here, a general case of uniform mixture estimation is solved. The statistics for estimation can be selected either according to the maximum likelihood method or method of moments. Although the likelihood method is more commonly used, here the method of moments is more convenient.

9 Appendix

9.1 Textbook to Stochastic Systems

9.2 Textbook to Mathematical Methods for Data Analysis

9.3 Textbook for PhD students