Parameter Estimation for Linear Dynamical Systems

Zoubin Ghahramani
Geoffrey E. Hinton
Department of Computer Science
University of Toronto
6 King's College Road
Toronto, Canada M5S 1A4

Email: zoubin@cs.toronto.edu

Technical Report CRG-TR-96-2
February 22, 1996

Abstract

Linear systems have been used extensively in engineering to model and control the behavior of dynamical systems. In this note, we present the Expectation Maximization (EM) algorithm for estimating the parameters of linear systems (Shumway and Stoffer, 1982). We also point out the relationship between linear dynamical systems, factor analysis, and hidden Markov models.

Introduction

The goal of this note is to introduce the EM algorithm for estimating the parameters of linear dynamical systems (LDS). Such linear systems can be used both for supervised and unsupervised modeling of time series. We first describe the model and then briefly point out its relation to factor analysis and other data modeling techniques.

The Model

Linear time-invariant dynamical systems, also known as linear Gaussian state-space models, can be described by the following two equations:

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + \mathbf{w}_t \tag{1}$$

$$\mathbf{y}_t = C\mathbf{x}_t + \mathbf{v}_t. \tag{2}$$

Time is indexed by the discrete index t. The output $\mathbf{y}_t$ is a linear function of the state, $\mathbf{x}_t$, and the state at one time step depends linearly on the previous state. Both state and output noise, $\mathbf{w}_t$ and $\mathbf{v}_t$, are zero-mean normally distributed random variables with covariance matrices Q and R, respectively. Only the output of the system is observed, the state and all the noise variables are hidden.

Rather than regarding the state as a deterministic value corrupted by random noise, we combine the state variable and the state noise variable into a single Gaussian random

variable; we form a similar combination for the output. Based on (1) and (2) we can write the conditional densities for the state and output,

$$P(\mathbf{y}_t|\mathbf{x}_t) = \exp\left\{-\frac{1}{2}[\mathbf{y}_t - C\mathbf{x}_t]'R^{-1}[\mathbf{y}_t - C\mathbf{x}_t]\right\}(2\pi)^{-p/2}|R|^{-1/2} \quad (3)$$

$$P(\mathbf{x}_t|\mathbf{x}_{t-1}) = \exp\left\{-\frac{1}{2}[\mathbf{x}_t - A\mathbf{x}_{t-1}]'Q^{-1}[\mathbf{x}_t - A\mathbf{x}_{t-1}]\right\}(2\pi)^{-k/2}|Q|^{-1/2} \quad (4)$$

A sequence of T output vectors $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$ is denoted by $\{\mathbf{y}\}$; a subsequence $(\mathbf{y}_{t_0}, \mathbf{y}_{t_0+1}, \ldots, \mathbf{y}_{t_1})$ by $\{\mathbf{y}\}_{t_0}^{t_1}$; similarly for the states.

By the Markov property implicit in this model,

$$P(\{\mathbf{x}\}, \{\mathbf{y}\}) = P(\mathbf{x}_1)\prod_{t=2}^{T} P(\mathbf{x}_t|\mathbf{x}_{t-1})\prod_{t=1}^{T} P(\mathbf{y}_t|\mathbf{x}_t). \quad (5)$$

Assuming a Gaussian initial state density

$$P(\mathbf{x}_1) = \exp\left\{-\frac{1}{2}[\mathbf{x}_1 - \boldsymbol{\pi}_1]'V_1^{-1}[\mathbf{x}_1 - \boldsymbol{\pi}_1]\right\}(2\pi)^{-k/2}|V_1|^{-1/2}. \quad (6)$$

Therefore, the joint log probability is a sum of quadratic terms,

$$\begin{aligned}\log P(\{\mathbf{x}\}, \{\mathbf{y}\}) = & -\sum_{t=1}^{T}\left(\frac{1}{2}[\mathbf{y}_t - C\mathbf{x}_t]'R^{-1}[\mathbf{y}_t - C\mathbf{x}_t]\right) - \frac{T}{2}\log|R| \\ & -\sum_{t=2}^{T}\left(\frac{1}{2}[\mathbf{x}_t - A\mathbf{x}_{t-1}]'Q^{-1}[\mathbf{x}_t - A\mathbf{x}_{t-1}]\right) - \frac{T-1}{2}\log|Q| \\ & -\frac{1}{2}[\mathbf{x}_1 - \boldsymbol{\pi}_1]'V_1^{-1}[\mathbf{x}_1 - \boldsymbol{\pi}_1] - \frac{1}{2}\log|V_1| - \frac{T(p+k)}{2}\log 2\pi. \quad (7)\end{aligned}$$

Often the inputs to the system can also be observed. In this case, the goal is to model the input–output response of a system. Denoting the inputs by $\mathbf{u}_t$, the state equation is

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t + \mathbf{w}_t. \quad (8)$$

where B is the input matrix relating inputs linearly to states. We will present the learning algorithm for the output-only case, although the extensions to the input–output case are straightforward.

If only the outputs of the system can be observed the problem can be seen as an *unsupervised* problem. That is, the goal is to model the unconditional density of the observations. If both inputs and outputs are observed, the problem becomes *supervised*, modeling the conditional density of the output given the input.

Related Methods

In its unsupervised incarnation, this model is an extension of maximum likelihood factor analysis (Everitt, 1984). The factor, $\mathbf{x}_t$, evolves over time according to linear dynamics. In factor analysis, a further assumption is made that the output noise along each dimension

is uncorrelated, i.e. that R is diagonal. The goal of factor analysis is therefore to compress the correlational structure of the data into the values of the lower dimensional factors, while allowing independent noise terms to model the uncorrelated noise. The assumption of a diagonal R matrix can also be easily incorporated into the estimation procedure for the parameters of a linear dynamical system.

The linear dynamical system can also be seen as a continuous-state analogue of the hidden Markov model (HMM; see Rabiner and Juang, 1986, for a review). The forward part of the forward-backward algorithm from HMMs is computed by the well-known Kalman filter in LDSs; similarly, the backward part is computed by using Rauch's recursion (Rauch, 1963). Together, these two recursions can be used to solve the problem of inferring the probabilities probabilities of the states given the observation sequence (known in engineering as the *smoothing* problem). These posterior probabilities form the basis of the E step of the EM algorithm.

Finally, linear dynamical systems can also be represented as graphical probabilistic models (sometimes referred to as belief networks). The Kalman-Rauch recursions are special cases of the probability propagation algorithms that have been developed for graphical models (Lauritzen and Spiegelhalter, 1988; Pearl, 1988).

The EM Algorithm

Shumway and Stoffer (1982) presented an EM algorithm for linear dynamical systems where the observation matrix, C, is known. Since then, many authors have presented closely related models and extensions, also fit with the EM algorithm (Shumway and Stoffer, 1991; Kim, 1994; Athaide, 1995). Here we present a basic form of the EM algorithm with C unknown, an obvious modification of Shumway and Stoffer's original work. This note is meant as a succinct review of this literature for those wishing to implement learning in linear dynamical systems.

The E step of EM requires computing the expected log likelihood,

$$\mathcal{Q} = E[\log P(\{\mathbf{x}\}, \{\mathbf{y}\})|\{\mathbf{y}\}]. \tag{9}$$

This quantity depends on three expectations—$E[\mathbf{x}_t|\{\mathbf{y}\}]$, $E[\mathbf{x}_t\mathbf{x}_t'|\{\mathbf{y}\}]$, $E[\mathbf{x}_t\mathbf{x}_{t-1}'|\{\mathbf{y}\}]$—which we will denote by the symbols:

$$\hat{\mathbf{x}}_t \equiv E[\mathbf{x}_t|\{\mathbf{y}\}] \tag{10}$$

$$P_t \equiv E[\mathbf{x}_t\mathbf{x}_t'|\{\mathbf{y}\}] \tag{11}$$

$$P_{t,t-1} \equiv E[\mathbf{x}_t\mathbf{x}_{t-1}'|\{\mathbf{y}\}]. \tag{12}$$

Note that the state estimate, $\hat{\mathbf{x}}_t$, differs from the one computed in a Kalman filter in that it depends on past *and future* observations; the Kalman filter estimates $E[\mathbf{x}_t|\{\mathbf{y}\}_1^t]$ (Anderson and Moore, 1979). We first describe the M step of the parameter estimation algorithm before showing how the above expectations are computed in the E step.

1 The M step

The parameters of this system are A, C, R, Q, $\boldsymbol{\pi}_1$, V_1. Each of these is re-estimated by taking the corresponding partial derivative of the expected log likelihood, setting to zero, and solving. This results in the following:

- Output matrix:

$$\frac{\partial \mathcal{Q}}{\partial C} = -\sum_{t=1}^{T} R^{-1}\mathbf{y}_t\hat{\mathbf{x}}_t' + \sum_{t=1}^{T} R^{-1}CP_t = 0 \tag{13}$$

$$C^{\text{new}} = \left(\sum_{t=1}^{T}\mathbf{y}_t\hat{\mathbf{x}}_t'\right)\left(\sum_{t=1}^{T}P_t\right)^{-1} \tag{14}$$

- Output noise covariance:

$$\frac{\partial \mathcal{Q}}{\partial R^{-1}} = \frac{T}{2}R - \sum_{t=1}^{T}\left(\frac{1}{2}\mathbf{y}_t\mathbf{y}_t' - C\hat{\mathbf{x}}_t\mathbf{y}_t' + \frac{1}{2}CP_tC'\right) = 0 \tag{15}$$

$$R^{\text{new}} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{y}_t\mathbf{y}_t' - C^{\text{new}}\hat{\mathbf{x}}_t\mathbf{y}_t'\right) \tag{16}$$

- State dynamics matrix:

$$\frac{\partial \mathcal{Q}}{\partial A} = -\sum_{t=2}^{T} Q^{-1}P_{t,t-1} + \sum_{t=2}^{T} Q^{-1}AP_{t-1} = 0 \tag{17}$$

$$A^{\text{new}} = \left(\sum_{t=2}^{T}P_{t,t-1}\right)\left(\sum_{t=2}^{T}P_{t-1}\right)^{-1} \tag{18}$$

- State noise covariance:

$$\begin{aligned}\frac{\partial \mathcal{Q}}{\partial Q^{-1}} &= \frac{T-1}{2}Q - \frac{1}{2}\sum_{t=2}^{T}\left(P_t - AP_{t-1,t} - P_{t,t-1}A' + AP_{t-1}A'\right) = 0 \\ &= \frac{T-1}{2}Q - \frac{1}{2}\left(\sum_{t=2}^{T}P_t - A^{\text{new}}\sum_{t=2}^{T}P_{t-1,t}\right) \qquad (19)\end{aligned}$$

$$Q^{\text{new}} = \frac{1}{T-1}\left(\sum_{t=2}^{T}P_t - A^{\text{new}}\sum_{t=2}^{T}P_{t-1,t}\right) \tag{20}$$

- Initial state mean:

$$\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\pi}_1} = (\hat{\mathbf{x}}_1 - \boldsymbol{\pi}_1)V_1^{-1} = 0 \tag{21}$$

$$\boldsymbol{\pi}_1^{\text{new}} = \hat{\mathbf{x}}_1 \tag{22}$$

- Initial state covariance:

$$\frac{\partial \mathcal{Q}}{\partial V_1^{-1}} = \frac{1}{2} V_1 - \frac{1}{2}(P_1 - \hat{\mathbf{x}}_1 \boldsymbol{\pi}_1' - \boldsymbol{\pi}_1 \hat{\mathbf{x}}_1' + \boldsymbol{\pi}_1 \boldsymbol{\pi}_1') \tag{23}$$

$$V_1^{\text{new}} = P_1 - \hat{\mathbf{x}}_1 \hat{\mathbf{x}}_1' \tag{24}$$

The above equations can be readily generalized to multiple observation sequences, with one subtlety regarding the estimate of the initial state covariance. Assume N observation sequences of length T, let $\hat{\mathbf{x}}_t^{(i)}$ be the estimate of state at time t given the i^{th} sequence, and

$$\bar{\hat{\mathbf{x}}}_t = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{x}}_t^{(i)}.$$

Then the initial state covariance is

$$V_1^{\text{new}} = P_1 - \bar{\hat{\mathbf{x}}}_1 \bar{\hat{\mathbf{x}}}_1' + \frac{1}{N} \sum_{i=1}^{N} [\hat{\mathbf{x}}_1^{(i)} - \bar{\hat{\mathbf{x}}}_1] [\hat{\mathbf{x}}_1^{(i)} - \bar{\hat{\mathbf{x}}}_1]'. \tag{25}$$

2 The E step

Using $\mathbf{x}_t^\tau$ to denote $E(\mathbf{x}_t | \{\mathbf{y}\}_1^\tau)$, and V_t^τ to denote $\text{Var}(\mathbf{x}_t | \{\mathbf{y}\}_1^\tau)$, we obtain the following Kalman filter forward recursions:

$$\begin{aligned}
\mathbf{x}_t^{t-1} &= A\mathbf{x}_{t-1}^{t-1} && (26)\\
V_t^{t-1} &= AV_{t-1}^{t-1}A' + Q && (27)\\
K_t &= V_t^{t-1}C'(CV_t^{t-1}C' + R)^{-1} && (28)\\
\mathbf{x}_t^t &= \mathbf{x}_t^{t-1} + K_t(\mathbf{y}_t - C\mathbf{x}_t^{t-1}) && (29)\\
V_t^t &= V_t^{t-1} - K_t C V_t^{t-1}, && (30)
\end{aligned}$$

where $\mathbf{x}_1^0 = \boldsymbol{\pi}_1$ and $V_1^0 = V_1$. Following Shumway and Stoffer (1982), to compute $\hat{\mathbf{x}}_t \equiv \mathbf{x}_t^T$ and $P_t \equiv V_t^T + \mathbf{x}_t^T \mathbf{x}_t^{T'}$ one performs a set of backward recursions using

$$\begin{aligned}
J_{t-1} &= V_{t-1}^{t-1}A'(V_t^{t-1})^{-1} && (31)\\
\mathbf{x}_{t-1}^T &= \mathbf{x}_{t-1}^{t-1} + J_{t-1}(\mathbf{x}_t^T - A\mathbf{x}_{t-1}^{t-1}) && (32)\\
V_{t-1}^T &= V_{t-1}^{t-1} + J_{t-1}(V_t^T - V_t^{t-1})J_{t-1}'. && (33)
\end{aligned}$$

We also require $P_{t,t-1} \equiv V_{t,t-1}^T + \mathbf{x}_t^T \mathbf{x}_{t-1}^{T'}$, which can be obtained through the backward recursions

$$V_{t-1,t-2}^T = V_{t-1}^{t-1} J_{t-2}' + J_{t-1}(V_{t,t-1}^T - AV_{t-1}^{t-1})J_{t-2}', \tag{34}$$

which is initialized $V_{T,T-1}^T = (I - K_T C)A\, V_{T-1}^{T-1}$.

References

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering.* Prentice-Hall, Englewood Cliffs, NJ.

Athaide, C. R. (1995). *Likelihood Evaluation and State Estimation for Nonlinear State Space Models.* Ph.D. Thesis, Graduate Group in Managerial Science and Applied Economics, University of Pennsylvania, Philadelphia, PA.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models.* Chapman and Hall, London.

Kim, C.-J. (1994). Dynamic linear models with Markov-switching. *J. Econometrics*, 60:1–22.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistical Society B*, pages 157–224.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

Rabiner, L. R. and Juang, B. H. (1986). An Introduction to hidden Markov models. *IEEE Acoustics, Speech & Signal Processing Magazine*, 3:4–16.

Rauch, H. E. (1963). Solutions to the linear smoothing problem. *IEEE Transactions on Automatic Control*, 8:371–372.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis*, 3(4):253–264.

Shumway, R. H. and Stoffer, D. S. (1991). Dynamic linear models with switching. *J. Amer. Stat. Assoc.*, 86:763–769.