RECONSTRUCTION AND PREDICTION OF NONLINEAR DYNAMICAL SYSTEMS : A HIERARCHICAL BAYES APPROACH WITH NEURAL NETS

T.Matsumoto, M.Saito, Y.Nakajima, J.Sugi and H.Hamagishi

Department of Electrical, Electronics and Computer Engineering
Waseda University
3-4-1 Ohkubo, Sinjuku-ku, Tokyo, Japan, 169-8555
takashi@mse.waseda.ac.jp

ABSTRACT

When nonlinearity is involved, time series prediction becomes a rather difficult task where the conventional linear methods have limited successes for various reasons.

One of the greatest challenges stems from the fact that typical observation data is a *scalar* time series so that dimension of the nonlinear dynamical system (*embedding dimension*) is unknown.

This paper proposes a Hierarchical Bayesian approach to nonlinear time series prediction problems. This class of schemes considers a *family* of prior distributions parameterized by hyperparameters instead of a single prior so that it enables algorithms less dependent on a particular prior. One can estimate posterior of weight parameters, hyperparameters and embedding dimension by marginalization with respect to the weight parameters and hyperparameters.

The proposed scheme is tested against two examples;
(i) chaotic time series, and
(ii) building air-conditioning load prediction.

1. FORMULATION

Problem A:
Given data set $D := \{x_t\}_{t=0}^{N} \subset \mathbb{R}$, predict $\{x_t\}_{t=N+1}^{T}$

Hypothesis $\mathcal{H}$

Hypothesis or model consists of the following:

(i) **Architecture**:
e.g., three-layer perceptron with h hidden units and a particular sigmoid function.

(ii) **Likelihood**:

$$P\left(\{x_t\}_{t=\tau}^{N}, \{x_{\tau-1}, \ldots, x_0\} \mid w, \beta, \mathcal{H}\right)$$
$$:= \underbrace{\prod_{t=0}^{N-\tau+1} \frac{1}{Z_D(\beta)} \exp\left(-\beta E_D(x_{t+\tau} \mid x_{t+\tau-1}, \ldots, x_t; w)\right)}_{\text{noisy dynamics}} \times \underbrace{P(x_{\tau-1}, \ldots, x_0 \mid \mathcal{H})}_{\text{initial state uncertainty}} \quad (1.1)$$

$$E_D(x_{t+\tau} \mid x_{t+\tau-1}, \ldots, x_t; w) := \frac{1}{2}(x_{t+\tau} - f(x_{t+\tau-1}, \ldots, x_t; w))^2 \quad (1.2)$$

where $f(\cdot)$ is neural net output, $w \in \mathbb{R}^k$ the weight parameters of a particular architecture, β (unknown) uncertainty level, $Z_D(\beta)$ the normalization constant, and τ is *embedding dimension* (the order of the dynamics) which is unknown. Equation(1.1) looks at $\{x_t\}$ as a τ-th order Markov process whose state transition probability density is given by the first factor whereas the second factor is the initial state probability density.

(iii) **Prior** for w:

$$P(w \mid \alpha, \mathcal{H}) := \prod_{c=1}^{C} \frac{1}{Z_W(\alpha_c)} \exp(-\alpha_c E_{W_c}(w_c)) \quad (1.3)$$

$$E_{w_c}(w_c) := \frac{1}{2} \| w_c \|^2 \quad (1.4)$$

where w is decomposed into groups:

$$w := (w_1, \ldots, w_C), \ w_c \in \mathbb{R}^{k_c}, \quad (1.5)$$
$$\alpha := (\alpha_1, \ldots, \alpha_C), \ \alpha_c \in \mathbb{R} \quad (1.6)$$

$\exp(-\alpha_c E_{W_c}(w_c))/Z_W(\alpha_c)$ represents the prior belief on how w_c should be distributed with (unknown) α_c and $Z_W(\alpha_c)$ is the normalization constant.

(iv) **Prior** for (α, β), hyperparameters: $P(\alpha, \beta \mid \mathcal{H})$

(v) **Prior** for $\mathcal{H}$: $P(\mathcal{H})$

The goal of the prediction problem is to compute the predictive distribution(density) $P(\{x_t\}_{t=N+1}^{T} \mid D)$ under (i) – (v). This paper first computes three levels of posterior distributions as shown in Fig. 1.1 and use them to compute the predictive distribution.

The most difficult parameter to be estimated is τ, the embedding dimension. In order to explain this, let us first consider the linear dynamical system

$$y_{t+1} = F y_t, \quad y_t \in \mathbb{R}^K, \qquad x_t = G^T y_t, \quad x_t \in \mathbb{R} \quad (1.7)$$

i.e., G^T represents a linear observation, T being matrix transpose. One can show that generically, that there is a nonsingular matrix Φ such that

$$(x_t, x_{t-1}, \ldots, x_{t-K+1}) = \Phi y_t$$

so that the K-dimensional *delay coordinate system* $x_t = (x_t, x_{t-1}, \ldots, x_{t-K+1})$ preserves various properties of (1.7). Well known AR model is described by

$$x_{t+1} = \sum_{i=0}^{K-1} w_i x_{t-i} + \nu \tag{1.8}$$

where ν is a noise process. Note that (1.1) contains (1.8) as a special case where

$$f(w; x_t, \ldots, x_{t-K+1}) = \sum_{i=0}^{K-1} w_i x_{t-i}, \quad \nu \sim i.i.d.\ N(0, 1/\beta)$$

Since AR model demands $\{w_i\}$ be (asymptotically) stable, the origin is the **only** meaningful **invariant set**. In contrast, nonlinear dynamical system

$$y_{t+1} = F(y_t), \quad y_t \in \mathbb{R}^K \tag{1.9}$$

can naturally admit non-trivial stable periodic orbits, invariant closed curves and even chaotic attractors which typically have Cantor structure. Let $Y \subset \mathbb{R}^K$ be an invariant set and let $x_t = G(y_t),\ x_t \in \mathbb{R}$ be observation. Determining the number of delay coordinates $(x_t, x_{t-1}, \ldots, x_{t-\tau+1})$ is non-trivial. The following is due to Sauer and others [1].

Fact 1.1 Let the invariant set Y be a compact subset of an open set $U \subset \mathbb{R}^K$, with **box counting dimension** d [1]. If

$$\tau > 2d \tag{1.10}$$

then for almost every smooth observation G, the delay coordinate map $y_t \longmapsto (x_t, x_{t-1}, \ldots, x_{t-\tau+1})$ is

(i) One-to-one on Y;

(ii) An immersion on each compact subset of a smooth manifold contained in Y, provided that several regularity conditions are met on periodic points.

Since $y_t \longmapsto (x_t, x_{t-1}, \ldots, x_{t-\tau+1})$ is one-to-one (for almost every G), delay coordinate system suffices for prediction purposes. Positive Lyapunov exponents can be computed since unstable manifold is preserved. Note, however, that the result is for noiseless dynamics. Note also that (1.10) is a sufficient condition so that $\tau \leq 2d$ may "work".

Decomposition (1.5) of weight parameters and associated decomposition (1.6) of hyperparameters are important. Typically a subvector w_c consists of those weights between each input variable to feedforward neural net and hidden units so that dim $w_c = h$, the number of hidden units. Another typical $w_{c'}$ consists of the biases for hidden units, and finally the bias for output unit together with the weights between hidden units and the output. Thus a typical dimension of α is $\tau + 2$, where τ is the hypothesized order of the Markov process.

[1] Let $N(\varepsilon)$ be the number of K-cubes needed to cover Y. Box counting dimension of Y is given by

$$d := \lim_{\varepsilon \to 0} \frac{\log N(\varepsilon)}{\log \frac{1}{\varepsilon}}$$

provided it exists which can be non-integer.

2. PREDICTIONS

Fact 2.1 (Level 1: Posterior for w)

The posterior of w given $(D, \alpha, \beta, \mathcal{H})$ is

$$P(w \mid D, \alpha, \beta, \mathcal{H}) = \frac{\frac{\exp(-M(w;\alpha,\beta))}{Z_D(\beta) Z_W(\alpha)}}{P(D \mid \alpha, \beta, \mathcal{H})} \tag{2.1}$$

$$M(w; \alpha, \beta) := \beta E_D(w) + \sum_{c=1}^{C} \alpha_c E_{W_c}(w_c) \tag{2.2}$$

and hence the most probable w, called w_{MP}, is given by

$$\boxed{w_{\mathrm{MP}} = \arg\min_w M(w; \alpha, \beta)} \tag{2.3}$$

Fact 2.2 (Level 2: Posterior for (α, β))

If $P(\alpha, \beta \mid \mathcal{H})$ is independent and flat, then the most probable hyperparameters are given by

$$\boxed{(\alpha_{\mathrm{MP}}, \beta_{\mathrm{MP}}) = \arg\max_{\alpha,\beta} P(D \mid \alpha, \beta, \mathcal{H})} \tag{2.4}$$

so that the following gradient information can be used for finding $(\alpha_{\mathrm{MP}}, \beta_{\mathrm{MP}})$:

$$\frac{\partial}{\partial \beta} \log P(D \mid \alpha, \beta, \mathcal{H}) \approx -E_D(w_{\mathrm{MP}}) - \frac{1}{2} \mathrm{Tr} A^{-1} B_D - \frac{\partial}{\partial \beta} \log Z_D(\beta) \tag{2.5}$$

where A is the Hessian of M evaluated at w_{MP}, Tr stands for a trace of a matrix, E_D is defined by (1.2) and B_D is the Hessian of E_D at w_{MP}.

$$\frac{\partial}{\partial \alpha_c} \log P(D \mid \alpha, \beta, \mathcal{H}) \approx -E_{W_c}(w_{c\mathrm{MP}}) - \frac{\partial}{\partial \alpha_c} \log Z_W(\alpha) - \frac{1}{2} \mathrm{Tr} A^{-1} B_C \tag{2.6}$$

where B_C is the Hessian of E_{W_c} at w_{MP}.

Fact 2.3 (Level 3: Posterior for $\mathcal{H}$ (model comparison))

If $P(\mathcal{H})$ is flat, then the most probable model is given by

$$\boxed{\mathcal{H}_{\mathrm{MP}} = \arg\max_{\mathcal{H}} P(D \mid \mathcal{H})} \tag{2.7}$$

Fact 2.4 (Predictive Distribution)

$$P(\{x_t\}_{N+1}^T \mid D) = \sum_{\mathcal{H}} \int\int\int P(\{x_t\}_{N+1}^T \mid w, \beta, \mathcal{H}) \times P(w, \alpha, \beta, \mathcal{H} \mid D)\, dw\, d\alpha\, d\beta \tag{2.8}$$

If $P(\{x_t\}_{N+1}^T \mid w, \beta_{\mathrm{MP}}, \mathcal{H}) \approx \prod_t \frac{1}{Z_D(\beta_{\mathrm{MP}})}$

$$\times \exp\left\{ -\frac{\beta_{\mathrm{MP}}}{2} \left(x_{t+1} - f(x_t, \ldots, x_{t-\tau+1}; w_{\mathrm{MP}}) - \frac{\partial f}{\partial w}^T (w - w_{\mathrm{MP}})\right)^2 \right\} \tag{2.9}$$

$$P(w \mid D, \alpha, \beta, \mathcal{H}) \approx \frac{1}{(2\pi)^{h/2} \det A^{-1/2}} \times \exp\left(-\frac{1}{2}(w - w_{\mathrm{MP}})^T A (w - w_{\mathrm{MP}})\right) \quad (2.10)$$

then the predictive mean $x_{t,\mathrm{MP}}$ is given by

$$\boxed{x_{t+1,\mathrm{MP}} = f(x_{t,\mathrm{MP}}, \ldots, x_{t-\tau+1,\mathrm{MP}}, w_{\mathrm{MP}}), \quad N \le t \le T-1 . \quad (2.11)}$$

Log marginal likelihood $-2\log P(D \mid \alpha, \beta, \mathcal{H})$ is sometimes called ABIC [2] or evidence for hyperparameters [3], and marginal likelihood at the next hierarchy $P(D \mid \mathcal{H})$ is sometimes called evidence for model [3]. The quantity proposed in [2], $-2\log P(D \mid \alpha, \beta, \mathcal{H}) + 2\dim(\alpha, \beta)$ is different from $-2\log P(D \mid \mathcal{H})$, however.

3. DEMONSTRATIONS

3.1. Chaotic Time Series

Consider the Rössler System

$$\begin{cases} \dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= bx - cz + xz \end{cases} \quad (3.1)$$

with $(a, b, c) = (0.36, 0.4, 4.5)$ (Fig. 3.1).

Consider

$$\begin{cases} \dot{x} &= -y - z + \nu_t^1 \\ \dot{y} &= x + ay + \nu_t^2 \\ \dot{z} &= bx - cz + xz + \nu_t^3 \end{cases} \quad (3.2)$$

where ν_t^1, ν_t^2, ν_t^3 are noise processes. To avoid technical difficulties associated with stochastic process with continuous parameters, let us consider the discrete version of (3.2):

$$\begin{cases} x_{(t+1)\delta} &= f(x_{t\delta}, y_{t\delta}, z_{t\delta}) + \nu_{t\delta}^1 \\ y_{(t+1)\delta} &= g(x_{t\delta}, y_{t\delta}, z_{t\delta}) + \nu_{t\delta}^2 \\ z_{(t+1)\delta} &= h(x_{t\delta}, y_{t\delta}, z_{t\delta}) + \nu_{t\delta}^3 \end{cases} \quad (3.3)$$

where $f(\cdot)$, $g(\cdot)$, $h(\cdot)$ represent a numerical integration scheme , e.g., Runge-Kutta, with step size δ, and $\nu_{t\delta}^i \sim i.i.d.\ N(0, \sigma^2), i = 1, 2, 3$.

Let $\{x_{t\delta}\}_{t\ge 0}$ be the observation. There are two parameters to be estimated. One is the sampling period η, i.e., how often $x_{t\delta}$ should be sampled. Another is the embedding dimension τ(see (1.1)). There are several different algorithms for each of them. One of our main purposes in this paper is to estimate τ so that we assume that η is already estimated. Figure.3.2 shows $(x_{t\delta\eta}, x_{(t-1)\delta\eta}, x_{(t-2)\delta\eta})$ with $\delta = 0.01$, $\eta = 50$, and $\sigma = 0.02$, $t = 0, \ldots, 499$. This data was used as the training data set and the scheme described in the previous section was applied.

Figure.3.3 shows $\log P(D \mid \alpha_{\mathrm{MP}}, \beta_{\mathrm{MP}}, \mathcal{H})$ against (τ, h). The model with the highest marginal likelihood was selected $(\tau = 4, h = 5)$, and used for prediction.

Fig. 3.4 shows $(x_{t\delta\eta,\mathrm{MP}}, x_{(t-1)\delta\eta,\mathrm{MP}}, x_{(t-2)\delta\eta,\mathrm{MP}})$-trajectory, and figure 3.5 shows prediction capability of the learned system where the initial condition was not in the training data. These figures indicate that the present approach may give rise to a new means for inferring embedding dimension of a chaotic attractor when system noise is present.

3.2. Air-conditioning Load Prediction

Saving energy and reduction of CO_2 emissions are becoming critical for conservation of global and regional environments. The cost of electricity during night hours is typically much less than that of the daytime. Therefore, in electrically operated HVAC (Heating, Ventilation, and Air-Conditioning) systems, introduction of thermal energy storage systems can help level off electricity demand throughout the day and thus increase the overall operation efficiency of the power plants run by utility companies. Very good prediction algorithms are needed for predicting air-conditioning loads in order to decide the amount ice to be produced.

"The First International Benchmark Test of Air-conditioning Load Prediction Methods for Optimum Operation of Thermal Energy Storage Systems" was organized by SHASE (Society of Heating, Air-conditioning, and Sanitary Engineers in Japan) [6] which we participated.

Problem B:

Let data set $D := \left(\{x_t\}_{t=0}^N \ , \ \{u_t\}_{t=0}^N\right) \subset \mathbb{R} \times \mathbb{R}^m$ be given, where u_t are the inputs and x_t is the output. Given additional input data $\{u_t\}_{t=N+1}^T$, predict $\{x_t\}_{t=N+1}^T$.

The air-conditioning load prediction problem belongs to Problem B where u_t represent meteological data including temperature, humidity, windspeed, solar flux, and so on, and x_t is the total load at time t.

The details are omitted due to limitation in space. Our result was the first among the seventeen participating groups [6].

4. REFERENCES

[1] Sauer, T., Yorke,J. and M. Casdagli, "Embeddology", J. Stat. Phys. vol. 65, pp.579-616, 1991.

[2] Akaike, H. [1980], "Likelihood and the Bayes procedure", In *Bayesian Statistics*, J.M.Bernardo, M.H. DeGroot, D.V.Lindley and A.F.M.Smith, eds, University Press, Valencia, Spain, 143-166.

[3] MacKay, D.J.C. [1991], "Bayesian Methods for Adaptive Models, PhD thesis", California Inst.Tech. Pasadena 1991.

[4] T.Matsumoto, Y.Nakajima, H.Hamagishi, J.Sugi and M.Saito [1998], "From Data to Nonlinear Dynamics : A Hierarchical Bayes Approach with Neural Nets", *IEEE Workshop on Neural Networks for Signal Processing* VIII, pp.333-342.

[5] Y.Nakajima, J.Sugi, M.Saito, H.Hamagishi, D.Hattori and T.Matsumoto [1998], "Hierarchical Bayesian Neural Nets for Air-conditioning Load Prediction : Nonlinear Dynamics Approach", *Proceedings of the 1998 International Joint Conference on Neural Networks*, 1948-1953.

[6] Society of Heating, Air-conditioning and Sanitary Engineers in Japan (SHASE) [1997], " International Benchmark Test of Air-conditioning Load Prediction Methods for Optimum Operation of Thermal Energy Storage Systems", *http://www.t3.rim.or.jp/~bmtest.*

Level 1

$$P(w \mid D, \alpha, \beta, \mathcal{H}) = \frac{P(D \mid w, \beta, \mathcal{H})\ P(w \mid \alpha, \mathcal{H})}{P(D \mid \alpha, \beta, \mathcal{H})}$$

Level 2

$$P(\alpha, \beta \mid D, \mathcal{H}) = \frac{P(D \mid \alpha, \beta, \mathcal{H})\ P(\alpha, \beta \mid \mathcal{H})}{P(D \mid \mathcal{H})} \quad \Longrightarrow$$

Level 3

$$P(\mathcal{H} \mid D) = \frac{P(D \mid \mathcal{H})\ P(\mathcal{H})}{P(D)}$$

Predictive Distribution

$$P(\{x_t\}_{N+1}^{T} \mid D)$$
$$= \sum_{\mathcal{H}} \iiint P(\{x_t\}_{N+1}^{T} \mid w, \beta, \mathcal{H})$$
$$\times\ P(w, \alpha, \beta, \mathcal{H} \mid D) dw d\alpha d\beta$$

Figure 1.1

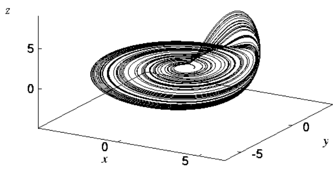

Figure 3.1: Rössler system

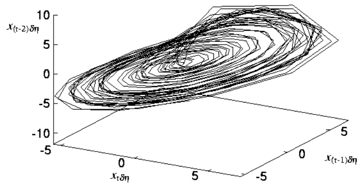

Figure 3.2 Training data: $(x_{t\delta\eta}, x_{(t-1)\delta\eta}, x_{(t-2)\delta\eta})$

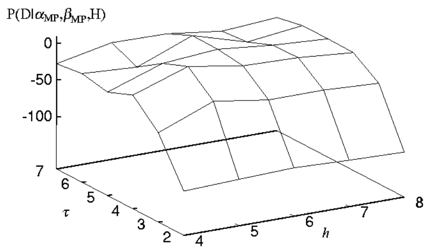

Figure 3.3: $\log P(D \mid \alpha_{\mathrm{MP}}, \beta_{\mathrm{MP}}, \mathcal{H})$ against (τ, h)

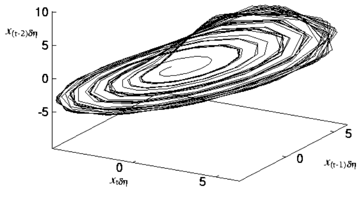

Figure 3.4. Predicted $(x_{t\delta\eta,\mathrm{MP}}, x_{(t-1)\delta\eta,\mathrm{MP}}, x_{(t-2)\delta\eta,\mathrm{MP}})$ trajectory

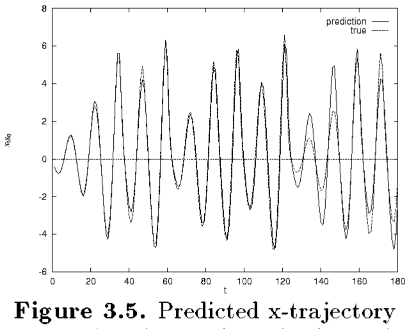

Figure 3.5. Predicted x-trajectory compared with true (noiseless) Rössler trajectory