Asymptotic Properties of Just-in-Time Models

Anders Stenman, Alexander V Nazin and Fredrik Gustafsson

Department of Electrical Engineering
Linköping University, S-581 83 Linköping, Sweden

URL: <http://www.control.isy.liu.se>

Email: {stenman,fredrik}@isy.liu.se

LiTH-ISY-R-1949

18 April 1997

ASYMPTOTIC PROPERTIES OF JUST-IN-TIME MODELS

Anders Stenman * Alexander V. Nazin ** Fredrik Gustafsson *

* *Department of Electrical Engineering, Linköping University,
S-581 83 Linköping, Sweden*

** *Institute of Control Sciences, Profsoyuznaya, 65, 117806 Moscow,
Russia*

Abstract: The concept of Just-in-Time models has been introduced for models that are not estimated until they are really needed. The prediction is taken as a weighted average of neighboring points in the regressor space, such that an optimal bias/variance trade-off is achieved. The asymptotic properties of the method are investigated, and are compared to the corresponding properties of related statistical non-parametric kernel methods. It is shown that the rate of convergence for Just-in-Time models at least is in the same order as traditional kernel estimators, and that better rates probably can be achieved.

Keywords: Non-parametric identification, Nonlinear systems

1. INTRODUCTION

We consider the problem of predicting the output from a non-linear dynamical system

$$y_t = m(\varphi_t) + e_t, \quad (1)$$

where $y_t \in \mathbb{R}$, $\varphi_t \in \mathbb{R}^n$ is a regression vector, and e_t are i.i.d. random variables with zero means and variances λ . Given a new operating point, φ_t , we seek to estimate $m(\varphi_t)$ using a sample set of noisy observations $\{(y_k, \varphi_k)\}_{k=1}^N$, and prior knowledge of the application. The regression vector φ_t can be interpreted as a vector of lagged inputs and outputs in a nonlinear ARX model fashion. However in the analysis to follow, the regressor will be restricted to be scalar-valued.

Traditionally in system identification literature and statistics, the regression problem has been solved by *global* modeling methods, like kernel methods (Wand and Jones, 1995), neural networks or other non-linear parametric models (Sjöberg *et al.*, 1995), but when dealing with very large data sets, this approach becomes less attractive to deal with because of the computational complexity. For real industrial applications, for example in the chemical process industry, the volume of data may occupy several Gigabytes.

The global modeling process is in general associated with an optimization step. This optimization problem is typically non-convex and will have a number of local minima which makes the solution difficult. Although the global model has the appealing feature of giving a high degree of data compression, it seems both inefficient and unnecessary to spend a large amount of calculations to optimize a model which is valid over the whole regressor space, while in most cases it is more likely that we only will visit a very restricted subset of it.

Inspired by ideas and concepts from the database research area, we have taken a conceptually different point of view. We assume that all observations are stored in a database, and that the models are built *dynamically* as the actual need arises. When a model is really needed in a neighborhood of an operating point φ_t , a subset of the data closest to the operating point is retrieved from the database, and a local modeling operation is performed on that subset, see Figure 1. For this concept, we have adopted the name *Just-in-Time* models, suggested by (Cybenko, 1996).

As in the related field of kernel estimation (Härdle, 1990; Wand and Jones, 1995) it is assumed that the Just-in-Time predictor is formed as a weighted average of the output variables in a neighborhood around

Fig. 1. The Just-in-Time concept.

φ_t ,

$$\hat{m}_{\text{JIT}}(\varphi_t) = \sum_{k=-\infty}^{t-1} w_k y_k, \quad (2)$$

where the weights are constructed such that they give measurements located close to φ_t more influence than those located far away from it. The performance is optimized locally by minimizing the pointwise mean square error (MSE),

$$\text{MSE}\{\hat{m}_{\text{JIT}}(\varphi_t)\} = E(\hat{m}_{\text{JIT}}(\varphi_t) - m(\varphi_t))^2, \quad (3)$$

subject to the properties of the weights w_k . It is a well known fact that the MSE can be decomposed into two parts, variance error and squared bias error. The optimal weights are thus selected as a trade-off between these two parts.

To our knowledge, the concept of Just-in-Time models is new in the control community. However, local non-parametric models in the same fashion are well-known in the statistical literature, although there always seems to be a global optimization in some step.

The outline is as follows; A review of related statistical non-parametric methods is given in Section 2. Section 3 describes the Just-in-Time method. Section 4 presents an analysis of the asymptotic properties of Just-in-Time models, and Section 5 finally, describes the problem of Hessian and noise variance estimation.

2. NON-PARAMETRIC ESTIMATION

Local non-parametric regression models have been discussed and analyzed in the statistical literature the last two decades, starting with (Stone, 1977) and (Cleveland, 1979). A special class of such models is *local polynomial kernel estimators*. These estimate the regression function at a particular point φ_t , by “locally” fitting a p th degree polynomial to the data via weighted least squares, where the weights are chosen to follow a *kernel function*, $K : [-1, 1] \rightarrow \mathbb{R}$, satisfying

$$\int_{-1}^1 K(u) du = 1, \quad \int_{-1}^1 uK(u) du = 0, \quad (4)$$

$$K(u) \geq 0, \quad \forall u. \quad (5)$$

One such kernel, is the *Epanechnikov kernel*,

$$K(u) = \begin{cases} 0.75(1 - u^2) & |u| < 1 \\ 0 & |u| > 1 \end{cases} \quad (6)$$

which has been proved to have optimal properties (Epanechnikov, 1969).

A commonly used kernel estimator, corresponding to $p = 0$, is the *Nadaraya-Watson estimator* (Nadaraya, 1964; Watson, 1964)

$$\hat{m}_{\text{NW}}(\varphi_t, h_N) = \frac{\sum_{k=1}^N K\left(\frac{\varphi_k - \varphi_t}{h_N}\right) y_k}{\sum_{k=1}^N K\left(\frac{\varphi_k - \varphi_t}{h_N}\right)}. \quad (7)$$

Here h_N denotes the *bandwidth*, and it can be interpreted as a scaling factor that controls the size of the neighborhood around φ_t .

The bandwidth h_N is usually optimized by minimizing the MSE subject to h_N . However, the MSE formula depends on the bandwidth in a complicated way, which makes it difficult to analyze the influence of the bandwidth on the performance of the kernel estimator. One way to overcome this problem is to use large sample approximations for the bias and variance terms. For the Nadaraya-Watson estimator (7) we have the following asymptotic result in the univariate case.

Proposition 2.1. Consider the fixed design case where $\varphi_i = i/N$ and e_i are identically distributed random variables with zero means and variances λ . Let $\hat{m}_{\text{NW}}(\varphi_t, h_N)$ be a Nadaraya-Watson estimator as in (7), and assume that

- (i) The second order derivative $m''(\varphi_t)$ is continuous on $[0, 1]$.
- (ii) The kernel function K is symmetric about 0, and has support on $[-1, 1]$.
- (iii) The bandwidth $h = h_N$ is a sequence satisfying $h_N \rightarrow 0$ and $Nh_N \rightarrow \infty$ as $N \rightarrow \infty$.
- (iv) The estimation point φ_t is an interior point satisfying $h < \varphi_t < 1 - h$.

Then

$$\begin{aligned} \text{MSE}\{\hat{m}_{\text{NW}}(\varphi_t)\} \\ \sim \underbrace{\left(\frac{m''(\varphi_t)}{2} h_N^2 \mu_2(K)\right)^2}_{\text{bias}^2} + \underbrace{\frac{1}{Nh_N} R(K) \lambda}_{\text{variance}}, \end{aligned} \quad (8)$$

with

$$\mu_2(K) = \int u^2 K(u) du, \quad R(K) = \int K^2(u) du.$$

Furthermore,

$$\inf_{h_N} \text{MSE}\{\hat{m}_{\text{NW}}(\varphi_t)\} \sim C(m''(\varphi_t), \lambda, K) \cdot N^{-4/5}, \quad (9)$$

with optimal bandwidth

$$h_N \sim \left(\frac{R(K) \lambda}{H^2 \mu_2^2(K)}\right)^{1/5} N^{-1/5}. \quad (10)$$

Proof Omitted. See (Wand and Jones, 1995) for details. $\square$

Here $\sim$ denotes “asymptotic equivalence”, that is, $a_n \sim b_n$ if and only if $a_n/b_n \rightarrow 1$ as $N \rightarrow \infty$.

This shows that the best obtainable rate of convergence for kernel estimators satisfying (4) and (5) is of order $N^{-4/5}$, which is slower than the typical rate of order N^{-1} for parametric models in system identification (Ljung, 1987). For the optimal Epanechnikov kernel (6), which is the minimum of (8) w.r.t. $K(\cdot)$, we have that

$$\mu_2(K) = \frac{1}{5}, \quad \text{and} \quad R(K) = \frac{3}{5},$$

which yields that

$$\inf_{h_N} \text{MSE}\{\hat{m}_{\text{NW}}(\varphi_t)\} \sim \frac{3}{4} \left(\frac{(m''(\varphi_t))^2 \lambda^4}{15} \right)^{1/5} N^{-4/5} \quad (11)$$

with asymptotic optimal bandwidth

$$h_N \sim \left(\frac{15\lambda}{(m''(\varphi_t))^2} \right)^{1/5} N^{-1/5}. \quad (12)$$

3. JUST-IN-TIME MODELS

This section gives a brief summary of the Just-in-Time method. More detailed descriptions are given in (Stenman *et al.*, 1996) and in (Stenman, 1997).

It is assumed that all observations $\{(y_i, \varphi_i)\}$ of the system are stored in a database. When a model is needed at φ_t , a subset Ω_M of the data located close to this point is retrieved from the database, and a local modeling operation is performed on the subset, see Figure 2.

Fig. 2. The Just-in-Time idea: A subset Ω_M of the data closest to the operating point φ_t is retrieved, and is used to compute an estimate of $m(\varphi_t)$.

The Just-in-Time estimator is formed as a local linear regression,

$$\hat{m}_{\text{JIT}}(\varphi_t) = \varphi_t^T \hat{\theta} + \hat{\phi}, \quad (13)$$

where $\hat{\theta}$ and $\hat{\phi}$ are the solution to the weighted least squares problem

$$\arg \min_{\theta, \phi} \sum_{\varphi_i \in \Omega_M} w_i (y_i - \varphi_i^T \theta - \phi)^2. \quad (14)$$

Here Ω_M denotes a neighborhood of φ_t containing M data, i.e.,

$$\Omega_M = \{\varphi_i : \|\varphi_i - \varphi_t\| \leq h\}$$

for some arbitrary $h \in \mathbb{R}$ and some suitable norm $\|\cdot\|$. The weight sequence $\{w_i\}$ is assumed to satisfy

$$\sum_{\varphi_i \in \Omega_M} w_i = 1, \quad \sum_{\varphi_i \in \Omega_M} w_i (\varphi_i - \varphi_t) = 0, \quad (15)$$

which are standard for kernel estimation and smoothing windows in spectral analysis (Kay, 1988), and which are consistent with (4).

Given the weight constraints (15), the estimator (13) reduces to a weighted average of the outputs in the neighborhood Ω_M , that is

$$\hat{m}_{\text{JIT}}(\varphi_t) = \sum_{\varphi_i \in \Omega_M} w_i y_i. \quad (16)$$

The optimal weights in (16) is determined by minimizing the mean square prediction error (MSE)

$$\text{MSE}\{\hat{m}_{\text{JIT}}(\varphi_t)\} = E(\hat{m}_{\text{JIT}}(\varphi_t) - m(\varphi_t))^2 \quad (17)$$

subject to the constraints (15). Since this is a quadratic problem with linear constraints, it has an explicit solution for $\{w_i\}$ (see eq. (34) in the proof below). The resulting optimal weights are depending on the noise variance λ , the Hessian $\mathcal{H}_m(\varphi_t)$, and the distance $\varphi_i - \varphi_t$. The Hessian is normally unknown, but by using a Taylor series expansion, it can be estimated from data using least squares theory, as shown in Section 5.

Figure 3 illustrates the weight sequences of the Just-in-Time estimator and the Nadaraya-Watson kernel estimator (7) (using the Epanechnikov kernel), when estimating $m(\varphi) = \sin(2\pi\varphi)$ at $\varphi = 0.27$ using the observations

$$y_i = m(\varphi_i) + e_i, \quad i = 1, \dots, 100$$

where $\varphi_i = i/100$ and $e_i \in N(0, \sqrt{0.05})$. The Just-in-Time weights are represented by the dashed line and the effective Nadaraya-Watson weights by the dashed-dotted line. The corresponding estimates are indicated with a circle and a cross respectively. As indicated, the Just-in-Time estimator gives a slightly better result than the corresponding kernel estimator. This is due to fact that the Just-in-Time weights are optimized locally.

4. ASYMPTOTIC PROPERTIES OF JUST-IN-TIME MODELS

In this section we investigate the asymptotical properties of Just-in-Time models for the univariate (scalar) case. The consistency of (16) and the speed of which the MSE (17) tends to zero as a function of the sample size N , are given in the following proposition. For simplicity and to enable comparison with the kernel estimator result presented in Proposition 2.1, it is stated for the fixed design case, but it can be shown that it also holds for the random design case where the φ_i 's are uniformly distributed on $[0, 1]$.

Fig. 3. A comparison between the Just-in-Time estimator and the Nadaraya-Watson estimator. True regression function (solid), simulated data (dots), Just-in-Time estimate (circle) and Nadaraya-Watson estimate (cross). The weight sequences for the Just-in-Time estimator (dashed) and the Nadaraya-Watson estimator (dash-dotted) are also plotted.

Proposition 4.1. Consider the fixed equally spaced design regression model where $\varphi_i = i/N$ and e_i are i.i.d. random variables with zero mean and variance λ . Assume that:

- (i) The second order derivative $m''(\varphi_t)$ is continuous on $[0, 1]$.
- (ii) The neighborhood Ω_M around φ_t is defined as
$$\Omega_M = \{\varphi_i : |\varphi_i - \varphi_t| \leq h_N\},$$
and contains $M_N = 2h_N N$ data.
- (iii) The neighborhood size parameter $h = h_N$ is a sequence satisfying $h_N \rightarrow 0$ and $Nh_N \rightarrow \infty$ as $N \rightarrow \infty$.
- (iv) The estimation point φ_t is located at the grid and is an interior point of the interval, i.e., $\varphi_t = l/N$ for some integer l satisfying

$$h_N N \leq l \leq (1 - h_N)N.$$

Let $\hat{m}_{\text{JIT}}(\varphi_t)$ denote the Just-in-Time estimate according to (16) with weights satisfying (15).

Then

$$\inf_w \text{MSE}\{\hat{m}_{\text{JIT}}(\varphi_t, w)\} \sim \frac{9\lambda((m''(\varphi_t))^2 M_N^5 + 320\lambda N^4)}{4M_N(720\lambda N^4 + (m''(\varphi_t))^2 M_N^5)}, \quad (18)$$

with optimal weights

$$w_i \sim c \left(1 - d(h_N) \left(\frac{\varphi_i - \varphi_t}{h_N}\right)^2\right), \quad (19)$$

where

$$d(h_N) = \frac{5}{3} \frac{(m''(\varphi_t))^2 h_N^3 N}{(m''(\varphi_t))^2 h_N^3 N + 10\lambda/h_N^2}. \quad (20)$$

Proof Introduce vectors

$$\begin{aligned} e &= (1, \dots, 1)^T, \\ \alpha &= (\alpha_1, \dots, \alpha_{M_N})^T, \\ \beta &= (\beta_1, \dots, \beta_{M_N})^T, \\ w &= (w_1, \dots, w_{M_N})^T, \end{aligned}$$

where

$$\alpha_i = \varphi_i - \varphi_t, \quad \text{and} \quad \beta_i = \frac{1}{2}m''(\varphi_t)\alpha_i^2. \quad (21)$$

The mean square error (17) is then given by

$$\begin{aligned} \text{MSE}\{\hat{m}_{\text{JIT}}(\varphi_t, w)\} &= \left(\sum_{\varphi_i \in \Omega} w_i m(\varphi_i) - m(\varphi_t) \right)^2 + \lambda \sum_{\varphi_i \in \Omega} w_i^2 \\ &\sim (\beta^T w)^2 + \lambda w^T w, \quad (22) \end{aligned}$$

where the similarity follows as a consequence of the constraints (15) and a second order Taylor expansion of $m(\cdot)$ at φ_t . The error made in the Taylor expansion vanishes asymptotically since $h_N \rightarrow 0$.

We now want to minimize the right hand side of (22) subject to the constraints (15). Define a Lagrange function $\mathcal{L}$ as

$$\mathcal{L} = \frac{1}{2} \text{MSE}\{\hat{m}_{\text{JIT}}(\varphi_t, w)\} + \mu(e^T w - 1) + \rho(\alpha^T w). \quad (23)$$

Then

$$\frac{\partial \mathcal{L}}{\partial w_i} = \lambda w_i + (\beta^T w)\beta_i + \mu + \alpha_i \rho = 0, \quad \forall i. \quad (24)$$

Introduce the notation

$$\kappa = \beta^T w, \quad (25)$$

for the bias error term in (22). Hence

$$w_i = -\frac{1}{\lambda}(\kappa\beta_i + \mu + \alpha_i\rho), \quad (26)$$

and we get the equation system

$$\begin{cases} e^T w = -\frac{1}{\lambda}(\kappa\beta^T e + \mu M_N + \rho e^T \alpha) = 1 \\ \alpha^T w = -\frac{1}{\lambda}(\kappa\beta^T \alpha + \mu e^T \alpha + \rho \alpha^T \alpha) = 0 \\ \beta^T w = -\frac{1}{\lambda}(\kappa\beta^T \beta + \mu e^T \beta + \rho \beta^T \alpha) = \kappa \end{cases} \quad (27)$$

for κ and the Lagrange multipliers μ and ρ .

The odd moments of α_i in (27) vanish asymptotically since

$$\begin{aligned} e^T \alpha &= \sum_{\varphi_i \in \Omega} (\varphi_i - \varphi_t) \\ &= O(M_N/N) \rightarrow 0 \quad \text{as } N \rightarrow \infty, \end{aligned} \quad (28)$$

and since

$$\begin{aligned} \beta^T \alpha &= \frac{m''(\varphi_t)}{2} \sum_{\varphi_i \in \Omega} (\varphi_i - \varphi_t)^3 \\ &= O(M_N^3/N^3) \rightarrow 0 \quad \text{as } N \rightarrow \infty. \end{aligned} \quad (29)$$

For the even moments of α_i we have

$$\begin{aligned} e^T \beta &= \frac{m''(\varphi_t)}{2} \sum_{\varphi_i \in \Omega} (\varphi_i - \varphi_t)^2 \\ &= m''(\varphi_t) \sum_{k=1}^{M_N/2} \left(\frac{k}{N}\right)^2 \\ &\sim \frac{m''(\varphi_t) M_N^3}{24N^2}, \end{aligned} \quad (30)$$

$$\alpha^T \alpha = \sum_{\varphi_i \in \Omega} (\varphi_i - \varphi_t)^2 \sim \frac{M_N^3}{12N^2}, \quad (31)$$

and

$$\begin{aligned} \beta^T \beta &= \frac{(m''(\varphi_t))^2}{4} \sum_{\varphi_i \in \Omega} (\varphi_i - \varphi_t)^4 \\ &= \frac{(m''(\varphi_t))^2}{2} \sum_{k=1}^{M_N/2} \left(\frac{k}{N}\right)^4 \\ &\sim \frac{(m''(\varphi_t))^2 M_N^5}{320N^4}. \end{aligned} \quad (32)$$

Hence, when inserting equations (28) to (32), the equation system (27) has the asymptotic solution

$$\begin{cases} \rho = 0 \\ \kappa = \frac{30 m''(\varphi_t) M_N^2 \lambda N^2}{720 \lambda N^4 + (m''(\varphi_t))^2 M_N^5} \\ \mu = -\frac{9 \lambda ((m''(\varphi_t))^2 M_N^5 + 320 \lambda N^4)}{4 M_N (720 \lambda N^4 + (m''(\varphi_t))^2 M_N^5)}. \end{cases} \quad (33)$$

From (26) it follows that

$$\begin{aligned} w &= -\lambda^{-1}(\kappa\beta + \mu e + \rho\alpha) \\ &\sim -\lambda^{-1}(\kappa\beta + \mu e). \end{aligned} \quad (34)$$

The variance error term in (22) is thus given by

$$\lambda w^T w \sim -(\kappa w^T \beta + \mu w^T w) = -(\kappa^2 + \mu). \quad (35)$$

Hence the MSE formula (22) simplifies to

$$\begin{aligned} \inf_w \text{MSE}\{\hat{m}_{\text{JIT}}(\varphi_t, w)\} &= \kappa^2 - (\kappa^2 + \mu) = -\mu \\ &= \frac{9 \lambda ((m''(\varphi_t))^2 M_N^5 + 320 \lambda N^4)}{4 M_N (720 \lambda N^4 + (m''(\varphi_t))^2 M_N^5)}. \end{aligned} \quad (36)$$

and (18) is proved. From (34) we have that

$$\begin{aligned} w_i &\sim -\lambda^{-1}(\kappa\beta_i + \mu) = -\frac{\mu}{\lambda} \left(1 + \frac{\kappa}{\mu}\beta_i\right) \\ &= -\frac{\mu}{\lambda} \left(1 + \frac{\kappa m''(\varphi_t)}{2\mu}(\varphi_i - \varphi_t)^2\right) \\ &= c \left(1 - d(h_N) \left(\frac{\varphi_i - \varphi_t}{h_N}\right)^2\right), \end{aligned} \quad (37)$$

with

$$d(h_N) = \frac{5}{3} \frac{(m''(\varphi_t))^2 h_N^3 N}{(m''(\varphi_t))^2 h_N^3 N + 10\lambda/h_N^2}, \quad (38)$$

and (19) and (20) are proved. $\square$

The MSE formula (18) is a decreasing function of M_N . This can be explained as follows: If $m(\varphi_t)$ is a quadratic function, i.e., the second order derivative $m''(\varphi_t)$ is constant for all $\varphi_t \in [0, 1]$, the Taylor series expansion of $m(\cdot)$ used in (22) will be valid over the entire interval. The optimal neighborhood should thus be chosen as $M_N = N$, i.e., the entire data set. However, the Taylor expansion is in general only valid locally, which requires that $M_N < N$ in order to guarantee that $h_N = \frac{M_N}{2N} \rightarrow 0$ as $N \rightarrow \infty$.

If the neighborhood size is chosen as

$$h_N \sim \left(\frac{15\lambda}{(m''(\varphi_t))^2}\right)^{1/5} N^{-1/5}. \quad (39)$$

which corresponds to $d(h_N) = 1$, exactly the same convergence rate

$$\begin{aligned} \text{MSE}\{\hat{m}_{\text{JIT}}(\varphi_t, w)\} \\ \sim \frac{3}{4} \left(\frac{(m''(\varphi_t))^2 \lambda^4}{15}\right)^{1/5} N^{-4/5}, \end{aligned} \quad (40)$$

as in the kernel estimation case is obtained. This is equivalent to an Epanechnikov type weight sequence with $w_i > 0$.

A larger M_N leads to that $d(h_N) > 1$, i.e., the weight sequence takes both positive and negative values. This will decrease the bias error even more. It is thus expected that the Just-in-Time method has better convergence rate than traditional kernel estimators.

Determining an exact expression for this rate will require that higher order derivatives are taken into account in the analysis.

5. ESTIMATION OF HESSIAN AND THE NOISE VARIANCE

As stated in Section 3, we need to know the Hessian and the noise variance in order to compute the optimal weights. A 2nd order Taylor expansion of $m(\cdot)$ at φ_t yields

$$\begin{aligned} m(\varphi_i) &= m(\varphi_t) + \mathcal{D}_m^T(\varphi_t)(\varphi_i - \varphi_t) + \\ &\quad + \frac{1}{2}(\varphi_i - \varphi_t)^T \mathcal{H}_m(\varphi_t)(\varphi_i - \varphi_t), \end{aligned} \quad (41)$$

where $\mathcal{D}_m(\cdot)$ and $\mathcal{H}_m(\cdot)$ denotes the Jacobian and Hessian of $m(\cdot)$ respectively. Since $\mathcal{D}_m$ and $\mathcal{H}_m$ enter linearly in (41), $\mathcal{H}_m$ can be estimated by least squares theory as

$$V_M(\vartheta, \Omega_M) = \frac{1}{M} \sum_{\varphi_i \in \Omega_M} (y_i - \phi_i^T \vartheta)^2, \quad (42)$$

$$\hat{\vartheta} = \arg \min_{\vartheta} V_M(\vartheta, \Omega_M), \quad (43)$$

where ϑ contains $m(\varphi_t)$ and the entries of $\mathcal{D}_m(\varphi_t)$ and $\mathcal{H}_m(\varphi_t)$, and ϕ_i is the corresponding regression vector. By this approach it is also possible to get an

estimate of the noise variance λ . From (Ljung, 1987) we have

$$\hat{\lambda} = V_M(\hat{\vartheta}, \Omega_M) \frac{1}{1 - d/M}, \quad (44)$$

where $d = \dim \vartheta$.

When estimating the Hessian using the least squares approach (42) and (43), it is clear that a small neighborhood would give the measurement noise a large influence on the resulting estimate $\hat{\vartheta}$. On the other hand, a large neighborhood would make the Taylor expansion (41) inappropriate and would introduce a bias error. A reasonable choice of region size could therefore be obtained as a trade-off between the bias error and the variance error when performing the Hessian estimation.

A commonly used approach in statistics and system identification is to evaluate the loss function (42) on completely new datasets Ω'_M , and choose M_{opt} as the M that minimizes $V_M(\hat{\vartheta}, \Omega'_M)$ (so called *cross-validation*). Adopting this concept to our framework, we get

$$M_{opt} = \arg \min_M V_M(\hat{\vartheta}, \Omega'_M) \quad (45)$$

In this context it would be more desirable to determine M by evaluating $V_M(\cdot)$ on the same data Ω_M as used in estimation, since we do not want to waste measurements without cause. However, this approach yields a problem, since the estimate for small values of M will adapt to the local noise realization. Hence when applying the estimate to the same data as used for estimation, the loss function will become an increasing function in M , i.e., the optimal M will be the smallest one. A number of methods has therefore been developed that penalize the loss function $V_N(\hat{\vartheta}, \Omega_M)$ for small M such that it imitates what we would have obtained if we had applied the evaluation on fresh data. One such method is Akaike's *Final Prediction Error (FPE)* (Akaike, 1969),

$$V_M^{FPE} = V_M(\hat{\vartheta}, \Omega_M) \frac{1 + d/M}{1 - d/M}, \quad (46)$$

where $d = \dim \vartheta = 1 + n + \frac{1}{2}n(n+1)$. We thus have a method of determining the region size M_{opt} as

$$M_{opt} = \arg \min_M V_M(\hat{\vartheta}, \Omega_M) \frac{1 + d/M}{1 - d/M}. \quad (47)$$

Note that this function is minimized w.r.t. M , and not w.r.t. the number of parameters d as usual in the area of model structure selection which is its originally intended application. In (Cleveland, 1979) and (Ruppert *et al.*, 1995), the related Mallows's C_p criterion is used to get a good bias/variance trade-off.

6. CONCLUSIONS

The asymptotical properties of Just-in-Time models has been investigated for the univariate (scalar) case.

It has been shown that we get at least the same rate of convergence $N^{-4/5}$ as for traditional kernel estimators, but it is expected that higher rates can be achieved since the Just-in-Time weights are not restricted to be positive. The assumption of equidistant points φ_i does of course not hold for general dynamical systems, and must be further investigated.

7. REFERENCES

- Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243–247.
- Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.
- Cybenko, G. (1996). Just-in-time learning and estimation. In: *Identification, Adaption, Learning* (S. Bittani and G. Picci, Eds.). NATO ASI series. Springer. pp. 423–434.
- Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications* **14**, 153–158.
- Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.
- Kay, S.M. (1988). *Modern Spectral Estimation*. Prentice-Hall, Englewood Cliffs, N.J.
- Ljung, L. (1987). *System Identification – Theory for the user*. Prentice Hall, Englewood Cliffs, N.J.
- Nadaraya, E. (1964). On estimating regression. *Theory of Probability and Its Applications* **10**, 186–190.
- Ruppert, D., S.J. Sheather and M.P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*.
- Sjöberg, J., Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P.-Y. Glorennec, H. Hjalmarsson and A. Juditsky (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica* **31**, 1691–1724.
- Stenman, A. (1997). Just-in-time models with applications to dynamical systems. Licentiate thesis 601. Dept. of EE, Linköping University. S-581 83 Linköping, Sweden.
- Stenman, A., F. Gustafsson and L. Ljung (1996). Just in time models for dynamical systems. In: *Proceedings of the 35th IEEE Conference on Decision and Control, Kobe, Japan*.
- Stone, C.J. (1977). Consistent nonparametric regression. *The annals of statistics* **5**, 595–620.
- Wand, M.P. and M.C. Jones (1995). *Kernel Smoothing*. Chapman & Hall.
- Watson, G. (1964). Smooth regression analysis. *Sankhyā* **A(26)**, 359–372.