CIRCUITS SYSTEMS SIGNAL PROCESSING
VOL. 21, NO. 1, 2002, PP. 11–21

PREDICTION ERROR ESTIMATION METHODS*

Lennart Ljung[1]

Abstract. This contribution describes a common family of estimation methods for system identification, viz, *prediction-error methods*. The basic ideas behind these methods are described. An overview of typical model structures to which they can be applied is also given, as well as the most fundamental asymptotic properties of the resulting estimates.
Key words: System identification, estimation, prediction, prediction errors, maximum likelihood, convergence, asymptotic covariance, closed loop identification.

1. Basic idea

System identification is about building mathematical models of dynamical systems using measured input-output data. This can be done using a number of different techniques, as evidenced in this special issue. *Prediction-error methods* (PEMs) are a broad family of parameter estimation methods that can be applied to quite arbitrary model parameterizations. These methods have a close kinship with the maximum likelihood method, originating from [4] and introduced into the estimation of dynamical models and time series by [2] and [1].

This article describes the basic properties of PEMs, applied to typical models used for dynamical systems and signals. See [5] or [8] for thorough treatments along the same lines.

Some basic notation is as follows. Let the input and output to the system be denoted by u and y, respectively. The output at time t will be $y(t)$, and similarly for the input. These signals may be vectors of arbitrary (finite) dimension. The case of no input (dim $u = 0$) corresponds to a *time series* or *signal model*. Let $Z^N = \{u(1), y(1), u(2), y(2), \ldots u(N), y(N)\}$ collect all past data up to time N. For the measured data, we always assume that they have been sampled at discrete

* Received January 5, 2001.
[1] Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden. E-mail: ljung@isy.liu.se

time points (here just enumerated for simplicity). However, we may very well deal with continuous-time *models*, anyway.

The basic idea behind the prediction error approach is very simple. Describe the model as a predictor of the next output:

$$\hat{y}_m(t|t-1) = f(Z^{t-1}). \tag{1}$$

Here $\hat{y}_m(t|t-1)$ denotes the one-step ahead prediction of the output, and f is an arbitrary function of past, observed data.

Parameterize the predictor in terms of a finite-dimensional parameter vector θ:

$$\hat{y}(t|\theta) = f(Z^{t-1}, \theta). \tag{2}$$

Some regularity conditions may be imposed on the parameterization, see, e.g., Chapter 4 in [5].

Determine an estimate of θ (denoted $\hat{\theta}_N$) from the model parameterization and the observed data set Z^N, so that the distance between $\hat{y}(1|\theta), \ldots, \hat{y}(N|\theta)$ and $y(1), \ldots, y(N)$ is minimized in a suitable norm.

If the above-mentioned norm is chosen in a particular way to match the assumed probability density functions, the estimate $\hat{\theta}_N$ will coincide with the maximum likelihood estimate.

The PEM has a number of advantages:

- It can be applied to a wide spectrum of model parameterizations (see Section 2).
- It gives models with excellent asymptotic properties, due to its kinship with maximum likelihood (see Sections 4 and 5).
- It can handle systems that operate in closed loop (the input is partly determined as output feedback, when the data are collected) without any special tricks and techniques (see Section 4).

It also has some drawbacks:

- It requires an explicit parameterization of the model. To estimate, say, an arbitrary linear, fifth-order model, some kind of parameterization, covering all fifth-order models, must be introduced.
- The search for the parameters that gives the best output prediction fit may be laborious and involve search surfaces that have many local minima.

2. Model parameterizations

The general predictor model is given by (2):

$$\hat{y}(t|\theta) = f(Z^{t-1}, \theta).$$

To give a concrete example, the underlying model could be a simple linear difference equation:

$$y(t) + a_1 y(t-1) + \cdots + a_n y(t-n) = b_1 u(t-1) + \cdots + b_m u(t-m). \quad (3)$$

Ignoring any noise contribution to this equation, or assuming that such a noise term would be unpredictable, the natural predictor becomes

$$\hat{y}(t|\theta) = -a_1 y(t-1) - \cdots - a_n y(t-n) + b_1 u(t-1) + \cdots + b_m u(t-m) \quad (4)$$

$$\theta = \begin{bmatrix} a_1 & \dots & a_n & b_1 \dots & b_m \end{bmatrix}^T, \quad (5)$$

which corresponds to

$$f(Z^{t-1}, \theta) = \theta^T \varphi(t) \quad (6)$$

$$\varphi(t) = \begin{bmatrix} -y(t-1) & \dots & -y(t-n) & u(t-1) & \dots & u(t-m) \end{bmatrix}^T. \quad (7)$$

It is natural to distinguish some specific characteristics of (2):

- *Linear time-invariant (LTI) models.* $f(Z^{t-1}, \theta)$ linear in Z^{t-1}, and not depending explicitly on time, which means that we can write

$$f(Z^{t-1}, \theta) = W_y(q, \theta) y(t) + W_u(q, \theta) u(t) \quad (8)$$

$$= \sum_{k=1}^{t-1} w_y(k) y(t-k) + \sum_{k=1}^{t-1} w_u(k) u(t-k) \quad (9)$$

  for some LTI filters W_y and W_u that both start with a delay. Here, q is the shift operator.
- *Linear regression models.* $f(Z^{t-1}, \theta)$ linear in θ, but possibly nonlinear in Z. Clearly (3) is both a linear model and a linear regression model.
- *Nonlinear models.* $f(Z^{t-1}, \theta)$ is nonlinear in Z.

We shall comment on these cases more in the following sections.

2.1. *Linear models*

The linear predictor model (8) is equivalent to the assumption that the data have been generated according to

$$y(t) = G(q, \theta) u(t) + H(q, \theta) e(t), \quad (10)$$

where e is white noise (unpredictable), and H is monic (that is, its expansion in q^{-1} starts with the identity matrix). We also assume that G contains a delay. The equivalence can be seen by rewriting (10) as

$$y(t) = [I - H^{-1}(q, \theta)] y(t) + H^{-1}(q, \theta) G(q, \theta) u(t) + e(t).$$

The first term on the right-hand side only contains $y(t-k)$, $k \geq 1$, so the natural predictor of $y(t)$, based on past data, will be given by (8) with

$$W_y(q,\theta) = [I - H^{-1}(q,\theta)], \qquad W_u(q,\theta) = H^{-1}(q,\theta)G(q,\theta). \tag{11}$$

It will be required that θ are constrained to values such that the filters $H^{-1}G$ and H^{-1} are stable. Note that the parameterization of G and H is otherwise quite arbitrary. For example, it could be based on a continuous-time, state-space model with known and unknown physical parameters in the matrix entries:

$$\dot{x}(t) = A(\theta)x(t) + B(\theta)u(t) \tag{12}$$

$$y(t) = Cx(t) + v(t). \tag{13}$$

Here the states x may have physical interpretations, e.g., positions and velocities, and θ corresponds to unknown material constants, etc. Sampling this model and then converting it to input-output form gives a model of the type (10), where G depends on θ in a well-defined (but possibly complicated) way.

2.1.1. Linear black-box models.

Sometimes we are faced with systems or subsystems that cannot be modeled based on physical insights. The reason may be that the function of the system or its construction is unknown, or that it would be too complicated to sort out the physical relationships. It is then possible to use standard models, which by experience are known to be able to handle a wide range of different system dynamics.

A very natural approach is to describe G and H in (10) as rational transfer functions in the shift (delay) operator with unknown numerator and denominator polynomials.

We would then have

$$G(q,\theta) = \frac{B(q)}{F(q)} = \frac{b_1 q^{-nk} + b_2 q^{-nk-1} + \cdots + b_{nb} q^{-nk-nb+1}}{1 + f_1 q^{-1} + \cdots + f_{nf} q^{-nf}}. \tag{14}$$

Then

$$\eta(t) = G(q,\theta)u(t) \tag{15}$$

is a shorthand notation for the relationship

$$\begin{aligned} \eta(t) + f_1\eta(t-1) + \cdots + f_{nf}\eta(t-nf) \\ = b_1 u(t-nk) + \cdots + b_{nb} u(t-(nb+nk-1)). \end{aligned} \tag{16}$$

Here, there is a time delay of nk samples.

In the same way, the disturbance transfer function can be written as

$$H(q,\theta) = \frac{C(q)}{D(q)} = \frac{1 + c_1 q^{-1} + \cdots + c_{nc} q^{-nc}}{1 + d_1 q^{-1} + \cdots + d_{nd} q^{-nd}}. \tag{17}$$

The parameter vector θ thus contains the coefficients b_i, c_i, d_i, and f_i of the transfer functions. This model is thus described by five structural parameters: nb, nc, nd, nf, and nk and is known as the *Box-Jenkins (BJ) model.*

An important special case is when the properties of the disturbance signals are not modeled, and the noise model $H(q)$ is chosen to be $H(q) \equiv 1$; that is, $nc = nd = 0$. This special case is known as an *output error (OE) model* because the noise source $e(t) = v(t)$ will then be the difference (error) between the actual output and the noise-free output.

A common variant is to use the same denominator for G and H:

$$F(q) = D(q) = A(q) = 1 + a_1 q^{-1} + \cdots + a_{na} q^{-na}. \tag{18}$$

Multiplying both sides of (14)–(17) by $A(q)$ then gives

$$A(q)y(t) = B(q)u(t) + C(q)e(t). \tag{19}$$

This model is known as the *ARMAX model.* The name is derived from the fact that $A(q)y(t)$ represents an AutoRegression and $C(q)e(t)$ a Moving Average of white noise, while $B(q)u(t)$ represents an eXtra input (or with econometric terminology, an eXogenous variable).

The special case $C(q) = 1$ gives the much-used *ARX model* (3).

2.2. *Nonlinear models*

There is clearly a wide variety of nonlinear models. One possibility that allows inclusion of detailed physical prior information is to build, nonlinear state space-models, analogous to (12). Another possibility, sometime called "semiphysical modeling" is to come up with new inputs, formed by nonlinear transformations of the original, measured u and y, and then deal with models, linear in these new inputs. A third possibility is to construct black-box models by general function expansions.

2.2.1. *Nonlinear black-box models*

The mapping f can be parameterized as a function expansion,

$$f(Z^{t-1}, \theta) = \sum_{k=1}^{d} \alpha_k \kappa(\beta_k(\varphi(t) - \gamma_k)), \qquad \varphi(t) = \varphi(Z^{t-1}). \tag{20}$$

Here, φ is an arbitrary function of past data. However, in the most common case, φ is given by (7). Moreover, κ is a "mother basis function," from which the actual functions in the function expansion are created by *dilation* (parameter β) and *translation* (parameter γ). For example, with $\kappa = \cos$ we would get a Fourier series expansion with β as frequency and γ as phase. More common are cases where κ is a unit pulse. With that choice, (20) can describe any piecewise constant

function, where the granularity of the approximation is governed by the dilation parameter β. A related choice is a soft version of a unit pulse, such as the Gaussian bell. Alternatively, κ could be a unit step (which also gives piecewise constant functions), or a soft step, such as the sigmoid.

Typically, κ is in all cases a function of a scalar variable. When φ is a column vector, the interpretation of the argument of κ can be made in different ways:

- If β is a row vector, $\beta(\varphi - \gamma)$ is a scalar, so the term in question is constant along a hyperplane. This is called the *ridge* approach, and is typical for sigmoidal neural networks.
- Interpreting the argument as $\|\varphi - \gamma\|_\beta$ as a quadratic norm with the positive semidefinite matrix β as a quadratic form gives terms that are constant on spheres (in the β norm) around γ. This is called the *radial* approach. Radial basis neural networks are common examples of this.
- Letting κ be interpreted as the product of κ-functions applied to each of the components of φ gives yet another approach, known as the *tensor* approach. The functions used in (neuro-)fuzzy modeling are typical examples of this approach.

See [5, Chapter 5] or [7] for more details on this interpretation of basis functions.

3. Estimation techniques

Once the model structure, i.e., the parameterized function $f(Z^t, \theta)$ has been defined, and a data set Z^N has been collected, the estimation of the parameter θ is conceptually simple: Minimize the distance between the predicted outputs (according to parameter θ) and the measured outputs,

$$\hat{\theta}_N = \arg\min_\theta V_N(\theta) \tag{21}$$

$$V_N(\theta) = \sum_{t=1}^{N} \ell(y(t) - f(Z^{t-1}, \theta)). \tag{22}$$

Here ℓ is a suitable distance measure, such as $\ell(\varepsilon) = \|\varepsilon\|^2$. The connection to the celebrated *maximum likelihood method* is obtained by a particular choice of norm: Assume that the data are produced by the mechanism

$$y(t) = f(Z^{t-1}, \theta) + e(t), \tag{23}$$

where $\{e(t)\}$ is a sequence of independent random variables with probability density function $p(x)$. Then, with $\ell(x) = -\log p(x)$, the criterion (22) is the negative logarithm of the likelihood function for the estimation problem (apart from θ-independent terms). This makes $\hat{\theta}_N$ equal to the maximum likelihood estimate.

3.1. *Numerical issues*

The actual calculation of the minimizing argument can be complicated, with substantial computations, and possibly a complex search over a function with several local minima. The numerical search is typically carried out using the *damped Gauss-Newton* method. For a scalar output and $\ell(\varepsilon) = \frac{1}{2}\varepsilon^2$, this takes the form

$$\begin{aligned}
\hat{\theta}^{(i+1)} &= \hat{\theta}^{(i)} - \mu_i R_i^{-1} \hat{g}_i \\
\hat{g}_i &= V_N'(\hat{\theta}^{(i)}) \\
V_N'(\theta) &= \frac{dV_N(\theta)}{d\theta} = -\frac{1}{N}\sum_{t=1}^{N}(y(t) - \hat{y}(t|\theta))\psi(t,\theta); \\
\psi(t,\theta) &= \frac{\partial}{\partial\theta}\hat{y}(t|\theta) \\
R_i &= V_N''(\hat{\theta}^{(i)}) \approx \frac{1}{N}\sum_{t=1}^{N}\psi(t,\hat{\theta}^{(i)})\psi^T(t,\hat{\theta}^{(i)}).
\end{aligned} \tag{24}$$

Here μ_i is a scalar, adjusted so that the criterion $V_N(\theta^{(i+1)}) < V_N(\theta^{(i)})$.

A thorough discussion of numerical issues of this minimization problem is given in [5, Chapter 10] and in [3].

4. Convergence properties

An essential question is: What will be the properties of the estimate resulting from (21)? These will naturally depend on the properties of the data record Z^N. It is in general a difficult problem to characterize the quality of $\hat{\theta}_N$ exactly. One normally has to be content with the asymptotic properties of $\hat{\theta}_N$ as the number of data, N, tends to infinity.

It is an important aspect of the general identification method (21) that the asymptotic properties of the resulting estimate can be expressed in general terms for arbitrary model parameterizations.

The first basic result is the following one:

$$\hat{\theta}_N \to \theta^* \quad \text{as} \quad N \to \infty, \tag{25}$$

where

$$\theta^* = \arg\min_{\theta} E\ell(\varepsilon(t,\theta)). \tag{26}$$

That is, as more and more data become available, the estimate converges to that value θ^* that would minimize the expected value of the "norm" of the prediction errors. This is in a sense *the best possible approximation* of the true system that is available within the model structure. The expectation E in (26) is taken

with respect to all random disturbances that affect the data and it also includes averaging over the input properties. This means, in particular, that θ^* will make $\hat{y}(t|\theta^*)$ a good approximation of $y(t)$ with respect to those aspects of the system that are enhanced by the input signal used.

The characterization of the limiting estimate can be more precise in the case of a linear model structure. We distinguish between the cases of open- and closed-loop data and in the remainder of this section will assume that the system is single-input, single-output.

4.1. Linear systems: Open-loop data

Suppose that the data actually have been generated by

$$y(t) = G_0(q)u(t) + v(t), \tag{27}$$

where u and v are independent. This means that the input u has been generated in open loop, i.e., independently of y. Let $\Phi_u(\omega)$ be the input spectrum and $\Phi_v(\omega)$ be the spectrum of the additive disturbance v. Then the prediction error can be written

$$\begin{aligned}\varepsilon_F(t,\theta) &= \frac{1}{H(q,\theta)}[y(t) - G(q,\theta)u(t)] \\ &= \frac{1}{H(q,\theta)}[(G_0(q) - G(q,\theta))u(t) + v(t)].\end{aligned} \tag{28}$$

By Parseval's relation, the prediction-error variance can also be written as an integral over the spectrum of the prediction error. This spectrum, in turn, is directly obtained from (28), so the limit estimate θ^* in (26) can also be defined as

$$\begin{aligned}\theta^* = \arg\min_{\theta} \Bigg[&\int_{-\pi}^{\pi} |G_0 - G_\theta|^2 \frac{\Phi_u(\omega)}{|H_\theta|^2}\, d\omega \\ &+ \int_{-\pi}^{\pi} \Phi_v(\omega)/|H_\theta|^2\, d\omega \Bigg].\end{aligned} \tag{29}$$

For brevity, here we used $G_\theta = G(e^{i\omega}, \theta)$, etc.

If the noise model $H(q,\theta) = H_*(q)$ does not depend on θ (as in the OE model), expression (29) thus shows that the resulting model $G(e^{i\omega}, \theta^*)$ will give that frequency function in the model set that is closest to the true one, in a quadratic frequency norm with weighting function

$$Q(\omega) = \Phi_u(\omega)/|H_*(e^{i\omega})|^2. \tag{30}$$

This shows clearly that the fit can be affected by the input spectrum Φ_u and the noise model H_*.

4.2. *Linear systems: Closed-loop data*

Assume now that the data has been generated from (27), but the input has been partly determined by output feedback, e.g., as

$$u(t) = r(t) - F_y(q)y(t). \tag{31}$$

Moreover, the noise is supposed to be described by

$$v(t) = H_0(q)e(t), \tag{32}$$

where e is white noise with variance λ. The reference (set point) signal r is supposed to be independent of the noise e. Using this fact, together with Parseval's relation as above, gives the following result:

$$\theta^* = \arg\min_{\theta} \int_{-\pi}^{\pi} [|G_0 + B_\theta - G_\theta|^2 \Phi_u + |H_0 - H_\theta|^2 \Phi_e^r]/|H_\theta|^2 \, \mathrm{d}\omega, \tag{33}$$

where

$$B_\theta = (H_0 - H_\theta)\bar{\Phi}_{ue}/\Phi_u \tag{34}$$

$$\Phi_e^r = \lambda - |\Phi_{eu}|^2/\Phi_u \tag{35}$$

Here Φ_{ue} is the cross spectrum between e and u, which in the case of (31)–(32) will be

$$\Phi_{ue}(\omega) = -\lambda \frac{F_y(e^{i\omega})H_0(e^{i\omega})}{1 + F_y(e^{i\omega})G_0(e^{i\omega})}. \tag{36}$$

The result (33) contains important information:

- If there exists a θ_0 such that $H_0 = H_{\theta_0}$ and $G_0 = G_{\theta_0}$, then this value is always a possible convergence point. If $\Phi_e^r > 0 \, \forall \omega$ (which, according to (36) means that u cannot be determined entirely from e by linear filtering), then this is the only possible convergence point.
- If H_θ cannot achieve the value H_0 (e.g., if H_θ is fixed as in an output error model), and $\Phi_{ue} \neq 0$, then there is a bias pull B_θ away from the true transfer function G_0. It is consequently necessary that the noise model also can be correctly described in the model structure in order to obtain an unbiased transfer function estimate in the case of closed-loop data.

However, the main conclusion is that the PEM, applied in a straightforward fashion, paying no attention to possible feedback effects, will provide unbiased estimates whenever the true system is contained in the model set. The only requirement is that the input u should not be formed from e only by linear time-invariant filtering.

5. Asymptotic distribution

Once the convergence issue has been settled, the next question is: How fast is the limit approached? This is dealt with by considering the asymptotic distribution of the estimate. The basic result is the following one: If $\{\varepsilon(t, \theta^*)\}$ is approximately white noise, then the random vector $\sqrt{N}(\hat{\theta}_N - \theta^*)$ converges in distribution to the normal distribution with zero mean, and the covariance matrix of $\hat{\theta}_N$ is approximately given by

$$P_\theta = \lambda[E\psi(t)\psi^T(t)]^{-1}, \tag{37}$$

where

$$\begin{aligned} \lambda &= E\varepsilon^2(t, \theta^*) \\ \psi(t) &= \frac{d}{d\theta}\hat{y}(t|\theta)|_{\theta=\theta^*}. \end{aligned} \tag{38}$$

This means that the convergence rate of $\hat{\theta}_N$ towards θ^* is $1/\sqrt{N}$. Think of ψ as the sensitivity derivative of the predictor with respect to the parameters. It is also used in the actual numerical search algorithm (24). Then (37) says that the covariance matrix for $\hat{\theta}_N$ is proportional to the inverse of the covariance matrix of this sensitivity derivative. This is a quite natural result.

The result (37), (38) is general and holds for all model structures, both linear and nonlinear, subject only to some regularity and smoothness conditions. These results are also fairly natural and provide the guidelines for all user choices involved in the process of identification. Of particular importance is that the asymptotic covariance matrix (37) equals the Cramér-Rao lower bound if the disturbances are Gaussian. That is, PEM, give the optimal asymptotic properties. See [5] for more details.

6. Use of prediction error methods (PEMs)

The family of PEMs has the advantage of being applicable to a wide variety of model structures. It also handles closed-loop data in a direct fashion and gives the best possible results (minimal covariance matrix), provided the model structure contains the true system. The approximation properties when the true system cannot be achieved in the model structure are also well understood.

Several software packages that implement these techniques are available, e.g., [6], and many successful applications have been reported.

The main drawback of the PEMs is that the numerical search in (24) may be laborious and require good initial parameter values. For multivariable, linear black-box state space models it is therefore very useful to combine the use of PEMs with what are called *subspace methods*, (e.g., [9]).

References

[1] K. J. Åström and T. Bohlin, Numerical identification of linear dynamic systems from normal operating records, in *IFAC Symposium on Self-Adaptive Systems*, Teddington, England, 1965.

[2] G. E. P. Box and D. R. Jenkins, *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco, CA, 1970.

[3] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[4] R. A. Fisher, On an absolute criterion for fitting frequency curves, *Mess. Math.*, 41:155, 1912.

[5] L. Ljung, *System Identification—Theory for the User*, Prentice-Hall, Upper Saddle River, NJ, 2nd ed., 1999.

[6] L. Ljung, *System Identification Toolbox for use with* MATLAB, *Version 5*, 5th ed., The MathWorks, Inc, Natick, MA, 2000.

[7] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P. Y. Glorennec, H. Hjalmarsson, and A. Juditsky, Nonlinear black-box modeling in system identification: A unified overview, *Automatica*, 31(12):1691–1724, 1995.

[8] T. Söderström and P. Stoica, *System Identification*, Prentice-Hall Int., London, 1989.

[9] P. van Overschee and B. de Moor, *Subspace Identification of Linear Systems: Theory, Implementation, Applications*, Kluwer Academic Publishers, Dordrecht, 1996.