ON-LINE ORDER SELECTION FOR COMMUNICATIONS

Hongmei Ni and Tülay Adalı

Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County, Baltimore, MD 21250

ABSTRACT

We address the problem of on-line order determination for communications and show that penalized partial likelihood criterion provides a suitable likelihood framework for the problem by allowing correlations among samples and on-line processing ability. An on-line, efficient order selection scheme is developed assuming that the observations can be modeled by a finite normal mixture model without imposing any additional conditions on the unknown system, such as linearity. Channel equalization by finite normal mixtures is considered as an example for which correct order determination is critical and examples are presented to show the application and effectiveness of the approach.

1. INTRODUCTION

On-line order selection is a difficult and important problem for real-time applications which has not been well studied. A truly on-line order estimation scheme which updates the order estimate as new samples arrive is highly desirable for real time communications as it can save training time, increase information transmission rate, and reduce the storage requirement and computational cost significantly. Very few publications in statistical estimation theory has dealt with this subject. In [4], a sequential Bayes learning and model selection approach is proposed and applied to radial basis function (RBF) networks. However, it is computationally intensive and the performance depends on selection of diffusion parameters. In [7], a minimal resource allocation network algorithm is used to grow and prune the RBF network's hidden neurons on-line. Again, selection of training parameters and thresholds is very critical to the performance.

Information theoretic criteria determine an optimal model order for a parameterized model such that a suitable criterion is minimized (or maximized). Use of information theoretic criteria [3, 11, 12, 13] for model selection eliminates the need for subjective judgment on the selection of threshold levels and hence has been very popular. The two most widely used information criteria are the Akaike's information criterion (AIC) [3] and the minimum description length (MDL) [11, 12]. Both criteria can be regarded as penalized maximum likelihood (ML) criteria and assume independent and identically distributed (i.i.d.) samples in their derivations, not a realistic assumption for most practical applications as correlations among samples typically do exist, which is also the case in most communications and signal processing applications. Moreover, on-line implementation of AIC or MDL is a difficult problem that has not been particularly addressed.

Partial likelihood (PL) [5, 14] allows for inclusion of dependent observations and sequential processing in a likelihood framework, hence it allows development of order selection schemes for real time signal processing using information theoretic criteria. In [10], we derived penalized partial likelihood (PPL) as the information theoretic criterion for order selection and proposed a sequential order selection scheme which increases the order estimate gradually. However, the procedure uses all past data samples to calculate the maximum PL values, increasing the storage requirement which is not desirable in real-time implementations. In this paper, we develop a new formulation of PPL that eliminates this need and hence is suitable for on-line implementation and show its successful application.

2. PL FORMULATION FOR FNM MODELS

Given a time series $\{x_n\}$, $n = 1, 2, \cdots$, and its time-dependent covariates $\{\mathbf{y}_n\}$, when the objective is to estimate the distribution of $\mathbf{y}_n$ given all the available information upto time n, we can define $\mathcal{F}_{n-1} = \sigma\{1, [x_n, \cdots, x_1], [\mathbf{y}_{n-1}, \cdots, \mathbf{y}_1]\}$ as the σ-field generated by all relevant events upto time n and choose a suitable probability model with parameter $\boldsymbol{\theta}$ to model the conditional distribution of $\mathbf{y}_n$ given $\mathcal{F}_{n-1}$, $p_\theta(\mathbf{y}_n|\mathcal{F}_{n-1})$. Then by a factorization of the likelihood [14], we can write the PL function relative to $\boldsymbol{\theta}$, $\mathcal{F}_n$ and the data sequence $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$ as

$$\mathcal{L}_N^p(\boldsymbol{\theta}) = \prod_{n=1}^{N} p_\theta(\mathbf{y}_n|\mathcal{F}_{n-1}). \qquad (1)$$

Research supported in part by the National Science Foundation Career Award, NSF NCR–9703161.

Note that the formulation above does not require the i.i.d. assumption that is typically invoked to write the likelihood in the product form. Also, the formulation does not assume conditioning on future samples which might be required for some conditional likelihood (CL) formulations again for characterization in the product form. These two properties in particular render PL particularly suitable for developing likelihood approaches for real-time signal processing [1]. However it is important to note that it is a generalization of ML and might coincide with ML or CL for special cases [1, 2]. Also, large sample optimality properties of PL such as consistency and asymptotic normality can be established under mild regularity conditions for the general case of dependent observations [1], which allows adaptive-structure and robust classifier designs by using modified likelihood functions and information-theoretic criteria. Next, we give the PL formulation for density estimation by the finite normal mixtures (FNM) model.

Let the d-dimensional observation vector $\mathbf{y}_n$ be written as $\mathbf{y}_n = [y_n, y_{n-1}, \cdots, y_{n-d+1}]^T$ where y_n is the system output at time n. For a channel of memory L, we have $y_n = f(x_n, x_{n-1}, \cdots, x_{n-L}) + \eta_n$ where $f(\cdot)$ is a linear or nonlinear mapping and η_n is the additive white system noise. Obviously the distribution of $\mathbf{y}_n$ is only dependent on $\mathbf{x}_n = [x_n, x_{n-1}, \cdots, x_{n-L-d+1}]^T$. We can thus write $p_\theta(\mathbf{y}_n|\mathcal{F}_{n-1}) = p_\theta(\mathbf{y}_n|\mathbf{x}_n)$. Then PL function becomes $\mathcal{L}^p_N(\boldsymbol{\theta}) = \prod_{n=1}^N p_\theta(\mathbf{y}_n|\mathbf{x}_n)$.

When x_n takes a value from a finite alphabet of size M, we can map $\mathbf{x}_n$ to a discrete variable z_n which takes values $\{1, \cdots, K\}$ where $K = M^{L+d}$. When η_n is normally distributed with variance σ^2, we have

$$p_\theta(\mathbf{y}_n|z_n) = \prod_{k=1}^{K} \left(\phi_k(\mathbf{y}_n,\ \boldsymbol{\theta})\right)^{\delta_k(z_n)} \qquad (2)$$

where $\delta_k(z_n) = 1$ when $z_n = k$ and 0 otherwise, and

$$\phi_k(\mathbf{y}_n,\ \boldsymbol{\theta}) = \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\parallel \mathbf{y}_n - \mathbf{m}_k \parallel^2}{2\sigma^2}\right) \qquad (3)$$

where $\mathbf{m}_k$ is the noiseless channel observation vector when $z_n = k$ and $\|\cdot\|$ denotes the Euclidean distance. This is a special case of the FNM model with discrete variable z_n in [9] as here we impose the constraint that the mixtures have a common covariance matrix $\Sigma = \sigma^2\mathrm{I}$, a condition typically satisfied for the channel equalization example that we consider.

Maximum PL estimation of the FNM model parameter vector $\boldsymbol{\theta} = [\mathbf{m}_1^T, \cdots, \mathbf{m}_K^T, \sigma^2]^T$ is shown to be very efficient, when the FNM model provides a good match to the data generation mechanism (see e.g. [9]). This is the case in channel equalization where the distribution of the output of a multipath channel corrupted by additive noise is a perfect match to the FNM model. Correct order determination for the FNM model, however, is very important as the efficiency of the algorithm is primarily due to the match of the model with the inherent structure of the data. This is possible only if the number of mixtures is correctly determined to be M^{L+d}, i.e., only if the channel order L is correctly estimated. In the next section, we introduce an on-line channel order selection procedure that uses penalized partial likelihood as the information theoretic criterion.

3. ON-LINE ORDER SELECTION BY PENALIZED PARTIAL LIKELIHOOD

We use penalized partial likelihood criterion for order selection in real-time processing [10]:

$$\mathrm{PPL}(i) = \ln \mathcal{L}^p_N(\boldsymbol{\theta}^*) - K_i \frac{\ln N}{2} \qquad (4)$$

where $\boldsymbol{\theta}^*$ is the maximum PL (MPL) estimate of the model parameter and K_i is the number of independently adjusted parameters for the ith model. The optimal model order is the one which maximizes the PPL criterion.

In [10], we followed Schwarz's approach [12] to derive the PPL criterion, while Rissanen [11] arrived at the same expression, the MDL criterion, from a totally different viewpoint, reformulating the problem as an information coding problem. The main difference of PPL criterion is that in its derivation, partial likelihood is used and hence it provides a more general formulation that allows for dependent observations.

In this section, we study on-line implementation of order detection using the PPL criterion given in (4). The implementation of equation (4) requires the use of all past data samples to calculate the maximum PL value, a situation undesirable in real-time implementations. Here, we obtain a new formulation of PPL that eliminates this need. Using the FNM model given in (2), we write

$$\begin{aligned}
&\ln \mathcal{L}^p_N(\boldsymbol{\theta}) \\
&= \ln \prod_{n=1}^{N} p(\mathbf{y}_n|\mathbf{x}_n) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \delta_k(z_n) \ln \phi_k(\mathbf{y}_n,\ \boldsymbol{\theta}) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \delta_k(z_n) \left(-\frac{d}{2}\ln(2\pi\sigma^2) - \frac{\parallel \mathbf{y}_n - \mathbf{m}_k \parallel^2}{2\sigma^2}\right) \\
&= -\frac{Nd}{2}\ln(2\pi\sigma^2) \\
&\quad - \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} \delta_k(z_n) \parallel \mathbf{y}_n - \mathbf{m}_k \parallel^2}{2\sigma^2} \qquad (5)
\end{aligned}$$

Maximizing (5) with respect to $\boldsymbol{\theta}$, i.e., solving the equations

$$\begin{aligned} \frac{\partial \ln \mathcal{L}_N^p(\boldsymbol{\theta})}{\partial \mathbf{m}_k} &= 0, \;\; k = 1, \cdots, K \\ \frac{\partial \ln \mathcal{L}_N^p(\boldsymbol{\theta})}{\partial \sigma^2} &= 0 \end{aligned} \qquad (6)$$

gives

$$\begin{aligned} \mathbf{m}_k^* &= \frac{\sum_{n=1}^N \delta_k(z_n)\mathbf{y}_n}{\sum_{n=1}^N \delta_k(z_n)}, \;\; k = 1, \cdots, K, \\ \sigma^{*2} &= \frac{1}{Nd}\sum_{n=1}^N \sum_{k=1}^K \delta_k(z_n) \| \, \mathbf{y}_n - \mathbf{m}_k^* \, \|^2 \end{aligned} \qquad (7)$$

Expectation-maximization (EM) algorithm [6] and its extensions have been widely used to compute the maximum likelihood parameter estimates by first obtaining current parameter estimates (E-step) and then updating these by generalized mean ergodic theorems (M-step). Since, we consider supervised learning, $\delta_k(z_n)$ is known, and hence the observations can be split into K classes depending on the value of z_n and the parameters of each class can be estimated separately. For the FNM model given in (2), the MPL parameter estimates are the corresponding sample mean vectors and a pooled sample covariance that combines the sample covariances for the K classes, as given in (7).

We can derive an on-line version of the estimates given in equation (7), by initializing the parameters, i.e., setting $\boldsymbol{\theta} = \mathbf{0}$ at $t = 0$, and performing the updates:

$$\begin{aligned} c_k^{(t)} &= c_k^{(t-1)} + \delta_k(z_t) \;\; k = 1, \cdots, K, \\ \mathbf{m}_k^{(t)} &= \mathbf{m}_k^{(t-1)} + \frac{\delta_k(z_t)}{c_k^{(t)}}(\mathbf{y}_t - \mathbf{m}_k^{(t-1)}) \;\; k = 1, \cdots, K, \\ \sigma^{2(t)} &= \sigma^{2(t-1)} + \frac{1}{td}\Big[\| \, \mathbf{y}_t \, \|^2 - d\,\sigma^{2(t-1)} \\ &\quad - \sum_{k=1}^K \Big(c_k^{(t)} \| \, \mathbf{m}_k^{(t)} \, \|^2 - c_k^{(t-1)} \| \, \mathbf{m}_k^{(t-1)} \, \|^2 \Big) \Big] \end{aligned} \qquad (8)$$

where $c_k^{(t)}$ is the counter for class $z_n = k$ at time t and is also initialized to 0.

Hence, the updates in (8) satisfy equation (7) for any N. In on-line implementation, it is desirable to keep the sample size N as small as possible while making sure that it is large enough to ensure desirable large sample properties of MPL estimation. In a practical implementation, starting with an initially small sample size and letting N increase proportionally with the number of model parameters can ensure reasonable parameter convergence and system performance, which has been verified by our simulation results. We include some of those results in the next section and discuss the effect of sample size in [10]. For the FNM model in (2), the number of parameters for channel order i is $K_i = d\,M^{i+d} + 1$. Thus the data sample size should increase exponentially with the order estimate and the practical rule of having samples 10–20 times the number of free parameters (K_i in this case) provides satisfactory performance.

We can write the PPL criterion for on-line estimation using equations (7) and (5) as follows:

$$\ln \mathcal{L}_N^p(\boldsymbol{\theta}^*) \;\; = \;\; -\frac{Nd}{2}\ln(2\pi\sigma^{*2}) - \frac{Nd}{2} \qquad (9)$$

Using (9), dividing (4) by N, excluding the term constant with respect to i, and finally using indexed N_i instead of N to indicate the fact that the sample size varies with i in our on-line parameter updates, we obtain the on-line version of PPL (OPPL) criterion as:

$$\text{OPPL}(i) = -\frac{d}{2}\ln(2\pi\sigma^{*2}) - K_i\frac{\ln N_i}{2N_i} \qquad (10)$$

Note that only the MPL parameter estimates are used to evaluate equation (10), and the previous samples do not need to enter the computation. Hence order detection by OPPL criterion can be implemented on-line.

For on-line implementation, we start with $d = 1$ for efficiency as explained in [10] and an initial order estimate, estimate the MPL parameters on-line, calculate the value of (10) after the parameters converge and check if (10) is maximized. If not, increase the order estimate and repeat the above procedure until the optimal order is found. The on-line channel order detection scheme we propose can be summarized as:

1. Initialize $L_e = 1$.
2. Using (8), on-line update the model parameters for channel order estimate L_e (and $L_e - 1$ if this step is executed the first time).
3. After the parameter estimates converge, calculate OPPL for L_e using (10).
4. If OPPL$(L_e) <$ OPPL$(L_e - 1)$, the channel order is estimated as $L_e - 1$. Stop.
   Else, OPPL$(L_e - 1) \leftarrow$ OPPL(L_e), $L_e \leftarrow L_e + 1$, and go to step 2.

4. SIMULATION RESULTS

We studied the properties of the OPPL criterion at different signal-to-noise ratio (SNR) levels for a number of channels. The OPPL criterion yielded satisfactory performance on a variety of channels. OPPL curves for two of these channels in shown in Figures 1 and 2 where the number of data samples for order L_e is chosen as $N_{L_e} = 40 \cdot 2^{L_e} + 20$ in these cases. We can observe that, for the channel in Figure 1, the OPPL criterion gives the correct channel order $L = 2$ at all tested SNR levels. For the channel in Figure 2, the correct order $L = 5$ is determined at high

SNRs, and there is slight underestimation at low SNRs. It is also worth noting the effect of noise as a function of the lowest multipath component. When the noise is comparable to the power of the lowest multipath component, ignoring this component does not significantly degrade the BER performance, as expected. We also tested the online order selection introduced for nonlinear channels as well as channels that are minimum phase and non-minimum phase. The correct channel order is obtained for most channels. For example, at all SNR levels, our scheme gives $L_e = 1$ for the linear channel $y_n = x_n + 0.5x_{n-1} + \eta_n$ within 100 samples, $L_e = 2$ for the nonlinear channel $y_n = y_{nl} - 0.2y_{nl}^2 + \eta_n$ where $y_{nl} = x_n + 0.5x_{n-1} + 0.3x_{n-2}$, within 300 samples, and $L_e = 2$ for the non-minimum phase nonlinear channel $y_n = y_{nl} - 0.2y_{nl}^2 + \eta_n$ where $y_{nl} = 0.3482x_n + 0.8704x_{n-1} + 0.3482x_{n-2}$, within 300 samples. For a channel with longer memory $y_n = x_n + 0.5x_{n-1} + 0.4x_{n-2} + 0.3x_{n-3} + 0.2x_{n-4} + 0.1x_{n-5} + \eta_n$, our scheme gives $L_e = 5$ within 3100 samples when SNR is greater than or equal to 10 dB, and $L_e = 4$ within 1500 samples when SNR is less than 10 dB. The number of samples used is increase exponentially with the channel order L as discussed in the previous section.

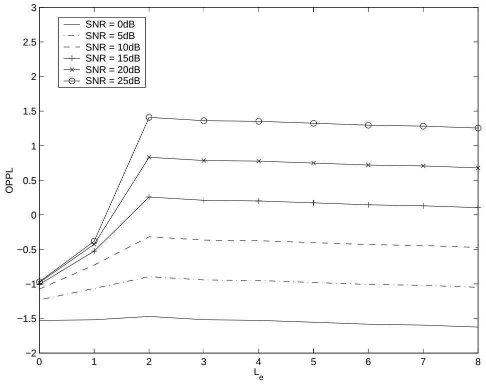

Fig. 1. OPPL curves for the channel $H(z) = 1 + 0.5z^{-1} + 0.3z^{-2}$

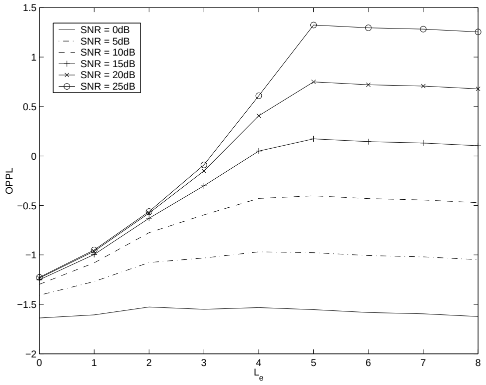

Fig. 2. OPPL curves for the channel $H(z) = 1 + 0.5z^{-1} + 0.4z^{-2} + 0.3z^{-3} + 0.2z^{-4} + 0.1z^{-5}$

Another point to note is that with the increase of the system memory L, the complexity of the FNM model and the number of required samples increase substantially, and the use of a posterior type modeling using logistic regression type models [1, 2] can be more attractive for these cases. An efficient subspace approach for estimation of effective channel order is given in [8] and is noted to be more robust to variations in SNR and number of data samples compared to the information theoretic criteria. The inherent assumption in the subspace decomposition approach is the linearity of the channel which is relaxed in the application of the FNM equalizer we introduced. The tradeoffs involved between these two approaches needs further study.

5. REFERENCES

[1] T. Adalı, X. Liu, and M. K. Sönmez, “Conditional distribution learning with neural networks and its application to channel equalization,” *IEEE Trans. Signal Processing*, vol. 45, no. 4, pp. 1051-1064, Apr. 1997.

[2] T. Adalı, H. Ni, and B. Wang, “Partial likelihood for estimation of multi-class posterior probabilities,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Phoenix, AZ, March 1999, vol. 2, pp. 1053-1056.

[3] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716-723, Dec. 1974.

[4] C. Andrieu and N.D. Freitas, “ Sequential monte carlo for model selection and estimation of neural networks, ” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Istanbul, Turkey, Jun. 2000.

[5] D.R. Cox, “Partial likelihood,” *Biometrika*, vol. 62, pp. 69-72, 1975.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Royal Statistical Soc., Ser. B*, vol. 39, no.1, pp. 1-28, 1977.

[7] P. C. Kumar, P. Saratchandran, and N. Sundararajan, “Communication channel equalization using minimal radial basis function neural networks,” *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, Cambridge, England, Sep. 1998, pp. 477-485.

[8] A. P. Liavas, P. A. Regalia, and J. P. Delmas, “Blind channel approximation: effective channel order determination,” *IEEE Trans. Signal Processing*, vol. 47, no. 12, pp. 3336-3344, Dec. 1999.

[9] H. Ni, T. Adalı, B. Wang, and X. Liu, “A general probabilistic formulation for supervised neural classifiers,” *Journal of VLSI Signal Processing Systems*, vol. 26, nos. 1/2, pp. 141-153, Aug. 2000.

[10] H. Ni and T. Adalı, “Sequential order selection for real time signal processing,” to appear *Proc. IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, Sydney, Australia, Dec. 2000.

[11] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, pp. 465-471, 1978.

[12] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, No. 2, pp. 461-464, 1978.

[13] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 387-392, Apr. 1985.

[14] W. H. Wong, “Theory of partial likelihood,” *Ann. Statist.*, 14, pp. 88-123, 1986.