Conditional NML Universal Models

Jorma Rissanen and Teemu Roos
Complex Systems Computation Group, Helsinki Institute for Information Technology
University of Helsinki and Helsinki University of Technology
Emails: jorma.rissanen@mdl-research.org, teemu.roos@cs.helsinki.fi

I. Introduction

The NML (Normalized Maximum Likelihood) universal model has certain minmax optimal properties but it has two shortcomings: the normalizing coefficient can be evaluated in a closed form only for special model classes, and it does not define a random process so that it cannot be used for prediction. We present a universal *conditional* NML model, which has minmax optimal properties similar to those of the regular NML model [9], [8], [1]. However, unlike NML, the conditional NML model defines a random process which can be used for prediction. It also admits a recursive evaluation for data compression. The conditional normalizing coefficient is much easier to evaluate, for instance, for tree machines than the integral of the square root of the Fisher information in the NML model. For Bernoulli distributions, the conditional NML model gives a predictive probability, which behaves like the Krichevsky-Trofimov predictive probability [3], [9], actually slightly better for extremely skewed strings. For some model classes, it agrees with the predictive probability found earlier by Takimoto and Warmuth, [10], as the solution to a different more restrictive minmax problem.

We also calculate the CNML models for the generalized Gaussian regression models, and in particular for the cases where the loss function is quadratic, and show that the CNML model achieves asymptotic optimality in terms of the mean ideal code length. Moreover, the quadratic loss, which represents fitting errors as noise rather than prediction errors, can be shown to be smaller than what can be achieved with the NML as well as with the so-called plug-in or the predictive MDL model.

II. Two minmax problems

Consider the model class $\mathcal{M}_k = \{f(x^n;\theta)\}$, $\theta = \theta_1,\dots,\theta_k$, and data sequences $x^n = x_1,\dots,n$, for $n = 1,2,\dots$. Let m be the smallest number t for which the ML estimate $\hat\theta_t = \hat\theta(x^t)$ can be computed. Actually, by letting k vary the number m could be reduced, but for the sake of simplicity we keep it fixed. The number

$$\log 1/f(x^n;\hat\theta_n)$$

has been considered as the ideal target for the code length obtainable with the model class, [1], which, however, is not attainable, because $f(x^n;\hat\theta_n)$ is not a probability distribution. This leads to the minmax problem

$$\min_q \max_{x^n} \log \frac{f(x^n;\hat\theta_n)}{q(x^n)},$$

with the solution due to Shtarkov, [9],

$$\begin{aligned} \hat f_{NML}(x^n;\mathcal{M}_\gamma) &= \frac{f(x^n;\hat\theta(x^n))}{C_n} \\ C_n &= \int f(y^n;\hat\theta(y^n))dy^n. \end{aligned} \tag{1}$$

This has been generalized to general parametric model classes in [7] to provide a universal *Normalized Maximum Likelihood*, NML, model with excellent properties. However, the normalizing coefficient can be evaluated easily only for restricted model classes, and the model does not define a random process. This means that it cannot be used for prediction and its evaluation for data compression is difficult.

Given a sequence of integers $t_0 = m+1 < t_1 < \dots, < t_s = n$ consider

$$L(x^n;\hat\theta_{t_0},\dots,\hat\theta_{t_s}) = \log 1/f(x^m;\hat\theta_m) + \sum_{j=0}^{s-1}\sum_{t=t_j}^{t_{j+1}-1} \log 1/f(x_{t+1}|x^t;\hat\theta_{t_{j+1}})$$

as the ideal target for the code length obtainable with the model class. This in general provides a shorter target for the attainable code length than the previous one, and in fact gives a larger likelihood than the traditional 'maximum likelihood'. The maximizing family of ML estimates $\{\hat\theta_t\}$ is obtained for $t_0 = m+1, t_1 = m+2,\dots,n$, or that the maximum likelihood is actually given by

$$f(x^n) = f(x^m;\hat\theta_m)\prod_{t=m+1}^{n} f(x_t|x^{t-1};\hat\theta_t). \tag{2}$$

This suggests the following minmax problem. For all $t > m$

$$\min_{q(x|x^{t-1})} \max_x \log \frac{f(x^{t-1},x;\hat\theta(x^{t-1},x))}{q(x|x^{t-1})}. \tag{3}$$

The solution is given by the *conditional NML* models

$$\begin{aligned} \hat f(x_t|x^{t-1}) &= \frac{f(x^t;\hat\theta(x^t)}{K_t} \\ K_t &= \int f(x^{t-1},x;\hat\theta(x^{t-1},x))dx. \end{aligned} \tag{4}$$

This is proved the same way as the solution to Shtarkov's problem: First, replacing the numerator by the density function (4) does not change the solution, and the maximized ratio of the two density functions (4) and $q(x|x^{t-1})$, which is not smaller than unity, is made unity when the latter is selected

equal to the former. We mention that there is another maxmin problem in terms of the mean code length with the same solution, [8], namely

$$\max_g \min_q E_g \log \frac{f(X|x^{t-1};\hat{\theta}(x^{t-1},X))}{q(X|x^{t-1})},$$

where the expectation is taken with respect to $g = g(x|x^{t-1})$ ranging over all distributions. The maxmin value equals the minmax value. Finally, these minmax–maxmin problems also hold unconditionally.

It is clear that the normalizing coefficient K_t, which in general is a function of x^{t-1}, is easier to calculate, at least numerically, than the normalizing coefficient in the NML universal model.

III. Markov Models

We begin with the Bernoulli class $\mathcal{B} = \{P(x;p)\}$, where the parameter $p = P(1)$. The ML estimate is given by $\hat{p}(x^n) = n_1/n$, where $n_1 = \sum_t x_t$ is the number of 1's in x^n. If $n_0 = n - n_1$ the maximized likelihood is

$$P(x^n; n_1/n) = \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}.$$

The conditional NML predictive probability can be written as

$$\hat{P}(1|x^n) = \frac{(n_1+1)\,e(n_1)}{(n_0+1)\,e(n_0)+(n_1+1)\,e(n_1)}, \qquad (5)$$

where $e(n_0) = (1+1/n_0)^{n_0}$ and $e(n_1) = (1+1/n_1)^{n_1}$; take $e(k) = 1$ for $k = 0$.

The same conditional probability function $\hat{P}(1|x^n)$ was found in [9], where it was shown to converge to the Krichevsky-Trofimov predictive probability

$$P_{KT}(1|x^n) = \frac{n_1 + 1/2}{n+1}.$$

It was also found later in [10], in effect, as the solution to the following minmax problem

$$\min_\theta \max_x \log \frac{f(x^{t-1},x;\hat{\theta}(x^{t-1},x))}{f(x|x^{t-1};\theta)}. \qquad (6)$$

This type of minmax problem is much harder to solve than the minmax problem (3), and the authors' derivation is quite complicated. Furthermore, the solution requires boundedness restrictions on the data x^n, even for the exponential family of models studied in the cited reference, unless the data are bounded as in the Bernoulli case. Since in the Bernoulli case the solution to the wider problem (3) lies in the same Bernoulli family it clearly has to coincide with the solution to (6).

Neither Krichevsky-Trofimov predictive probability nor the related Laplace probability,

$$P_L(1|x^n) = \frac{n_1+1}{n+2},$$

has been shown to have any particular optimality property. Takimoto and Warmuth [10] showed that for the Bernoulli models, the regret of the CNML model (4) satisfies for all sequences the inequality

$$R(\hat{f},x^n) := \ln 1/\hat{f}(x^n) - \ln 1/f(x^n;\hat{\theta}(x^n)) \le \frac{1}{2}\ln(n+1) + \frac{1}{2},$$

and that the worst case sequence is when the string of length 2n has n-1 ones, or, in effect, the random string.

For data compression the performance in the worst case sequence is less important than the per symbol code length as a function of the symbols' occurrence counts. The common performance index is the regret, which, however, taken alone gives a misleading picture of the performance of a code because its relevance depends on the per symbol code length. The CNML probabilities are not determined by the symbols' occurrence counts only, and the analysis appears to be difficult. Instead we calculate in Figure 1 its worst case regret as well as the per symbol code length for strings of length 30, and also show for the sake of comparison the well known analytically computed results of three other models, the Laplace and Krichevsky-Trofimov predictors as well as the NML universal model (1). We see clearly, that all the models give about equal per symbol code length, except for strings where the ratio of the count of symbol one to the length of the strings is close to zero or one. These are precisely the strings where significant compression can be obtained, and we see that the CNML code gives the best compression for them – even better than the Krichevsky-Trofimov predictor. We also see that although the Laplace predictor has by far the smallest regret for other strings, its significance is minor.

Shtarkov also gave the conditional CNML probabilities for Markov classes of models. For the sake of completeness and the reason that they solve the minmax problem (3) we rederive them for binary Markov models and tree machines. Since the Markov class does not belong to the exponential family the techniques given in [10] to solve the narrower problem (6) will not work. However, the solution to the wider problem happens to remain in the Markov class, and the same solution solves also the narrower minmax problem.

Consider a Markov model, either of a fixed or variable order, defined by a tree machine with state space $S = \{s\}$. The states are sequences of binary strings and the state transitions are defined as follows $s \mapsto (s, x_t)$, where (s, x_t) is the longest suffix of the concatenate of the string s and the symbol x_t that falls in S. For instance if $S = 0, 01, 11$, then $(01, 0) = 0$, and $(01, 1) = 11$, and so on. The model is defined by the states, the state transitions, and the binary probabilities $\theta = \{P(0|s), s \in S\}$ at the states. Hence, given an initial state s_0 and its probability $P(s_0)$, which we set to unity for it cancels in the following formulas, the probability of the string x^n is given by

$$P(x^n;\theta) = \prod_t P(x_t|s(x^{t-1})),$$

where the state $s(x^t)$ is the longest suffix of the string $x^t = x_1,\ldots,x_t$ that falls in S.

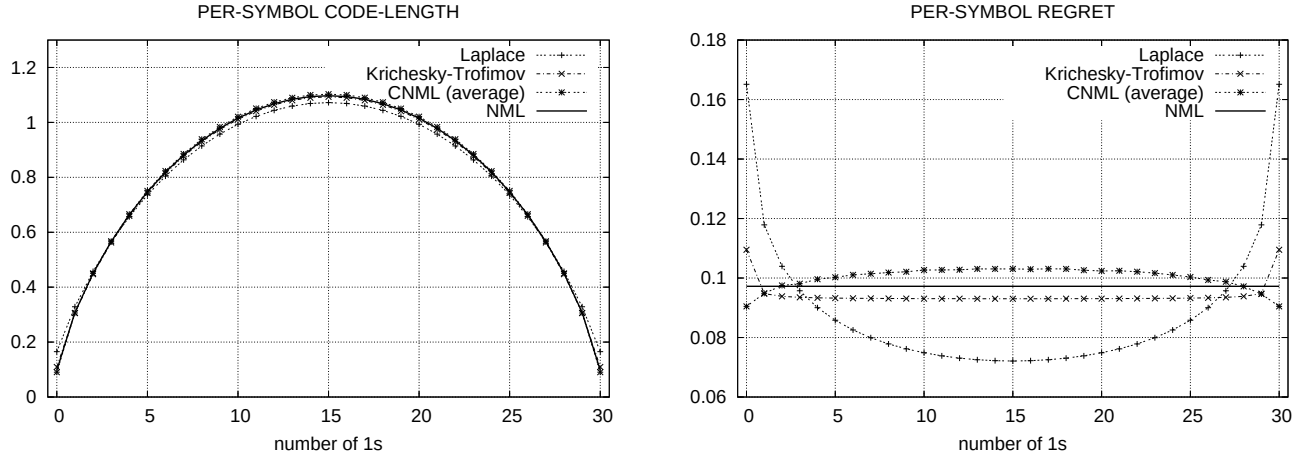

Fig. 1. The per-symbol code-length and regret for four universal models in the Bernoulli case.

The maximized likelihood is given by

$$P(x^n;\hat{\theta}(x^n)) = \prod_s \frac{n_s}{n} \prod_i \left(\frac{n_{i|s}}{n_s}\right)^{n_{i|s}}, \qquad (7)$$

where $n_{i|s} = n_{i|s}(x^n)$ denotes the number of times the sequence si occurs in x^n and $n_s = \sum_i n_{i|s}$. Also, $n = \sum_s n_s$. Putting $s = s(x^n)$ we then obtain with straightforward calculations the conditional

$$\hat{P}(1|x^n) = \frac{(n_{0|s}+1)e(n_{0|s})}{(n_{0|s}+1)e(n_{0|s}) + (n_{1|s}+1)e(n_{1|s})}.$$

We see that this generalizes the Bernoulli case in a natural way. The generalization to non-binary alphabets is straightforward as shown in [9].

IV. Generalized Gaussian Family

We consider the family of regression models

$$f(y^n|X_n;\lambda) = Z_\lambda^{-n} e^{-\lambda \sum_{t=1}^n |y_t - \hat{y}_t|^\alpha}, \qquad (8)$$

where $y^n = y_1,\ldots,y_n$ are real-valued data, $X_n = [\bar{x}_1,\ldots,\bar{x}_n]$ is the $k \times n$ regressor matrix of columns $\bar{x}_t$, $\hat{y}_t = F(\bar{x}_t,\eta)$ a regression function with a k-component vector parameter η, and λ and α positive parameters, the latter kept constant. The normalizing coefficient is given by the n'th power of

$$Z_\lambda = \frac{2}{\alpha}\lambda^{-1/\alpha}\Gamma(1/\alpha), \qquad (9)$$

and it is seen not to depend on η.

The maximum likelihood value of λ is given by

$$\hat{\lambda}_n = \frac{n}{\alpha \sum_{t=1}^n |y_t - \hat{y}_t|^\alpha}, \qquad (10)$$

which depends on all the past and the present values of $\hat{y}_t$. Let $\hat{\eta}_t = \hat{\eta}(y^t, X_t)$ denote the ML estimate of the parameter η; i.e. one that minimizes the sum

$$\sum_{i=1}^t |y_i - F(\bar{x}_i,\eta)|^\alpha.$$

The maximized likelihood is given by its negative logarithm for $t > m$ as

$$\ln 1/f(y^t|X_t;\hat{\lambda}_t) = \frac{t}{\alpha}\ln(e/\hat{\lambda}_t) + t\ln\frac{2\Gamma(1/\alpha)}{\alpha}, \qquad (11)$$

where $\hat{\lambda}_t = \hat{\lambda}(\hat{\eta}_t,\ldots,\hat{\eta}_m)$ depends on all the past ML estimates. Regard $\hat{y}_t = F(\bar{x}_t,\hat{\eta}_t)$ as a function of y_t, given the other variables. With $\hat{e}_i = y_i - \hat{y}_i$ put

$$\hat{s}_m = \sum_{i=1}^m |y_i - F(\bar{x}_t,\hat{\eta}_m)|^\alpha \qquad (12)$$

$$\hat{s}_t = \sum_{i=m+1}^t |\hat{e}_i|^\alpha = \hat{s}_{t-1} + |\hat{e}_t|^\alpha. \qquad (13)$$

By (3) define the conditional density functions for $t > m$

$$\hat{f}(y_t|y^{t-1}) = \frac{1}{K_t}\left(1 + \frac{|y_t - \hat{y}_t|^\alpha}{\hat{s}_{t-1}}\right)^{-t/\alpha} \qquad (14)$$

$$K_t = \int \left(1 + \frac{1}{\hat{s}_{t-1}}|y_t - \hat{y}_t|^\alpha\right)^{-t/\alpha} dy_t,$$

where $m+1$ is the smallest value of t for which $\hat{\theta}(y^t, X_t)$ is defined. Notice that $\hat{y}_t$ for $t > m$ depends on y_t through the estimate $\hat{\eta}_t = \hat{\eta}(y^t|X_t)$, which makes $\hat{e}_t$ a fitting error called for in the minmax problem (3) rather than a prediction error. Given an initial density function $q(y^m)$ we get the density function

$$\hat{f}(y^n|X_n) = \prod_{t=m+1}^n \hat{f}(y_t|y^{t-1})q(y^m). \qquad (15)$$

We are mainly interested in the Gaussian case $\alpha = 2$ and the absolute value case, $\alpha = 1$, where the normalizing integrals can be evaluated in a closed form.

A. Gaussian family

We consider the linear-quadratic regression problem, where the data y^n, X_n are modeled as follows

$$y_t = b'\bar{x}_t + \epsilon_t = \sum_{i=1}^k b_i x_{t,i} + \epsilon_t, \qquad (16)$$

$\{\epsilon_t\}$ being an iid sequence from a normal distribution of zero mean and variance σ^2. The regressor matrix X_t consists either of fixed numbers, not given by y^n, or as in AR models it is given by the columns of $\bar{x}_t = col\{y_{t-1}, \ldots, y_{t-k}\}$. Consider the representation of the data

$$y_t = b_t' \bar{x}_t + \hat{e}_t = \sum_{i=1}^{k} b_{t,i} x_{t,i} + \hat{e}_t, \tag{17}$$

where the ML estimates, written now as row vectors $b_i' = b_{i,1}, \ldots, b_{i,k}$, are given by

$$\begin{aligned} b_t &= \hat{\theta}(y^t|X_t) = V_t \sum_{j=1}^{t} \bar{x}_j y_j \qquad (18)\\ V_t &= (X_t X_t')^{-1} \end{aligned}$$

the prime ' indicating the transpose. For the sake of comparison consider also the representations

$$\begin{aligned} y_t &= b_{t-1}' \bar{x}_t + e_t \qquad (19)\\ y_t &= b_n' \bar{x}_t + \hat{\epsilon}_t(n). \qquad (20) \end{aligned}$$

The predictor $\bar{x}_t' b_{t-1}$ of y_t is sometimes called the 'plug-in' predictor, because the parameters b of the process are replaced by the ML estimates from the latest past data, not including y_t. The resulting model (19) is widely studied, [2], [6], [4], [11], and it defines the linear quadratic PMDL (Predictive MDL) model or the Least Squares model if σ is kept fixed.

Write in (8) $\lambda = 1/(2\sigma^2)$, which gives the maximized likelihood $(2\pi e \hat{\sigma}_t^2)^{-t/2}$, where

$$\hat{\sigma}_t^2 = (1/t) \sum_{i=1}^{t} (y_i - \bar{x}_i' b_t)^2. \tag{21}$$

The conditional density function for $t > m$ is by (14)

$$\begin{aligned} \hat{f}(y_t|y^{t-1}) &= \frac{1}{K_t} \left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{s}_{t-1}}\right)^{-t/2}\\ \hat{s}_t &= \sum_{i=1}^{t} (y_i - \hat{y}_i)^2\\ K_t &= \int_{-\infty}^{\infty} \left(1 + \frac{(y_t - \hat{y}_t)^2}{\hat{s}_{t-1}}\right)^{-t/2} dy_t. \end{aligned}$$

To get the normalizing integral we write first

$$\begin{aligned} \hat{y}_t &= \bar{x}_t' b_t = d_t y_t + \bar{y}_t \qquad (22)\\ d_t &= \bar{x}_t' V_t \bar{x}_t \qquad (23)\\ \bar{y}_t &= \bar{x}_t' V_t \sum_{i=1}^{t-1} y_i \bar{x}_i, \qquad (24) \end{aligned}$$

where $\bar{y}_t$ does not depend on y_t. Then

$$K_t = \int_{-\infty}^{\infty} \left[1 + \frac{(1-d_t)^2}{\hat{s}_{t-1}} \left(y - \frac{\hat{y}_t}{1-d_t}\right)^2\right]^{-t/2} dy.$$

By change of variables

$$z = [y - \hat{y}_t/(1-d_t)](1-d_t)/\sqrt{\hat{s}_{t-1}}$$

the integral becomes

$$\begin{aligned} K_t &= \frac{\sqrt{\hat{s}_{t-1}}}{1-d_t} \int_{-\infty}^{\infty} (1+z^2)^{-t/2} dz\\ &= \frac{\sqrt{\hat{s}_{t-1}}}{1-d_t} \sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)/\Gamma(n/2), \qquad (25) \end{aligned}$$

the second equality by the fact that z is seen to have Student's z-distribution.

The conditional density function is then given by

$$\hat{f}(y_t|y^{t-1}) = \frac{1}{K_t} \left[1 + \frac{(1-d_t)^2}{\hat{s}_{t-1}} \left(y - \frac{\bar{y}_t}{1-d_t}\right)^2\right]^{-t/2}. \tag{26}$$

With a density function $q(y^m|X_m)$ for the initial data, which we do not pick here, we get

$$\hat{f}(y^n|X_n) = q(y^m|X_m) \prod_{t=m+1}^{n} \hat{f}(y_t|y^{t-1})/K_t. \tag{27}$$

We give without proof the asymptotic mean ideal code length for the case where the data are generated by (16), and the regressor variables $\bar{x}_t$ are nonrandom satisfying

$$\frac{1}{n} \sum_{1}^{n} \bar{x}_i \bar{x}_i' \to \Sigma, \tag{28}$$

the limit being a positive definite matrix. For all positive δ and all large enough n

$$\frac{1}{n} E \ln 1/\hat{f}(y^n|X_n) \le \frac{1}{2} \ln \sigma^2 + \frac{k+\delta}{2n} \ln n. \tag{29}$$

Further, under the assumption (28) even for a random regressor matrix

$$\begin{aligned} \sum_{t=m+1}^{n} \hat{e}_t^2 &= \sum_{t=m+1}^{n} e_t^2 (1-d_t)^2 \qquad (30)\\ \sum_{t=m+1}^{n} e_t^2 (1-d_t) &= \sum_{t=m+1}^{n} \hat{\epsilon}_t^2(n) \qquad (31)\\ \sum_{t=m+1}^{n} \hat{e}_t^2 &< \sum_{t=m+1}^{n} \hat{\epsilon}_t^2(n) < \sum_{t=m+1}^{n} e_t^2, \qquad (32) \end{aligned}$$

where $\hat{\epsilon}_t(n) = y_t - \bar{x}_t' b_n$. Moreover, when the regressor matrix is constant

$$\begin{aligned} \sum_{t=m+1}^{n} E\hat{e}_t^2 &= \sigma^2 \left((n-m) - \sum_{t=m+1}^{n} (1-d_t)\right) \qquad (33)\\ \sum_{t=m+1}^{n} E e_t^2 &= \sigma^2 \left((n-m) + \sum_{t=m+1}^{n} 1/(1-d_t)\right) \qquad (34) \end{aligned}$$

$$\sum_{t=m+1}^{n} E\hat{\epsilon}_t^2(n) = \sigma^2 (n-m). \tag{35}$$

We see that the fitting errors are the smallest under the representation (17), which defines the CNML model, not only for the worst case sequence or in the mean but for all sequences. However, only the representation (19) and its

prediction errors define a code length for the data, and if we add the necessary code lengths to the CNML and NML fitting errors representing noise we get code lengths, and the resulting probabilities of course will have to intersect. Finally, Also the CNML model defines a predictor for the data, which agrees with that obtained with the representation (19).

B. Laplace distribution

The second important case is the absolute value loss function, $\alpha = 1$. The main difficulty in the applications is that the ML estimates are difficult to obtain. For this reason, one may settle for linear estimates, which we do as well.

The conditional density functions for $t > m$ are given by

$$\hat{f}(y_t|y^{t-1}) = \frac{1}{K_t}\left(1 + \frac{(1-d_t)|y_t - \bar{y}_t/(1-d_t)|}{\hat{s}_{t-1}}\right)^{-t},$$

where $\hat{s}_t = \sum_{i=1}^{t} |y_i - \hat{y}_i|$. The normalizing constant becomes

$$K_t = 2\int_0^\infty \left(1 + \frac{1-d_t}{\hat{s}_{t-1}}|y_t - \bar{y}_t/(1-d_t)|\right)^{-t} dy_t$$
$$= 2\frac{\hat{s}_{t-1}/(t-1)}{1-d_t}(1 - \bar{y}_t/s_{t-1})^{1-t},$$

where we changed the variables

$$u = 1 + (1-d_t)|y_t - \bar{y}_t/(1-d_t)|/\hat{s}_{t-1}.$$

Again $m+1$ is the smallest value of t for which $\hat{\theta}(y^t, X_t) = b_t$ is defined.

We get further

$$\hat{f}(y^n|X_n) = 2^{m-n} q(y^m|X_m)\hat{s}_n^{-n}\hat{s}_m^m \prod_{t=m+1}^{n} \frac{1-d_t}{t-1},$$

for some $q(y^m|X_m)$ to be chosen. We get then

$$\hat{f}(y^n|X_n)$$
$$= 2^{-(n-m)} q(y^m|X_m)\hat{s}_n^{-n} s_m^m \prod_{t=m+1}^{n} (1-d_t)(t-1)\hat{s}_{t-1}^{t-1}.$$

ACKNOWLEDGMENTS

This work was supported in part by the Finnish Technology Agency under project KUKOT, and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

REFERENCES

[1] Barron, A.R., Rissanen, J., and Yu, B. (1998), 'The MDL Principle in Modeling and Coding', *IEEE Trans. Information Theory*, Vol. **IT-44**, No. 6, pp 2743–2760

[2] Davis, M.H.A. and Hemerly, E.M. (1990), 'Order Determination and Adaptive Control of ARX Models Using the PLS Criterion', *Proceedings of the Fourth Bad Honnef Conference on Stochastic Differential Systems. Lecture Notes in Control and Information Sci.* (N. Christopeit, ed.) Springer, New York

[3] Krichevsky, R.E. and Trofimov V.K. (1981), 'The Performance of Universal Encoding', *IEEE Trans. Information Theory*, Vol. **IT-27**, No. 2, pp. 199–207

[4] Lai, T.L. and Wei, C.Z. (1982), 'Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems', *Annals of Statistics*, Vol 10, **1**, 154–166

[5] Rissanen, J. (1984), 'Universal Coding, Information, Prediction, and Estimation', *IEEE Trans. Information Theory*, Vol. **IT-30**, No. 4, 629–636

[6] Rissanen, J. (1986), 'A Predictive Least Squares Principle', IMA J. Math. Control Inform. **3**, 211–222

[7] Rissanen, J. (1996), 'Fisher Information and Stochastic Complexity', *IEEE Trans. Information Theory*, Vol. **IT-42**, No. 1, pp 40–47

[8] Rissanen, J. (2007), *Information and Complexity in Statistical Modeling*, Springer Verlag, Series Information Science and Statistics, 140 pages

[9] Shtarkov, Y. (1987), 'Universal Sequential Coding of Single Messages', *Problems of Information Transmission*, Vol. **23**, No. 3, pp. 175–186

[10] Takimoto, E. and Warmuth, M. (2000), 'The Last-Step Minimax Algorithm', *Proceedings of the 11'th International Conference on Algorithmic Learning Theory*

[11] Wei, C.Z. (1992), 'On Predictive Least Squares Principles', *Annals of Statistics*, Vol 20, **1**, 1–42