Second International Conference
Modelling and Development of Intelligent Systems
Sibiu - Romania, September 29 - October 02, 2011

Supervised Approach to Learning Multivariate Linear Systems

Luminiţa State, Iuliana Paraschiv-Munteanu

Abstract

We consider the proble of developing a learning fro data sche e for the unknown input-output dependency of a syste **S** of linear type in the sense that the m-di ensional output of **S** results by co bining in a linear way the effects of n observable variables and the effects of several unobservable latent variables. The effects of the latent variables on the output is treated as additive noise, that is being given the observable vector x, the syste co putes the output $y = \beta^T \begin{pmatrix} 1 \\ x \end{pmatrix} + \varepsilon$, where β is a $(n+1) \times m$ atrix and ε is a m-di ensional Gaussian variable. In the paper the athe atical argu ents for the esti ation sche e based exclusively on a finite size set of observations is provided. We present an experi ental evaluation of the quality of the resulted learning sche e in order to establish conclusions concerning their accuracy and generalization capacities, the evaluation being perfor ed in ter s of etric, probabilistic and infor ational criterion functions.

1 Introduction

The tremendous growth in practical applications of machine learning over the past decade has been accompanied by a wide variety of important developments in the underlying algorithms and techniques that make use of concepts and results coming from several areas as mathematical statistics, computer science and engineering.

Since the main aim of machine learning is to obtain computer programs that are able to extract information from samples of data and as well as knowledge from the past experience and include them in the process of solving problems of high complexity, the research methodology in the field of machine learning is essentially based on a large class of concepts and results coming from mathematical statistics, neural and evolutionary computation, brain models, adaptive control theory and so on.

The aim of the research was to develop a model free learning methodology in order to predict a system behavior, conventionally denoted by **S** on the basis of finite size sequence of observations. In real life applications, the input data sequence are either obtained in a controlled way, that is the observer knows in advance the generating mechanism, or, in an uncontrolled way when the generating mechanism is ignored by the observer. We denote by **G** (Generator) the component that supplies a series of samples from an n-dimensional space loaded as inputs to us. The learning environment is assumed to be of supervised type, that is the output of **S** is available to the observer. The goal is to develop a learning component **L** on the basis of a finite size set of input-output observations, that is to infer the unknown the input-output dependency of **S** and use it for further predictions. The general scheme of our model is presented in Figure 1 ([3]).

The learning component **L** is responsible for a class of possible models for the unknown dependency corresponding to **S**. In other words, the learning component **L** implements a class of hypothesis (models) Ω, such that to each particular hypothesis $\omega \in \Omega$ corresponds a function $\varphi_\omega : \mathcal{X} \to \mathcal{Y}$ defined on the space

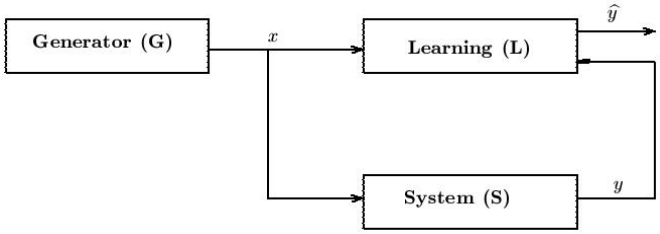

Figure 1: The learning environment.

of inputs $\mathcal{X}$ and taking values in the space of outputs $\mathcal{Y}$. For each particular input x_0, $\widehat{y}_0 = \varphi_\omega(x_0)$ is the estimate of the **S**'s output corresponding to x_0 in case of the model ω. Being given a criterion function $\mathcal{C}$ that expresses numerically the fitness of each model with respect to the available evidence E, about **S**, the best model $\omega_0(E)$ is a solution of the optimization problem

$$\arg\left(\text{optimize}_{\omega\in\Omega}\,\mathcal{C}(\omega, E)\right) . \tag{1}$$

In the case of supervised learning the available evidence E is represented by a finite set of pairs $\{(x_i, y_i)\ ,\ 1 \le i \le N\} \subset \mathcal{X} \times \mathcal{Y}$, where each y_i is the actual output of **S** for the input x_i. If we assume that the unknown dependency is of deterministic type, that is the inputs and the outputs of **S** are functionally related a reasonable choice of the criterion function $\mathcal{C}$ is the arithmetic mean of the square errors, that is for each $\omega \in \Omega$,

$$\mathcal{C}(\omega, E) = \frac{1}{N}\sum_{i=1}^{N} \|y_i - \varphi_\omega(x_i)\|^2 . \tag{2}$$

The optimization problem (1) becomes

$$\arg\left(\min_{\omega\in\Omega} \mathcal{C}(\omega, E)\right) , \tag{3}$$

and its solutions are called the Minimum Square Errors (MSE) models computed on the basis of $\{(x_i, y_i)\ ,\ 1 \le i \le N\}$ ([14]).

A more realistic approach is to accept that besides the inputs, the outputs of **S** are also influenced by a series of unknown number of unobservable factors referred as latent variables. In this case the cumulated effects of the latent variables can by only modelled in probabilistic terms, assuming some class of multivariate probability distributions, each hypothesis corresponding to a certain input-output dependency combined with a probabilistic model for the latent vector. For simplicity sake, we consider that the latent vector is a continuous random vector, that is to each hypothesis $\omega \in \Omega$ corresponds a conditional density function $f(\cdot|\cdot\,;\omega)$ ([9], [11]). Put in other words, for each $\omega \in \Omega$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $f(y|x\,;\omega)$ expresses 'the chance' of getting the output y for the input x in case of the model ω. If the available evidence about **S** is $\{(x_i, y_i)\ ,\ 1 \le i \le N\}$ then a reasonable choice of $\mathcal{C}(\omega, E)$ is the likelihood function. If we assume that the inputs $x_1, \ldots, x_N$ are independently generated by **G** then

$$\mathcal{C}(\omega, E) = \prod_{i=1}^{N} f\left(y_i|x_i\,;\omega\right) , \tag{4}$$

and the optimization problem (1) becomes

$$\arg\left(\max_{\omega\in\Omega} \mathcal{C}(\omega, E)\right) . \tag{5}$$

The solutions of (5) are the MLE (Maximum Likelihood Estimation) models computed on the basis of $\{(x_i, y_i)\ ,\ 1 \le i \le N\}$.

2 A MLE-based approach in learning multivariate linear systems

In our work we assume that the inputs come from a n-dimensional real space and for each input x the output y of $\mathbf{S}$ is a tuple of m real variables. It is also assumed that the output dependency on the input is of linear type, that is the influence of the input is given by $\beta^T x$ where β is an unknown $(n+1)\times m$ matrix. The effects of the latent variables can by taken as noise additively superimposed on the influence of the inputs, that is the output of $\mathbf{S}$ is $y = \beta^T \begin{pmatrix} 1 \\ x \end{pmatrix} + \varepsilon$ where ε is a m-dimensional Gaussian distributed random vector, $\varepsilon \sim \mathcal{N}(\mu, \Sigma)$. The parameters $\mu \in \mathbb{R}^m$ and $\Sigma \in \mathcal{M}_m(\mathbb{R})$, symmetric positive definite matrix, are unknown. Therefore, the dimension of the space of hypotheses Ω is $m(m+n+2)$, and for each particular tuple of $m(m+n+2)$ parameters, $\omega = (\beta, \mu, \Sigma)$ defines a model of $\mathbf{S}$. We denote by $f(y|x\,;\omega)$ the conditional density function on the output space corresponding to the model ω.

Several intuitively justified criterion functions can be considered in order to quantitatively express the quality of each possible model with respect to available sequence of observations. For each criterion function, the identification of the "fittest" model reduces to solving an constrained/unconstrained optimization problem ([6]).

Let $\mathcal{S}_N = \{(x_i, y_i)\ , 1 \le i \le N\}$ be the sequence of input-output observations taken on $\mathbf{S}$. Being given the model $\omega = (\beta, \mu, \Sigma)$, the estimate $\widetilde{y}$ of the actual output of $\mathbf{S}$ being given the input x is a random vector of density function

$$f(\widetilde{y}|x,\omega) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left\{-\frac{1}{2}\left(\widetilde{y} - \beta^T z - \mu\right)^T \Sigma^{-1} \left(\widetilde{y} - \beta^T z - \mu\right)\right\},$$

where $z = \begin{pmatrix} 1 \\ x \end{pmatrix}$.

Consequently, the log-likelihood function is

$$l(\beta, \mu, \Sigma, \mathcal{S}_N) = -\frac{mN}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}\left((y_i-\mu)^T - z_i^T\beta\right)\Sigma^{-1}\left((y_i-\mu) - \beta^T z_i\right),$$

and the best model from the point of view of maximum likelihood principle is a solution of the constrained optimization problem

$$\begin{cases} \min\limits_{\beta,\mu,\Sigma} \left(\Phi(\beta,\mu,\Sigma)\right) \\ \\ \Sigma \in \mathcal{M}_m(\mathbb{R}) \text{ symmetric and positive defined,} \end{cases}$$

where

$$\Phi(\beta,\mu,\Sigma) = N\ln|\Sigma| + \sum_{i=1}^{N}\left((y_i-\mu)^T - z_i^T\beta\right)\Sigma^{-1}\left((y_i-\mu) - \beta^T z_i\right).$$

Let us denote

$$\begin{aligned} &Y = (y_1, \ldots, y_N) \in \mathcal{M}_{m\times N}(\mathbb{R}),\ Z = (z_1, \ldots, z_N) \in \mathcal{M}_{(n+1)\times N}(\mathbb{R}), \\ &u = (1, \ldots, 1)^T \in \mathbb{R}^N,\ A = \mathbf{I}_N - \frac{1}{N}uu^T. \end{aligned} \tag{6}$$

For any matrix B, we denote by B^+ the generalized inverse (Penrose pseudo-inverse) of B.

Theorem 1 *The objective function $\Phi(\beta,\mu,\Sigma)$ has an unique critical point $(\beta_0, \mu_0, \Sigma_0)$ where*

$$\beta_0 = \left(Y(ZA)^+\right)^T,\ \mu_0 = \frac{1}{N}\left(Yu - Y(ZA)^+ Zu\right),\ \Sigma_0 = \frac{1}{N}Y\left(A - (ZA)^+(ZA)\right)Y^T, \tag{7}$$

and Σ_0 is a symmetric and positive semi-defined matrix.

Proof. The generalized gradients of $\Phi(\beta,\mu,\Sigma)$ with respect to β, μ, and Σ respectively, are

$$\begin{cases} \nabla_\beta \Phi(\beta,\mu,\Sigma) = -\left(ZY^T - Zu\,\mu^T - ZZ^T\beta\right)\Sigma^{-1}, \\ \nabla_\mu \Phi(\beta,\mu,\Sigma) = -\Sigma^{-1}\left(Yu - N\mu - \beta^T Zu\right), \\ \nabla_\Sigma \Phi(\beta,\mu,\Sigma) = \Sigma^{-1}D\Sigma^{-1} - \dfrac{1}{2}\operatorname{diag}\left(\Sigma^{-1}D\Sigma^{-1}\right), \end{cases}$$

where

$$D = N\Sigma - YY^T + \left(\mu\,(Yu)^T + (Yu)\,\mu^T\right) + \left(\left(ZY^T\right)^T\beta + \beta^T ZY^T\right) - N\left(\mu\mu^T\right) - \left(\mu\,(Zu)^T\beta + \beta^T(Zu)\,\mu^T\right) - \beta^T ZZ^T\beta\,.$$

From the system

$$\begin{cases} \nabla_\mu \Phi(\beta,\mu,\Sigma) = \mathbf{0}_m, \\ \nabla_\beta \Phi(\beta,\mu,\Sigma) = \mathbf{O}_{n+1,m}, \end{cases}$$

since $|\Sigma| \neq 0$, we get

$$\begin{cases} Yu - N\mu - \beta^T Zu = \mathbf{0}_m \\ ZY^T - Zu\,\mu^T - ZZ^T\beta = \mathbf{O}_{n+1,m}\,, \end{cases}$$

that is

$$\beta = \left(Y\,(ZA)^+\right)^T = \beta_0\,, \quad \mu = \frac{1}{N}\left(Yu - Y\,(ZA)^+ Zu\right) = \mu_0\,.$$

Replacing μ_0, β_0 in the system

$$\nabla_\Sigma \Phi(\beta,\mu,\Sigma) = \mathbf{O}_m$$

we obtain

$$\Sigma^{-1}D\Sigma^{-1} - \frac{1}{2}\operatorname{diag}\left(\Sigma^{-1}D\Sigma^{-1}\right) = \mathbf{O}_m,$$

where $\operatorname{diag}\left(\Sigma^{-1}D\Sigma^{-1}\right) \in \mathcal{M}_m(\mathbf{R})$ is the diagonal matrix that retains only the entries placed on the main diagonal of $\Sigma^{-1}D\Sigma^{-1}$. Since Σ is a positive definite matrix, we get $D = \mathbf{O}_m$ and consequently,

$$\Sigma = \frac{1}{N}\left(YY^T + \beta_0^T ZZ^T\beta_0 - YZ^T\beta_0 - \beta_0^T ZY^T + \frac{1}{N}\left(\beta_0^T Zuu^T Y^T - \beta_0^T Zuu^T Z^T\beta_0 + Yuu^T Z^T\beta_0 - Yuu^T Y^T\right)\right) =$$

$$\frac{1}{N}\left(YAY^T + \beta_0^T ZAZ^T\beta_0 - YAZ^T\beta_0 - \beta_0^T ZAY^T\right).$$

Using the well-known properties of the Penrose pseudo-inverse, the expression of Σ becomes

$$\begin{aligned} \Sigma &= \frac{1}{N}\left(YAY^T + Y\,(ZA)^+(ZA)\,AZ^T\left(Y\,(ZA)^+\right)^T - YAZ^T\left(Y\,(ZA)^+\right)^T - Y\,(ZA)^+ ZAY^T\right) = \\ &\frac{1}{N}\left(YAY^T + Y\left((ZA)^+(ZA)\right)^T(ZA)^T\left((ZA)^+\right)^T Y^T - Y\left((ZA)^+(ZA)\right)^T Y^T - Y(ZA)^+(ZA)\,Y^T\right) = \\ &\frac{1}{N}Y\left(A + \left((ZA)^+(ZA)\,(ZA)^+(ZA)\right)^T - (ZA)^+(ZA) - (ZA)^+(ZA)\right)Y^T = \\ &\frac{1}{N}Y\left(A - (ZA)^+(ZA)\right)Y^T \overset{not}{=} \Sigma_0\,. \end{aligned}$$

A further simplification can be obtained by noting that $B = A - (ZA)^+ ZA$ is a symmetric matrix and $B^2 = B$. Which finally yields to

$$\Sigma_0 = \frac{1}{N}YBB^TY^T\,.$$

Obviously, the estimate Σ_0 is a symmetric and positive semi-defined matrix. □

Theorem 2 *Let β_0 and μ_0 be given by Theorem 1. Then for any (β, μ, Σ) in the parameter space,*

$$l(\beta, \mu, \Sigma, \mathcal{S}_N) \leq l(\beta_0, \mu_0, \Sigma, \mathcal{S}_N). \tag{8}$$

Proof. Using (6) and $v^T v = \operatorname{tr}\left(vv^T\right)$ the expression of the log-likelihood function can be written as

$$l(\beta, \mu, \Sigma, \mathcal{S}_N) = -\frac{mN}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\left(Y - \mu u^T - \beta^T Z\right)\left(Y - \mu u^T - \beta^T Z\right)^T\right).$$

Therefore

$$l(\beta_0, \mu_0, \Sigma, \mathcal{S}_N) = -\frac{mN}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}Y\left(A - (ZA)^+ ZA\right)\left(A - (ZA)^+ ZA\right)^T Y^T\right).$$

Using the relations $A = A^2 = A^T$ we get

$$\begin{aligned}&\left(A - (ZA)^+ ZA\right)\left(A - (ZA)^+ ZA\right)^T = \\ &A^2 - A(ZA)^+(ZA) - (ZA)^+(ZA)A + (ZA)^+(ZA)(ZA)^+(ZA) = \\ &A - A(ZA)^+(ZA) - (ZA)^+(ZA) + (ZA)^+(ZA) = A - A(ZA)^+(ZA).\end{aligned}$$

therefore the term $Y - \mu u^T - \beta^T Z$ becomes

$$Y - \mu u^T - \beta^T Z = Y\left(A - (ZA)^+(ZA)\right) + \frac{1}{N}\left(Yu - Y(ZA)^+ Zu - N\mu\right)u^T + \left(Y(ZA)^+ - \beta^T\right)Z =$$

$$Y\left(A - (ZA)^+(ZA)\right) + (\mu_0 - \mu)u^T + (\beta_0 - \beta)^T Z.$$

Consequently,

$$\begin{aligned}l(\beta, \mu, \Sigma, \mathcal{S}_N) = &-\frac{mN}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}Y\left(A - (ZA)^+ ZA\right)\left(A - (ZA)^+ ZA\right)^T Y^T\right) - \\ &\operatorname{tr}\left(\Sigma^{-1}Y\left(A - (ZA)^+ ZA\right)\left((\mu_0 - \mu)u^T\right)^T\right) - \operatorname{tr}\left(\Sigma^{-1}Y\left(A - (ZA)^+ ZA\right)\left((\beta_0 - \beta)^T Z\right)^T\right) - \\ &\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(\mu_0 - \mu)u^T u(\mu_0 - \mu)^T\right) - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(\beta_0 - \beta)^T ZZ^T(\beta_0 - \beta)\right) =\end{aligned} \tag{9}$$

$$\begin{aligned}&l(\beta_0, \mu_0, \Sigma, \mathcal{S}_N) - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(\mu_0 - \mu)u^T u(\mu_0 - \mu)^T\right) - \frac{1}{2}\operatorname{tr}\left(\left(\Sigma^{-1}\beta_0 - \beta\right)^T ZZ^T(\beta_0 - \beta)\right) - \\ &\operatorname{tr}\left(\Sigma^{-1}Y\left(\mathbf{I}_N - (ZA)^+ Z\right)Au(\mu_0 - \mu)^T\right) - \operatorname{tr}\left(\Sigma^{-1}Y\left(\mathbf{I}_N - (ZA)^+ Z\right)AZ^T(\beta_0 - \beta)\right).\end{aligned}$$

Since

$$Au = \left(\mathbf{I}_N - \frac{1}{N}uu^T\right)u = u - \frac{1}{N}u\left(u^T u\right) = \mathbf{0}_N$$

we get

$$Y\left(\mathbf{I}_N - (ZA)^+ Z\right)Au(\mu_0 - \mu)^T = \mathbf{O}_m.$$

Also, using the properties of the Penrose pseudo-inverse, we get

$$\begin{aligned}\left(\mathbf{I}_N - (ZA)^+ Z\right)AZ^T &= (ZA)^T - (ZA)^+ ZAA^T Z^T = (ZA)^T - \left((ZA)^+(ZA)\right)^T(ZA)^T = \\ &(ZA)^T - \left((ZA)(ZA)^+(ZA)\right)^T = (ZA)^T - (ZA)^T = \mathbf{O}_{N,n+1},\end{aligned}$$

that is $Y\left(\mathbf{I}_N - (ZA)^+ Z\right)AZ^T(\beta_0 - \beta) = \mathbf{O}_m$.
Taking into account these arguments we finally obtain,

$$l(\beta, \mu, \Sigma, \mathcal{S}_N) = l(\beta_0, \mu_0, \Sigma, \mathcal{S}_N) - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(\mu_0 - \mu)u^T u(\mu_0 - \mu)^T\right) - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(\beta_0 - \beta)^T ZZ^T(\beta_0 - \beta)\right).$$

Obviously, since Σ a positive definite matrix,

$$\operatorname{tr}\left(\Sigma^{-1}(\mu_0-\mu)u^T u(\mu_0-\mu)^T\right) = N\operatorname{tr}\left((\mu_0-\mu)^T\Sigma^{-1}(\mu_0-\mu)\right) = N(\mu_0-\mu)^T\Sigma^{-1}(\mu_0-\mu) \geq 0$$

and

$$\operatorname{tr}\left(\Sigma^{-1}(\beta_0-\beta)^T ZZ^T(\beta_0-\beta)\right) = \operatorname{tr}\left(Z^T(\beta_0-\beta)\Sigma^{-1}(\beta_0-\beta)^T Z\right) = \operatorname{tr}\left(\left((\beta_0-\beta)^T Z\right)^T\Sigma^{-1}(\beta_0-\beta)^T Z\right) \geq 0$$

that is

$$l(\beta,\mu,\Sigma,\mathcal{S}_N) \leq l(\beta_0,\mu_0,\Sigma,\mathcal{S}_N). \quad \square$$

Remark. Although, a long series of tests pointed out that the estimate Σ_0 given by Theorem 1 is a positive matrix the mathematical proof is still an open problem. Also, it is not known whether the unique critical point (β_0,μ_0,Σ_0) corresponds to the best model in the sense of the maximum likelihood principle.

An adaptive learning procedure can be obtained using the gradient ascent method applied to the log-likelihood criterion function. The search developed by the adaptive procedure in a $m(m+n+2)$-dimensional space aims to adjust the model parameters β, μ, Σ in order to maximize the log-likelihood function or, equivalently, to minimize $\Phi(\beta,\mu,\Sigma)$. The procedure should be implemented using a control parameter $\delta > 0$ and a stopping condition $\mathcal{C}(\delta)$ usually expressed in terms of the magnitude of the displacement in the parameter space due to the current iteration. In our tests $\mathcal{C}(\delta) = true$ if

$$\|\beta^{new}-\beta^{old}\| < \delta\,,\ \|\mu^{new}-\mu^{old}\| < \delta\,,\ \|\Sigma^{new}-\Sigma^{old}\| < \delta$$

where $\|\cdot\|$ is a conventionally norm, for instance Euclidian norm.

Also, the implementation of the procedure can be done using either a constant learning rate $\rho > 0$ or a decreasing sequence of positive learning rates (ρ_k) that refines the search while the search advances.

Since during the search process the estimates of Σ are not guaranteed to be invertible, the implementation procedure `MLE_gradient_ascent` uses approximations of the actual generalized gradients where the generalized inverse is used instead.

`procedure MLE_gradient_ascent`
`Input:` $\{(x_1,y_1),\ldots,(x_N,y_N)\}$
`Initializations:` $\delta > 0$, $\rho > 0$, $\widetilde{\beta}$, $\widetilde{\mu}$, $\widetilde{\Sigma}$,
$Z = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_N \end{pmatrix}$, $Y = (y_1,\ldots,y_N)$, $u = (1,\ldots,1)^T$
$\beta^{old} \leftarrow \widetilde{\beta}$, $\mu^{old} \leftarrow \widetilde{\mu}$, $\Sigma^{old} \leftarrow \widetilde{\Sigma}$
`Compute` $S = ZZ^T$, $Q = ZY^T$, $P = YY^T$
$Z_1 = Zu$, $Y_1 = Yu$
`repeat`
$\Sigma_1 \leftarrow \left(\Sigma^{old}\right)^+$
$\beta^{new} \leftarrow \beta^{old} + \rho\left(Q - Z_1\left(\mu^{old}\right)^T - S\beta^{old}\right)\Sigma_1$
$\mu^{new} \leftarrow \mu^{old} + \rho\Sigma_1\left(Y_1 - \left(\beta^{old}\right)^T Z_1 - N\mu^{old}\right)$
$D = N\Sigma^{old} - P + Y_1\left(\mu^{old}\right)^T + \left(Y_1\left(\mu^{old}\right)^T\right)^T + Q^T\beta^{old} + \left(Q^T\beta^{old}\right)^T -$
$N\mu^{old}\left(\mu^{old}\right)^T - \mu^{old}(Z_1)^T\beta^{old} - \left(\mu^{old}(Z_1)^T\beta^{old}\right)^T - \left(\beta^{old}\right)^T S\beta^{old}$
$\Sigma^{new} \leftarrow \Sigma^{old} + \rho\left(-\Sigma_1 D\Sigma_1 + \frac{1}{2}\,\mathtt{diag}\left(\Sigma_1 D\Sigma_1\right)\right)$
`evaluate` $\mathcal{C}(\delta)$
$\beta^{old} \leftarrow \beta^{new}$, $\mu^{old} \leftarrow \mu^{new}$, $\Sigma^{old} \leftarrow \Sigma^{new}$
`until` $\mathcal{C}(\delta)$
`Output:` β^{new}, μ^{new}, Σ^{new}.

3 Experimental analysis

Because the model $(\beta_0, \mu_0, \Sigma_0)$ given by (7) is not theoretical guaranteed as the best model from the point of view of maximum likelihood principle, we have performed a long series of tests aiming to derive conclusions concerning the performance of the proposed method on experimental way. The test examples x_i's were random generated from n-dimensional Gaussian repartition $\mathcal{N}(\mu_1, \Sigma_1)$. The target responses y_i's were computed as $y_i = \widetilde{\beta}^T x_i + \varepsilon$ for given β and ε randomly generated from known Gaussian repartition $\mathcal{N}\left(\widetilde{\mu}, \widetilde{\Sigma}\right)$.

According to the previous arguments the expression of conditional density function on the output space corresponding to each example x_i being given the model $\omega = (\beta, \mu, \Sigma)$ is

$$f(y|x_i, \omega) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} \exp\left\{-\frac{1}{2}\left(y - \beta^T z_i - \mu\right)^T \Sigma^{-1} \left(y - \beta^T z_i - \mu\right)\right\},$$

where $z_i = \begin{pmatrix} 1 \\ x_i \end{pmatrix}$, therefore the most likely output predicted value is $y_i' = \beta^T z_i + \mu$.

In order to evaluate the quality of the resulted model we considered three indicators to evaluate the overall error.

The first indicator evaluates the overall mean error of miss-prediction for the given set of example $\{(x_i, y_i) \mid 1 \le i \le N\}$ corresponding to each possible model ω

$$error\_1 = \frac{1}{N} \sum_{i=1}^{N} \left(1 - f(y_i|x_i, \omega)\right) . \tag{10}$$

The second indicator is a mean error computed in terms of the actual responses and the most likely predicted values,

$$error\_2 = \frac{1}{N} \sum_{i=1}^{N} \|y_i - y_i'\|^2 = \frac{1}{N} \sum_{i=1}^{N} \|y_i - \beta^T z_i - \mu\|^2 . \tag{11}$$

The third measure, of informational type, aims to evaluate the informational correlation between the input and the computed output corresponding to each model, and it is expressed in terms of the empirical mutual information. Since $x_1, \ldots, x_N$ are randomly generated $\mathcal{N}(\mu_1, \Sigma_1)$, the probability distribution $\widetilde{p} = (\widetilde{p}(x_1), \ldots, \widetilde{p}(x_N))$ characterizes the collection of examples,

$$\widetilde{p}(x_j) = \frac{p(x_j)}{\sum_{i=1}^{N} p(x_i)} ,$$

where $p(x_j) = \dfrac{1}{\sqrt{(2\pi)^n |\Sigma_1|}} \exp\left\{-\dfrac{1}{2}(x_j - \mu_1)^T \Sigma_1^{-1} (x_j - \mu_1)\right\}$, $1 \le j \le N$.

The empirical entropy of the input samples $x_1, \ldots, x_N$ is given by the Shannon entropy corresponding to $\widetilde{p}$,

$$H(\widetilde{p}) = -\sum_{i=1}^{N} \widetilde{p}(x_i) \ln \widetilde{p}(x_i) . \tag{12}$$

Using the transition probabilities

$$\widetilde{p}(y_j | x_i, \omega) = \frac{f(y_j | x_i, \omega)}{\sum_{k=1}^{N} f(y_k | x_i, \omega)} , \quad 1 \le i, j \le N ,$$

we define the probability distribution $\widetilde{q} = (\widetilde{q}(y_1), \ldots, \widetilde{q}(y_N))$ on the set of target responses by $\widetilde{q}(y_j) = \sum_{i=1}^{N} \widetilde{p}(x_i)\, \widetilde{p}(y_j | x_i, \omega)$, $1 \le j \le N$, and let

$$H(\widetilde{q}) = -\sum_{i=1}^{N} \widetilde{q}(y_i) \ln \widetilde{q}(y_i) , \tag{13}$$

be the empirical Shannon entropy of the set of **S**'s outputs.

Using the well-known expression of the mutual information ([4]), the empirical mutual information can be defined as

$$\mathcal{I}(\mathcal{S}_N) = H(\widetilde{q}) - \sum_{i=1}^{N}\sum_{j=1}^{N} \widetilde{p}(x_i)\,\widetilde{p}(y_j|\,x_i,\omega)\ln\widetilde{p}(y_j|\,x_i,\omega)\ . \tag{14}$$

We evaluated the performance of the model $\omega_0 = (\beta_0, \mu_0, \Sigma_0)$ given by (7) for different sizes for learning samples $\mathcal{S}_N$.

Test 1. The settings are $n = 2$, $m = 1$, $\mu_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\widetilde{\beta} = \begin{pmatrix} 0 \\ 2 \\ 4 \end{pmatrix}$, $\widetilde{\mu} = 0.25$ and $\widetilde{\Sigma} = 1$. Some of the results for data of different sizes N, corresponding to the quasi-optimal model ω_0 computed for each data set are summarized in Table 1*a*.

Table 1: *Model evaluation in case $n = 2$, $m = 1$ and for different volumes of: a) learning data; b) test samples.*

a

N	$error\_1$	$error\_2$	$H(\widetilde{p})$	$H(\widetilde{q})$	$\mathcal{I}(\mathcal{S}_N)$
10	0.603	0.533	2.214	2.205	1.335
20	0.702	0.802	2.767	2.796	1.304
50	0.711	0.892	3.768	3.831	1.226
100	0.712	0.950	4.477	4.524	1.233
200	0.716	0.959	5.098	5.188	1.221
300	0.721	1.019	5.510	5.590	1.106
400	0.713	0.953	5.822	5.897	1.155
500	0.725	1.025	6.023	6.113	1.150

b

N_{test}	$error\_1$	$error\_2$	$H(\widetilde{p})$	$H(\widetilde{q})$	$\mathcal{I}(\mathcal{S}_N)$
15	0.706	1.149	2.458	2.582	1.346
30	0.710	1.257	3.243	3.276	1.367
50	0.683	1.039	3.752	3.829	1.573
70	0.650	1.136	4.146	4.172	1.573
90	0.707	1.292	4.309	4.400	1.637
100	0.709	1.276	4.376	4.483	1.662
200	0.715	1.361	5.124	5.235	1.616
300	0.719	1.375	5.547	5.633	1.579

Several tests aimed to establish conclusions concerning the generalization capacity of the quasi-optimal model. For instance, in case of a training sequence of volume $N = 100$, the computed quasi-optimal model is $\omega_0 = (\beta_0, \mu_0, \Sigma_0)$ is

$$\beta_0 = \begin{pmatrix} 0.000 \\ 2.614 \\ 4.077 \end{pmatrix} \quad \mu_0 = -0.218\,, \quad \Sigma_0 = 0.457\,,$$

the results are summarized in table 1*b* and the variations of the input, output empirical entropies and the empirical mutual information as functions of the test sample sizes are depicted in Figure 2*b*.

Test 2. The settings are $n = 2$, $m = 2$, $\mu_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\widetilde{\beta} = \begin{pmatrix} 0 & 0 \\ 2 & 1 \\ 4 & 5 \end{pmatrix}$, $\widetilde{\mu} = \begin{pmatrix} 0.25 \\ 0.25 \end{pmatrix}$ and $\widetilde{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Some of the results for data of different sizes N, corresponding to the quasi-optimal model ω_0 computed for each data set are summarized in Table 2*a*.

Several tests aimed to establish conclusions concerning the generalization capacity of the quasi-optimal model. For instance, in case of a training sequence of volume $N = 100$, the computed quasi-optimal model is $\omega_0 = (\beta_0, \mu_0, \Sigma_0)$ is

$$\beta_0 = \begin{pmatrix} -0.000 & -0.000 \\ 1.872 & 1.012 \\ 4.134 & 5.047 \end{pmatrix} \quad \mu_0 = \begin{pmatrix} 0.274 \\ 0.104 \end{pmatrix}, \quad \Sigma_0 = \begin{pmatrix} 0.677 & -0.007 \\ -0.007 & 0.899 \end{pmatrix},$$

the results are summarized in table 2*b* and the variations of the input, output empirical entropies and the empirical mutual information as functions of the test sample sizes are depicted in Figure 3*b*.

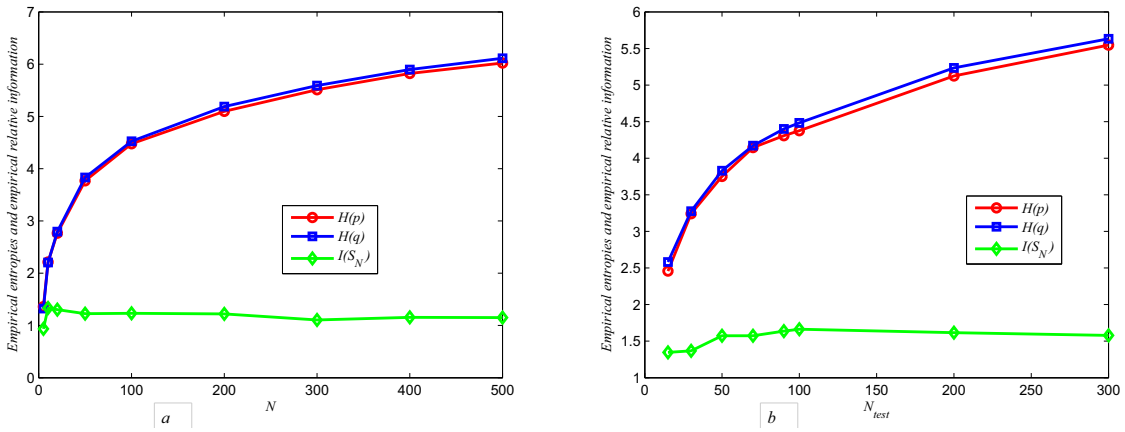

Figure 2: *The variation of the empirical input/output entropies and empirical mutual information: a) as function of N, volume of learning data; b) for $N = 100$ and different volumes of test samples.*

Table 2: *Model evaluation in case $n = 2$, $m = 2$ and for different volumes of: a) learning data; b) test samples.*

a

N	$error\_1$	$error\_2$	$H(\widetilde{p})$	$H(\widetilde{q})$	$\mathcal{I}(\mathcal{S}_N)$
15	0.887	1.436	2.500	2.483	1.610
20	0.877	1.282	2.816	2.720	1.521
50	0.892	1.473	3.748	3.675	1.810
100	0.926	2.030	4.441	4.369	1.583
200	0.908	1.774	5.090	5.065	1.704
300	0.923	2.072	5.513	5.495	1.638
400	0.919	1.982	5.810	5.768	1.661
500	0.918	1.954	6.014	5.987	1.763

b

N_{test}	$error\_1$	$error\_2$	$H(\widetilde{p})$	$H(\widetilde{q})$	$\mathcal{I}(\mathcal{S}_N)$
15	0.917	2.154	2.542	2.311	1.277
30	0.912	1.874	3.190	3.116	1.763
50	0.890	1.438	3.764	3.699	1.627
70	0.910	2.216	4.042	3.889	1.772
90	0.907	1.881	4.349	4.277	1.769
100	0.912	2.014	4.432	4.292	1.783
200	0.906	1.938	5.098	5.062	1.698
300	0.902	1.860	5.512	5.457	1.822

4 Conclusions

The aim of the report research was to develop a model free learning methodology in order to predict a system behavior, conventionally denoted by **S** on the basis of finite size sequence of observations. The learning environment is assumed to be of supervised type, that is a finite size sample of pairs of input-output values is available to the observer. The goal was to develop a learning component on the basis on the input-output sample, that is to infer the unknown the input-output dependency of **S** and use it for further predictions. In Section 2, the input-output dependency of **S** is modeled in terms of a Gaussian repartition of the output space. The model parameters are adjusted on the basis of the available sequence of observations using the Maximum Likelihood Principle. We managed to compute explicitly the quasi-optimal model ω_0 (Theorem 1) and establish its 'almost' optimality (Theorem 2).

The performance evaluation of the quasi-optimal model is evaluated in Section 3 aiming to establish conclusions concerning the quality of the predictions about **S** computed by the learning component. The performance is evaluated in terms of three indicators, on the basis of the available sequence of observations and new test samples. The tests confirm that the proposed methodology assures fast learning of good quality, the predicted values being 'close' to the actual values in case of learning samples as well as in case of new test data. A series of further developments, still in progress, aiming to extend and refine the proposed methodology by taking into account non-linear models are going to be published in the near future.

Acknowledgement: The paper reports a series of results of the research performed in the framework of the Doctoral School in Computer Science at the University of Piteşti, Romania.

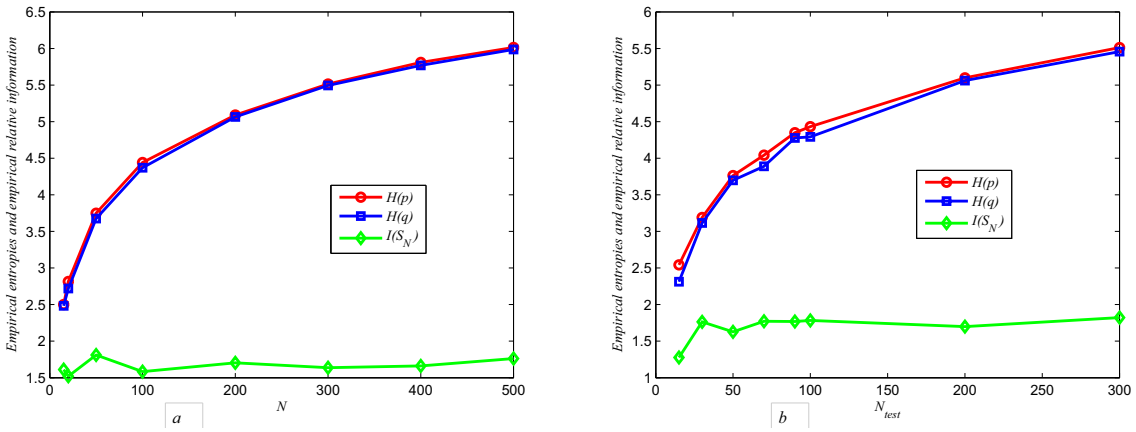

Figure 3: *The variation of the empirical input/output entropies and empirical mutual information: a) as function of N, volume of learning data; b) for $N = 100$ and different volumes of test samples.*

References

[1] E. Alpaydin, *Introduction to Machine Learning*, The MIt Press, Massachusetts, 2010.

[2] C. Bishop, *Pattern Recognition and Machine Learning*, Springer Verlag, 2006.

[3] V. Cherkassky, F. Mulier, *Learning from Data Concepts, Theory, and Methods*, Second Edition, John Wiley & Sons, Inc., 2007.

[4] T.M. Cover, J.A. Thomas, *Elements of information theory (Second Edition)*, John Wiley & Sons, Inc., 2006.

[5] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning - Data mining, Inference and Prediction (Second Edition)*, Springer-Verlag, 2009.

[6] A.J. Izenman, *Modern Multivariate Statistical Techniques: Regressian, Classification and Manifold Learning*, Springer, 2008.

[7] A.K. Jain, R. Duin, J. Mao, Statistical Pattern Recognition: A Review, *IEEE Transactions On Pattern Analysis and Machine Intelligence* 22(1): 4–37, 2000.

[8] S. Marsland, *Machine Learning: An Algorithmic Perspective*, CRC Press, Taylor & Francis Group, Boca Raton - London - New York, 2009.

[9] K.E. Muller, P.W. Stewart, *Linear Model Theory - Univariate, Multivariate and Mixed Models*, Wiley-Interscience, 2006.

[10] I. Paraschiv-Munteanu, Theoretical approach in performance evaluation of a classsification system, *University of Pitesti - Scientific Bulletin, Serie Mathematics and Computer Science*, 14: 139–150, 2008.

[11] A.C. Rencher, *Methods of Multivariate Analysis*, Wiley-Interscience, 2002.

[12] L. State, I. Paraschiv-Munteanu, *Introducere în teoria statistică a recunoaşterii formelor*, Editura Universităţii din Piteşti, Romania, 2009.

[13] S. Sharma, *Applied Multivariate Techniques*, John Wiley & Sons, Inc., 1996.

[14] K. Takeaki, K. Hiroshi, *Generalized Least Squares*, John Wiley & Sons, Inc., 2004.

Luminiţa State
University of Piteşti
Faculty of Mathematics and Computer Science
1 Târgu din Vale St., Piteşti 110040
ROMANIA
E-mail: *lstate@clicknet.ro*

Iuliana Paraschiv-Munteanu
University of Bucharest
Faculty of Mathematics and Computer Science
14 Academiei St., Bucharest 010014
ROMANIA
E-mail: *pmiulia@fmi.unibuc.ro*