Least Conservative Support and Tolerance Tubes

Kristiaan Pelckmans, Jos De Brabanter,
Johan A.K. Suykens, *Senior, IEEE*, Bart De Moor, *Fellow, IEEE*

***Abstract*—This[1] correspondence studies a distribution-free estimator of the conditional support and tolerance intervals of a distributions underlying a set of paired i.i.d. observations. The key ingredients are (a) an appropriate notion of risk which measures what probability mass is not captured by the estimate, (b) a uniform concentration inequality for the empirical risk based on a compression argument , and (c) the derivation of a lower-bound to the mutual information, dictating how to maximize the informativeness of the estimator. For this result we extend Fano's inequality to the bivariate case.**

***Index Terms*—Statistical Learning Theory, Fano's inequality, Mutual Information**

I. Introduction

GIVEN a set of paired observations $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ (for $d > 0$) which are i.i.d. copies of a random vector (X, Y) and possessing a fixed but unknown joint distribution F_{XY}, this letter concerns the question which values a random variable Y can possibly/likely take given a covariate $X \in \mathbb{R}^d$. We term the discussed estimator of this quantity *support and tolerance tubes*. This investigation is motivated as (a) one is often interested in other characteristics of the joint distribution than the conditional expectation (regression), while (b) one is not willing to make restrictive assumptions on the underlying distribution whatsoever: e.g. in econometrics one is often interested in the volatility of a market. In environmental sciences one is typically concerned with the extremal behavior (i.e. the min or max value) of a magnitude, and their relation to measurable environmental variables.

We argue that the present research objective can be approached naturally from a setting of statistical learning theory, see [1], [2] for an overview. A key quantity is the *risk* which is translated here as the probability mass falling outside the conditional support tube. A main conceptual difference with the literature on (parametric) statistical inference is that no attempt is made whatsoever to reveal the underlying conditional mean (as in regression) or the conditional quantile functions (as in parametric quantile regression). The precise question differs from the usual learning task as considered in the literature of machine learning as we do not aim for a single point-prediction. It is also paramount to differentiate the present contribution with the aim as set out in (nonparametric) quantile regression (see e.g. [3] for a global introduction, and e.g. [4], [5] for recent advances herein): even if there is no hope to recover the individual quantile functions (e.g. as they are not unique due to discontinuities), the present agnostic methods still yields a valid qualitative characterization of the conditional distribution. To see this, observe that our estimator has two important characteristics (i) what is the probability of *covering* the actual conditional support (i.e. what is the probability that a new sample can fall outside the estimated interval). (ii) to what extent is the estimated conditional support of the tube *conservative* (i.e. does it overestimate the actual conditional support?). Now, the difference with methods as (nonparametric) quantile estimation is that the risk of our estimate is only affected by the probability of covering the conditional support, and not by the amount of conservatism. This treatment of conservatism is especially useful whenever the underlying distributions cannot possibly be captured reasonably based on the given (small) sample. This reasoning differentiates the current work also from the literature on estimating level sets as introduced in [6]. As a practical consequence, our estimate will differ from the estimates given by quantile regression methods as e.g. the so-called quantile crossing problem [3] is avoided entirely.

Now, proper reduction of the conservatism of the estimate is motivated by extending Fano's inequality (see e.g. [7], p. 38), yielding a lower-bound to the mutual information of two random variables. This inequality was applied towards a different setting of classification (and other learning schemes), see e.g. [8] and citations. The derivation yields a non-parametric estimator of the Mutual Information (MI) possessing a probabilistic guarantee which is derived using a classical compression argument. The result differs from other estimators of entropy and the mutual information as e.g. using Fisher's information matrix [7] or based on Gaussian assumptions as e.g. in [9] as a distribution-free context is adopted. Furthermore, the result essentially improves on the classical

K.Pelckmans is Assistant Professor/Research Fellow (Forskarsassistent) at the University of Uppsala, Sweden, Department of Information Technology. E-mail: kp@it.uu.se

JDB is an associated docent at the K.U.Leuven, Belgium.

JS and BDM are (full) professors at K.U.Leuven, Belgium, ESAT, SCD/SISTA.

Manuscript received March, 2006; revised December 2008. This work was supported by the FWO Travel Grant and (Research Council KUL): GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; (Flemish Government): (FWO): PhD/postdoc grants, projects, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, G.0302.07. research communities (ICCoS, ANMMM, MLDM); (IWT): PhD Grants,GBOU (McKnow), Eureka-Flite2 - Belgian Federal Science Policy Office: IUAP P5/22,PODO-II,- EU: FP5-Quprodis; ERNSI; - Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard.

[1]This work resulted from discussions with U. von Luxburg, M. Pontil, J. Shawe-Taylor and M. Herbster.

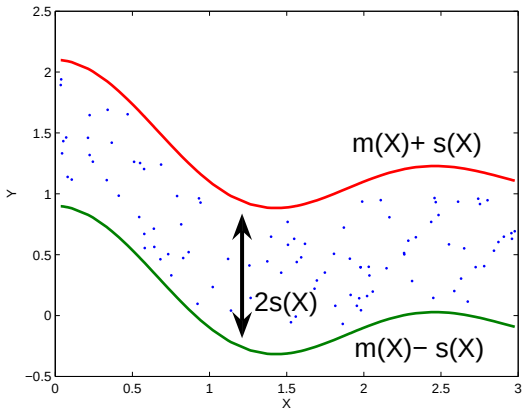


Fig. 1. *Example of a support vector tube based on a finite sample of a bivariate random variable (X, Y). A tube $\mathcal{T}_{m,s}$ is defined as the conditional interval $\mathcal{T}_{m,s}(x) = [m(x) - s(x),\ \ m(x) + s(x)]$ with fixed width $2s(x) = 2t > 0$.*

nonparametric estimators which were reviewed in [10], [11] as we do not impose restrictive (smoothness) assumptions on the underlying distributions whatsoever. As a consequence of this totally agnostic setting, the stated results are much weaker (only a lower-bound to the MI is given) and convergence properties are not directly in reach. But the result validates however our agnostic aims, particularly maximizing the (lower-bound to the) MI motivates optimization of a convenient measure of conservatism which is inversely proportional to the width of the tube.

Motivated by these derivations, (i) an estimator of the conditional support with maximal MI and controlling the risk is derived and its extension to the setting of conditional tolerance intervals is given; (ii) some theoretical guarantees for the estimate are provided; (iii) the relation to the method of the minimal convex hull is made explicit; and (iv) it is shown how the estimate can be computed efficiently by solving a linear program.

Section II proofs the main result, and explores the relation with the convex hull. From a practical perspective, Section III provides further insight in how the optimal estimate can be found efficiently by solving a linear program. We adopt the notation of representing random variables (or random vectors) as capital letters (e.g. $X, Y, ...$), while deterministic quantities (possibly in the same domain) are represented as lower case letters $(x, y, ...)$.

II. Support and Tolerance Tubes

A. Support Tubes and Risk

Definition 1 (Support and Tolerance Tubes) *Given a set of data $\mathcal{D}_n$ which are sampled i.i.d. from a fixed but unknown joint distribution F_{XY}. Let $\mathcal{H}_1 \subset \{m : \mathbb{R}^d \to \mathbb{R}\}$ and $\mathcal{H}_2 \subset \{s : \mathbb{R}^d \to \mathbb{R}^+\}$ be proper function spaces where the latter is restricted to positive functions and $\mathcal{H}_2 \subset \mathcal{H}_1$. Let $p(\mathbb{R})$ be the powerset of $\mathbb{R}$ such that $p(\mathbb{R}) = \{V \subset \mathbb{R}\}$. The class of tubes $\Gamma(\mathcal{H}_1, \mathcal{H}_2)$ is defined as*

$$\Gamma(\mathcal{H}_1, \mathcal{H}_2) = \Big\{\mathcal{T}_{m,s} : \mathbb{R}^d \to p(\mathbb{R}),\ m \in \mathcal{H}_1, s \in \mathcal{H}_2 \ \Big|\ \forall x \in \mathbb{R}^d : \mathcal{T}_{m,s}(x) = [m(x) - s(x),\ m(x) + s(x)]\Big\} \quad (1)$$

abbreviated as $\mathcal{T}_{m,s} = m \pm s$. A tube $\mathcal{T}_{m,s} \in \Gamma(\mathcal{H}_1, \mathcal{H}_2)$ is a true support tube (ST) of a joint distribution F_{XY} if the equality $P(Y \in \mathcal{T}_{m,s}(X)) = 1$ holds. Similarly a tube $\mathcal{T}_{m,s} \in \Gamma(\mathcal{H}_1, \mathcal{H}_2)$ is a true tolerance tube (TT) for F_{XY} of level $0 < \alpha < 1$ if $P(Y \in \mathcal{T}_{m,s}(X)) \geq 1 - \alpha$.

Let the indicator $\mathbb{I}(Y \notin \mathcal{T}_{m,s}(X))$ be equal to one if $Y \notin \mathcal{T}_{m,s}(X)$ and zero otherwise. We define the risk of a candidate ST for given joint distribution as follows

$$\mathcal{R}(\mathcal{T}_{m,s}; F_{XY}) = E\left[\mathbb{I}\left(Y \notin \mathcal{T}_{m,s}(X)\right)\right] = P\left(Y \notin \mathcal{T}_{m,s}(X)\right), \quad (2)$$

where the expectation is taken over the random variables X and Y with joint distribution F_{XY}. Its empirical counterpart becomes $\mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Y_i \notin \mathcal{T}_{m,s}(X_i)\right)$. The study of support tubes based on empirical samples will yield bounds of the form

$$P\left(\sup_{\mathcal{T}_{m,s} \in \Gamma} \mathcal{R}(\mathcal{T}_{m,s}; F_{XY}) \geq \epsilon\right) \leq \eta(\epsilon; \Gamma(\mathcal{H}_1, \mathcal{H}_2)), \quad (3)$$

where $0 < 1-\epsilon < 1$ is the *probability of covering* the tube and where the function $\eta(\cdot; \Gamma(\mathcal{H}_1, \mathcal{H}_2)) : [0, 1] \to [0, 1)$ expresses the *confidence level* in the probability of covering.

B. Generalization Bound

For now, we focus on the case of the ST, extensions specific to the TT are described in the next subsection. Assume a given hypothesis class $\Gamma(\mathcal{H}_1, \mathcal{H}_2)$ of STs. Consider an algorithm constructing a ST - say $\mathcal{T}_{m,s}$ - with zero empirical risk $\mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n) = 0$. The generalization performance can be bounded using a geometrical argument which was also used for deriving the compression bound outlined in [12], [13], and refined in various publications as e.g. [14].

Theorem 1 (Compression Bound on Risk of a ST) *Fix $\delta > 0$ and let $\mathcal{D}_n$ contain n i.i.d. samples from a fixed but unknown joint distribution F_{XY}. Consider the class of tubes Γ where each tube $\mathcal{T}_{m,s}$ is uniquely determined by D appropriate samples (i.e., $\mathcal{T}_{m,s}$ can be* 'compressed' *to D samples). Let $n_D = n - D$ denote the number of remaining samples. Fix $\delta > 0$, then with probability exceeding $1 - \delta$ one has that the following inequality holds for any $\mathcal{T}_{m,s}$ where $\mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n) = 0$:*

$$\sup_{\mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n) = 0} \mathcal{R}(\mathcal{T}_{m,s}; F_{XY}) \leq \frac{\log\left(K_{n,D}(\Gamma)\right) + \log\left(\frac{1}{\delta}\right)}{n - D} \triangleq c(\delta, D, n), \quad (4)$$

where we define $K_{n,D}(\Gamma)$ as

$$K_{n,D}(\Gamma) = \binom{n}{D}(2^{D-1} - 1) \leq \left(\frac{2ne}{D}\right)^{D}. \quad (5)$$

Proof: Let $\epsilon > 0$ be fixed and consider a ST determined by D samples - say the first D samples $\{(X_1, Y_1), \ldots, (X_D, Y_D)\}$ - denoted as $\mathcal{T}^D_{m,s}$. Assume F_{XY} is such that the actual risk of this tube is larger than a given value $0 < \epsilon < 1$ such that $\mathcal{R}(\mathcal{T}^D_{m,s}; F_{XY}) \geq \epsilon$. Then the probability that the remaining $n - D$ i.i.d. samples $\{(X_{D+1}Y_{D+1}), \ldots, (X_n, Y_n)\}$ are consistent with $\mathcal{T}^D_{m,s}$ by chance is lower than $\prod_{i=D+1}^{n} P\left(Y_i \in \mathcal{T}^D_{m,s}(X_i)\right) \leq (1-\epsilon)^{n-D}$. This can be bounded as follows

$$P\left(\mathcal{R}(\mathcal{T}^D_{m,s}; F_{XY}) \geq \epsilon\right) \leq (1-\epsilon)^{n-D} \leq e^{-(n-D)\epsilon}, \quad (6)$$

making use of the classical binomial bound, see e.g. [2]. The finite number of tubes which can be compressed without loss of information to D points can be bounded using a geometrical argument. Given D points, every point can be used to interpolate either the upper-function $m + s$, or the lower-function $m - s$. However, switching the assignments of all points simultaneously leads to the same ST, and the case of all points assigned to the same (upper- or lower-) function does not result in a unique tube neither. Therefor, the number of ST which can be determined using D samples out of n - denoted as $K_{n,D}(\Gamma)$ - can be bounded as follows:

$$\begin{aligned} K_{n,D}(\Gamma) &= \binom{n}{D}(2^{D-1} - 1) \\ &\leq \left(\frac{ne}{D}\right)^D (2^{D-1} - 1) \leq \left(\frac{2ne}{D}\right)^D \end{aligned} \quad (7)$$

where the inequality $\binom{n}{D} \leq (\frac{ne}{D})^D$ of the binomial coefficient is used. Combining (6) and (5), and solving the statement for δ as classical proofs the result. ∎

A crucial element for this result is that it is known a priori that such a tube with zero empirical risk exists independently from the data at hand (realizable case), this assumption is fulfilled by construction. Although combinatorial in nature (any found hypothesis Γ should be determined entirely by a subset of D chosen examples), it is shown in the next section how this property holds for a simple estimator which can be estimated efficiently as a standard linear program.

Example 1 (Confidence Level) *The following example indicates the practical use of this result: given $n = 200$ i.i.d. samples with a corresponding class of hypotheses each determined by three samples ($D = 3$ and thus $K_{n,D}(\Gamma) \leq 3 * 10^8$). Fixing the tolerance level as $\delta = 95\%$, one can state that the true risk will not be higher than 0.1049. This result can be used in practice as follows. Given an observed set of i.i.d. samples $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^{200} \subset \mathbb{R} \times \mathbb{R}$, compute the tube $\widehat{\mathcal{T}_{m,s}} = \{\hat{w}x \pm \hat{t} \ \forall x\}$ with $\hat{t} > 0$, $w \in \mathbb{R}$ and $\mathcal{R}_n(\widehat{\mathcal{T}_{m,s}}; \mathcal{D}_n) = 0$. When a new sample $X_j \in \mathbb{R}$ arrives, then predict that the corresponding $Y_j \in \mathbb{R}$ will lie in the interval $\hat{w}X_j \pm \hat{t}$. Then we are reasonably sure (with a probability of 0.95) that this assertion will hold in at least 89.51% of the cases when the number n_v of samples of data $\{X_j\}_{j=1}^{n_v}$ goes to infinity.*

A similar result can be obtained using the classical theory of non-parametric tolerance intervals, as initiated in [15], see e.g. [16].

Corollary 1 (Bound by Order Statistics) *Fix $\epsilon > 0$ and let $\mathcal{D}_n$ contain $n > 0$ i.i.d. samples from a fixed but unknown joint distribution F_{XY}. Consider the class of tubes Γ where each tube $\mathcal{T}_{m,s}$ is uniquely determined by D appropriate samples. Then, with probability higher than $1 - \delta < 1$, the following inequality holds:*

$$\begin{aligned} P\left(\sup_{\mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n)=0} \mathcal{R}(\mathcal{T}_{m,s}; F_{XY}) \geq \epsilon\right) \\ \leq K_{n,D}(\Gamma)\left(n(1-\epsilon)^{n-1} - (n-1)(1-\epsilon)^n\right), \end{aligned} \quad (8)$$

where $K_{n,D}(\Gamma)$ is defined as in Theorem 1.

Proof: Consider at first a fixed tube $\mathcal{T}^*_{m,s}$. After projecting all samples $\{(X_i, Y_i)\}_{i=1}^n$ to the univariate sample $R_i = m(X_i) - Y_i$, it is clear that a minimal tube with fixed m will have borders $\min(R_i)$ and $\max(R_i)$. Note that now $P(R \notin [\min(R_i), \max(R_i)])$ equals $\mathcal{R}(\mathcal{T}^*_{m,s}; F_{XY})$. Application of the standard results as in [15] for such tolerance intervals gives

$$\begin{aligned} P\Big(P(R \notin [\min(R_i), \max(R_i)]) \geq \epsilon\Big) \\ \leq n(1-\epsilon)^{n-1} - (n-1)(1-\epsilon)^n \end{aligned} \quad (9)$$

Application of the union bound over all hypothesis Γ as in (5) gives the result. ∎

Remark that this bound is qualitatively very similar to the previous one. As an interesting aside, the previous result implies a generalization bound on the minimal convex hull, i.e. a bound on the probability mass contained in the minimal Convex Hull (CH) of an i.i.d. sample. We consider the planar case, the extension to higher dimensional case follows straightforwardly. Formally, one may define the minimal planar convex hull $\mathrm{CH}(\mathcal{D}_n)$ of a sample $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ as the minimal subset of $\mathbb{R} \times \mathbb{R}$ containing all samples $(X_i, Y_i) \in \mathbb{R} \times \mathbb{R}$, and all convex combinations of any set of samples.

Theorem 2 (Probability Mass of the Convex Hull) *Fix $\delta > 0$ and let $\mathcal{D}_n$ contain i.i.d. samples of a random variable $(X, Y) \subset \mathbb{R} \times \mathbb{R}$. Then with probability exceeding $1 - \delta$, the probability mass outside the minimal convex hull $\mathrm{CH}(\mathcal{D}_n)$ is bounded as follows*

$$P\left((X, Y) \notin \mathrm{CH}(\mathcal{D}_n)\right) \leq \frac{3\log(n) - 1.5122 + \log(\frac{1}{\delta})}{n - 3}. \quad (10)$$

Proof: The key element of the proof is found in the fact that the CH is the intersection of all linear support tubes in Γ with minimal (constant) width having zero empirical risk. Let $\#\mathrm{CH}(\mathcal{D}_n)$ denote this intersection, formally,

$$\begin{aligned} &(X, Y) \in \#\mathrm{CH}(\mathcal{D}_n) \\ &\quad \Leftrightarrow Y \in \mathcal{T}_{m,s}(X), \quad \forall \mathcal{T}_{m,s} : \quad \mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n) = 0. \end{aligned} \quad (11)$$

Now we proof that $\#\mathrm{CH}(\mathcal{D}_n) = \mathrm{CH}(\mathcal{D}_n)$. On the hand, by definition $\#\mathrm{CH}(\mathcal{D}_n)$ is a convex set that contains all the points, and therefore it contains their convex hull.

Conversely, assume that $\mathrm{CH}(\mathcal{D}_n) \subset \#\mathrm{CH}(\mathcal{D}_n)$, then a point $(X, Y) \in \#\mathrm{CH}(\mathcal{D}_n)$ exist where $(X, Y) \notin \mathrm{CH}(\mathcal{D}_n)$,

and the point (X,Y) is included in all tubes $\mathcal{T}_{m,s}$ having $\mathcal{R}_n(\mathcal{T}_{m,s};\mathcal{D}_n) = 0$. By definition of the convex hull $(X,Y) \notin \mathcal{D}_n$, neither can it be a convex combination of any set of samples. Now, by the supporting hyperplane theorem (see e.g. [17]), there exists a linear hyperplane separating this point from the minimal convex hull. Constructing a tube $\mathcal{T}_{m,s}$ where $m+s$ equals this supporting plane, and with width large enough such that $\mathcal{R}_n(\mathcal{T}_{m,s};\mathcal{D}_n) = 0$ contradicts the assumption, proving the result.

Now, note that by definition the following inequality holds

$$P\Big((X,Y) \notin \#\mathrm{CH}(\mathcal{D}_n)\Big) \leq \sup_{\mathcal{R}_n(\mathcal{T}_{m,s};\mathcal{D}_n)=0} \mathcal{R}(\mathcal{T}_{m,s};F_{XY}). \tag{12}$$

Moreover, the set of linear tubes in $\mathbb{R}^2$ with fixed width can be characterized by a set containing exactly $D=3$ samples as proven in the following section. Finally, specializing the result of Theorem 1 in (4) gives the result. ∎

Note that classically the expected probability mass of a CH is expressed in terms of the expected number of extremal points of the data cloud [18]. Interestingly, the literature on statistical learning studies the number of extreme points in estimators as an (empirical) measure of complexity of an hypothesis space, note e.g. the correspondence between Theorem 12 in [1] and Theorem 2 in [18], and the coding interpretation of SVMs, see e.g. [1], [13], [14]. A disadvantage of the mentioned approach appears that the expected number of extremal points of the convex hull is a quantity which is difficult to characterize a priori (without seeing the data), without presuming restrictions on the underlying distribution [2]. The key observation of the previous theorem is that this number can be bounded by decomposing the minimal convex hull as the intersection of a set of linear tubes.

C. Support Tubes and Mutual Information

At first, a technical proposition is proven which will provide the cornerstone for the main result of the paper as stated below.

Proposition 1 (Upper-bound to the Conditional Entropy) *Let $\mathcal{T}_{m,s} : \mathbb{R}^d \to V \subset \mathbb{R}$ be a fixed tube, then one has*

$$H(Y|(X,Y) \in \mathcal{T}_{m,s}(X)) \leq \mathbb{E}[\log(2s(X))]. \tag{13}$$

Proof: For a fixed $x \in \mathbb{R}^d$ it holds that

$$H(Y|Y \in \mathcal{T}_{m,s}(x)) \leq \log(2s(x)) \tag{14}$$

following the fact that the uniform distribution has maximal entropy over all distributions in with fixed interval $2s(x)$. The conditional distribution is then defined as follows

$$\begin{aligned} &H(Y|(X,Y) \in \mathcal{T}_{m,s}(X)) \\ &\quad = \int H(Y|X=x, Y \in \mathcal{T}_{m,s}(x))\ dF_X(x) \\ &\quad \leq \int \log(2s(x))\ dF_X(x), \end{aligned} \tag{15}$$

hereby proving the result. ∎

In the case $\mathcal{H}_2\{s=t, t \in \mathbb{R}_0^+\}$, one has $H(Y|(X,Y) \in \mathcal{T}_{m,s}(X)) \leq \log(2t)$. The motivation for the analysis of the support tube is found in the following upper-bound to the mutual information based on a finite sample.

Theorem 3 (Lower-bound to the Mutual Information) *Fix $\delta > 0$, given an hypothesis class of tubes $\Gamma(\mathcal{H}_1,\mathcal{H}_2)$ and a set $\mathcal{D}_n$ containing n i.i.d. samples from F_{XY}. Let $c(\delta,D,n)$ as in equation (4) and assume that one has $c(\delta,D,n) < 0.5$. The following lower bound on the expected mutual information $I(Y|X)$ holds with probability exceeding $1-\delta$ that*

$$\begin{aligned} H(Y|X) \leq\ & c(\delta,D,n)H(Y) + \\ & (1-c(\delta,D,n))\mathbb{E}[\log(2s(X))] + h(c(\delta,D,n)), \end{aligned} \tag{16}$$

where $\mathbb{E}[\log(2s(X))]$ depends on F_{XY}, and equivalently

$$\begin{aligned} I(Y|X) \geq\ & (1-c(\delta,D,n))\Big(H(Y) \\ & - \mathbb{E}\left[\log(2s(X))\right]\Big) - h(c(\delta,D,n)), \end{aligned} \tag{17}$$

where F_X denotes the marginal distribution of X and $h(z)$ is the entropy of a Bernoulli random variable with parameter z.

Proof: The proof of this inequality follows roughly the derivation of Fano's inequality as in e.g. [7]. Let the random variable $U = g(X,Y,\mathcal{T}_{m,s}) \in \{0,1\}$ be defined as $U = \mathbb{I}(Y \notin \mathcal{T}_{m,s}(X))$ with n i.i.d. samples $\{U_i = \mathbb{I}(Y_i \notin \mathcal{T}_{m,s}(X_i))\}_{i=1}^n$. Twice the application of the chain rule on the conditional entropy gives

$$H(U,Y|X) = H(Y|X) + H(U|X,Y) = H(Y|X) \tag{18}$$

$$\begin{aligned} H(Y,U|X) &= H(U|X) + H(Y|U,X) \\ &\leq H(U) + H(Y|U,X), \end{aligned} \tag{19}$$

since U is a function of X and Y, the conditional entropy $H(U|X,Y) = 0$, and $H(U|X) \leq H(U)$. Theorem 1 states that for $\mathcal{T}_{m,s}$ with zero empirical risk, the actual risk satisfies $\mathbb{E}[U] = \mathcal{R}(\mathcal{T}_{m,s};F_{XY}) \leq c(\delta,D,n)$ with probability higher than $1-\delta$, such that the quantity $H(U)$ can be bounded with the same probability as

$$\begin{aligned} H(U) \leq\ & -c(\delta,D,n)\log(c(\delta,D,n)) \\ & -(1-c(\delta,D,n))\log(1-c(\delta,D,n)) \triangleq h(c(\delta,D,n)), \end{aligned} \tag{20}$$

because the entropy of a binomial variable is concave with maximum at 0.5 and $0 < \epsilon(\delta,D,n) < 0.5$ by assumption, see e.g. [7].

Now, the second term of the r.h.s. of (19) is considered. Note first that since $H(Y) \geq H(Y|X,U=0)$, it holds for all $0 < a < c(\delta,D,n) \leq 0.5$ that

$$\begin{aligned} &aH(Y) + (1-a)H(Y|X,U=0) \\ &\leq c(\delta,D,n)H(Y) + (1-c(\delta,D,n))H(Y|X,U=0). \end{aligned} \tag{21}$$

Hence,

$$\begin{aligned} H(Y|U,X) &= P(U=1)H(Y|X,U=1) \\ &\quad + P(U=0)H(Y|X,U=0) \\ &\leq P(U=1)H(Y) + P(U=0)H(Y|X,U=0) \qquad (22) \\ &\leq c(\delta,D,n)H(Y) + (1-c(\delta,D,n))H(Y|X,U=0) \\ &\leq c(\delta,D,n)H(Y) + (1-c(\delta,D,n))\mathbb{E}[\log(2s(X))], \qquad (23) \end{aligned}$$

where the first inequality follows from $H(Y|X,U=1) \leq H(Y)$, and the second one from (21) and since $P(U=1) < c(\delta,D,n)$. The third inequality constitutes the core of the proof, following from the previous Lemma. Combining this inequality with (20) and the definition of mutual information, $I(Y|X) = H(Y) - H(Y|X)$ yields inequality (17). ∎

In the case of the class of tubes with constant nonzero width $2t \in \mathbb{R}_0^+$, the inequality can be written as follows. With probability higher than $1-\delta$, the following lower-bound holds

$$I(Y|X) \geq (1-c(\delta,D,n))\,(H(Y) - \log(2t)) - h(c(\delta,D,n)), \qquad (24)$$

if $c(\delta,D,n) < 0.5$. Maximizing this lower-bound can be done by minimizing the width t and maximizing the probability of covering $(1-c(\delta,D,n))$, since the unconditional entropy is fixed.

From definition 1, it follows that a ST is not uniquely defined for a fixed F_{XY}. From the above derivation, a natural choice is to look for the most informative (and hence the least conservative) support tube as follows

$$\mathcal{T}^*_{m,s} = \underset{\mathcal{T}_{m,s}\in\Gamma(\mathcal{H}_1,\mathcal{H}_2)}{\arg\min} \|s\|_{\mathcal{H}_2} \text{ s.t. } \mathcal{T}_{m,s} \text{ is a ST to } F_{XY}. \qquad (25)$$

where $\|\cdot\|_{\mathcal{H}_2}$ denotes an appropriate (pseudo-) norm on the hypothesis space $\mathcal{H}_2$, proportional to the term $\mathbb{E}[\log 2s(X)]$ of equation (17). Let the theoretical risk of a ST on F_{XY} be defined as $\mathcal{R}(\mathcal{T}_{m,s}, F_{XY}) = \int P\left(Y \notin \mathcal{T}_{m,s}(x) \mid X = x\right) dF_X$. Given only a finite number of observations in $\mathcal{D}_n$, the empirical counterpart is studied

$$\widehat{\mathcal{T}_{m,s}} = \underset{\mathcal{T}_{m,s}\in\Gamma(\mathcal{H}_1,\mathcal{H}_2)}{\arg\min} \|s\|_{\mathcal{H}_2}, \text{ s.t. } \mathcal{R}_n(\mathcal{T}_{m,s};\mathcal{D}_n) = 0. \qquad (26)$$

and the above concentration results provide a guarantee that one will find an empirical solution which has also good theoretical properties.

D. Tolerance Tubes

The discussion can be extended to the case of tolerance tubes with a tolerance level of $0 < \alpha < 1$. Assume we have an estimator which for a sample $\mathcal{D}_n$ returns a tube $\widehat{\mathcal{T}_{m,s}}$ specified by exactly D samples such that at most $\lceil \alpha n \rceil$ samples violate the tube. The question how well this estimator behaves for novel samples is considered. Specifically, we bound the expected occurrence of a sample not contained in the tube $\widehat{\mathcal{T}_{m,s}}$ as follows using Hoeffding's inequality as classical.

Proposition 2 (Deviation Inequality for Tolerance Tubes) *Fix $\delta > 0$, let $\mathcal{D}_n$ contain n i.i.d. samples, and any hypothesis $\mathcal{T}_{m,s}$ can be represented (compressed) by exactly D samples, one has with probability exceeding $1-\delta$ that*

$$\mathcal{R}(\widehat{\mathcal{T}_{m,s}}; F_{XY}) \leq \mathcal{R}_n(\widehat{\mathcal{T}_{m,s}}; \mathcal{D}_n) + \sqrt{\frac{2D\log(\frac{2ne}{D}) + 2\log\left(\frac{8}{\delta}\right)}{n}}. \qquad (27)$$

This proof follows straightforwardly from the classical inequality for uniform convergence with $K_{n,D}(\Gamma) \leq \left(\frac{2ne}{D}\right)^D$ different hypotheses, see e.g. [1], [13] or [2]. It is a straightforward exercise to use this result to derive a bound on the mutual information in the case of tolerance tubes as previously.

III. Linear Support/Tolerance Vector Tubes

Motivated by the above derivations, this section elaborates on a practical estimator of both the support and tolerance tubes. Here we restrict ourselves to the linear model class $\mathcal{H}_1 = \{m : m(x) = x^T w \mid w \in \mathbb{R}^d\}$ and the class of parallel tubes $\mathcal{H}_2 = \{s : \ s(x) = t, \ t \in \mathbb{R}^+\}$ with constant width for clarity of explanation. Problem (26) with $\Gamma(\mathbb{R}^d, \mathbb{R}^+)$ can be casted as a linear programming problem as follows,

$$(\hat{w}, \hat{t}) = \underset{w,t>0}{\arg\min} \ t \text{ s.t. } -t \leq Y_i - w^T X_i \leq t \ \forall i = 1,\ldots,n. \qquad (28)$$

The more general case of TT requires an additional step, specifically, one has to be able to impose the a-priori property that one will end up with an estimate indeed capturing all but $\lceil \alpha n \rceil$ samples. This is guaranteed by imposing a proper term in the linear program with weighting term C which is to be set as $C = \lceil \alpha n \rceil$ for given $\alpha \geq 0$.

Lemma 1 (Tolerance Vector Tubes) *The following estimator (strictly) excludes at most $C = \lceil \alpha n \rceil$ observations (*quantile property*), while the functions $w^T x - t$ and $w^T x + t$ interpolate at least $d+1$ sample points (*Compression property*). If the underlying distribution F_{XY} is Lebesgue smooth and non-degenerate (hence no linear dependence between the variables and the vector of ones occur), exactly $d+1$ points are interpolated with probability 1.*

$$\begin{aligned} (\widehat{\mathcal{T}_{w,t}}, \xi_i) &= \underset{w,t,\xi_i}{\arg\min}\, \mathcal{J}_C(t,\xi_i) = Ct + \sum_{i=1}^n \xi_i \\ \text{s.t. } & -t - \xi_i \leq w^T X_i - Y_i \leq t + \xi_i, \xi_i \geq 0 \ \ \forall i = 1,\ldots,n. \end{aligned} \qquad (29)$$

*Moreover, the observations which satisfy the inequality constraints exactly determine the solution completely (*representer property*), hereby justifying the name of Support/Tolerance Vector Tubes in analogy with the nomenclature in support vector machines.*

Proof: The quantile property is proven as follows. Let $\alpha_i^+, \alpha_i^- \in \mathbb{R}^+$ be positive Lagrange multipliers $\forall i = 1,\ldots,n$. The Lagrangian of the constrained problem (33) becomes $\mathcal{L}_C(w,t,\xi_i;\alpha^+,\alpha^-,\beta) = \mathcal{J}_C(w,t,\xi_i) - \sum_{i=1}^n \beta_i\xi_i - \sum_{i=1}^n \alpha_i^+ \left(w^T X_i - Y_i + t + \xi_i\right)$

$-\sum_{i=1}^{n}\alpha_i^-\left(Y_i - w^T X_i + t + \xi_i\right)$. The first order conditions for optimality become

$$\begin{cases} \dfrac{\partial \mathcal{L}_C}{\partial t} = 0 \rightarrow C = \sum_{i=1}^{n}(\alpha_i^+ + \alpha_i^-) & (a) \\ \dfrac{\partial \mathcal{L}_C}{\partial w} = 0 \rightarrow 0_n = \sum_{i=1}^{n}\left(\alpha_i^- - \alpha_i^+\right)X_i & (b) \\ \dfrac{\partial \mathcal{L}_C}{\partial \xi_i} = 0 \rightarrow 1 = (\alpha_i^+ + \alpha_i^-) + \beta_i. & (c) \end{cases} \tag{30}$$

Following the complementary slackness conditions ($\beta_i\xi_i = 0\ \forall i = 1,\ldots,n$), if follows that $\beta_i = 0$ for data-points outside the tube ($\xi_i > 0$). This together with condition (30.a) and (30.c) proofs the quantile property.

The compression property follows from the fundamental lemma of a linear programming problem: the solution to the problem satisfies at least $d+1+n$ inequality constraints with equality. If $\hat{t} \neq 0$, then at least $d+1$ constraints $\xi_i = 0$ should be satisfied as at most n constraints of the $2n$ inequalities of the form $-t - \xi_i \leq (w^T X_i - Y_i)$ and $(w^T X_i - Y_i) \leq t + \xi_i$ can hold at the same time. If $\hat{t} = 0$, the problem reduces to the classical least absolute deviation estimator, possessing the above property. Let $\mathbf{x} = (X_1,\ldots,X_n)^T \in \mathbb{R}^{n\times d}$ be a matrix and $\mathbf{y} = (Y_1,\ldots,Y_n)^T \in \mathbb{R}^n$ be a vector. If the matrix $(1_N, \mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n\times(1+d+1)}$ is nonsingular (F_{XY} is non-degenerate) the solution to the problem (33) satisfies exactly $n+d+1$ inequalities, and any two functions $\{w^T x - t,\ w^T x + t\ \forall x \in \mathbb{R}^d\}$ can at most (geometrically) interpolate $d+1$ linear independent points.

Since a solution interpolates $d+1$ (linear independent) points exactly under the above conditions, knowledge of which points - say $\mathcal{S} \subset \{1,\ldots,n\}$ - implies the optimal solution $\hat{w}$ and $\hat{t}$ as

$$w^T X_i \pm t = Y_i, \quad \forall i \in \mathcal{S}, \tag{31}$$

where $\pm t$ denotes whether the specific sample interpolates the upper- or lower function. This means that the solution can be represented as the set $\mathcal{S}$ together with a one-bit flag indicating the sign. To represent the solution, one as such needs $(d+1)(\ln(n)+1)$ bits. The probability mass inside the tube is given by the value C which is known a priori. ∎

Note that a similar computational principle lies at the heart of the derivation of the ν-SVM [19]. The compression property is unlike the classical representer theorems for kernel machines, as no regularization term (e.g. $\|w\|$) occurs in the estimator. In the case of $C \rightarrow 0$, the estimator (33) results in the smallest support tube. When $C \rightarrow +\infty$, the robust L_1 norm is obtained [20], and when C is such that $t = \epsilon$, the ϵ-loss of the SVR is implemented. One has to keep in mind however that despite those computational analogies, the scope of interval estimation differentiates substantially from the L_1 and the SVR point predictors.

We now turn to the computationally more challenging task of estimating multiple condition tolerance intervals at the same time.

Proposition 3 (Multi-Tolerance Vector Tubes) *Consider the set of tubes with levels $(1-C_1,\ldots,1-C_l)^T \in \mathbb{R}_+^l$ for $l > 0$ defined as*

$$\mathcal{T}_{m,s}^{(m)} = \left\{ \mathcal{T}_{m,s}^{l} = \left[w^T x - \sum_{k=1}^{l} t_k^-, \quad w^T x + \sum_{k=1}^{l} t_k^+ \right] \right\}_{l=1}^{m} \tag{32}$$

where $m(x) = w^T x$. The parameters $w \in \mathbb{R}^d$, $t^+ = (t_0,\ldots,t_m)^T \in \mathbb{R}^{m+1}$ and $t^- = (t_0^-,\ldots,t_m^-)^T \in \mathbb{R}^{m+1}$ can be found by solving the following convex programming (LP) problem

$$\begin{aligned} \min_{w,t^+,t^-,\xi_i^m} \mathcal{J}_C(t^+,t^-,\xi_i^m) &= \sum_{l=1}^{m} C_l(t_l^+ + t_l^-) + \sum_{l=1}^{m}\sum_{i=1}^{n}(\xi_i^{l+} + \xi_i^{l-}) \\ \text{s.t.} \quad & \begin{cases} -\xi_i^{-l} - t_l^- \leq (w^T X_i - Y_i) \leq t_l^+ + \xi_i^{+l} \\ 0 \leq \xi_i^{l+}, \xi_i^{l-}, \quad 0 \leq t_l^+, t_l^- \end{cases} \\ & \forall l = 1,\ldots,m, \quad \forall i = 1,\ldots,n. \end{aligned} \tag{33}$$

*Then every solution excludes at most C_l datapoints (*generalized quantile property*), while the boundaries of all tubes pass through at most $d + 2(m+1)$ datapoints (compression property).*

Proof: The proof follows exactly the same lines as in Lemma 2, employing the fundamental theorem of linear programming and the first order conditions of optimality. Note that by construction, the different tubes are properly nested, i.e. not allowed to cross. ∎

Realizing that the result of Proposition 2 holds simultaneously for all tubes which can be compressed to D samples, the result can be extended to the case of multi-tolerance tubes as follows.

Corollary 2 (Multi-Tolerance Tubes) *Fix $\delta > 0$, let $\mathcal{D}_n$ contain n i.i.d. samples, and let our algorithm find a multi-tolerance vector tubes with m levels $\widehat{\mathcal{T}_{m,s}}$ which can be compressed to $(d+2(m+1))$ samples, one has with probability exceeding $1-\delta$ that for any $l = 1,\ldots,m$*

$$\begin{aligned} &\mathcal{R}(\widehat{\mathcal{T}_{m,s}^{l}}; F_{XY}) \leq \mathcal{R}_n(\widehat{\mathcal{T}_{m,s}^{l}}; \mathcal{D}_n) \\ &+ \sqrt{\frac{2(d+2(m+1))\log\left(\frac{2ne}{(d+2(m+1))}\right) + 2\log\left(\frac{8}{\delta}\right)}{n}}. \end{aligned} \tag{34}$$

Here we see that one pays a small price in terms of m, a term which originates from the fact that we have now a compression coefficient of $d+2(m+1)$ replacing the term $d+1$ before. Figure 2 gives an example of such a multi-tolerance tube with a nonlinear function m which is a linear combination of localized basis-functions. This computational mechanism of inferring and representing the empirically optimal tube $\widehat{\mathcal{T}_{m,s}}$ can be extended to data represented in a more complex metric (e.g. $x \subset \mathbb{R}^d$ where $d \rightarrow \infty$ or by using reproducing kernels). Hereto, it is seen that one needs another mechanism of restricting the hypothesis space $\mathcal{H}_1$. Consider for example the class $\mathcal{H}_{1\rho} = \{m(x) = w^T x\ |\|w\|_2^2 \leq \rho\}$, having a finite covering number (see e.g. [1]).

IV. Conclusion

This paper studied a simple estimator of the conditional support and tolerance intervals of a distribution. The result is

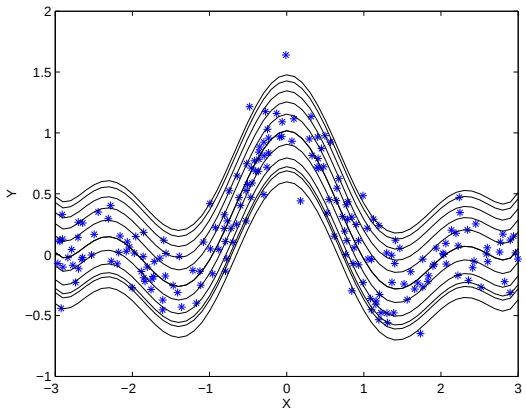


Fig. 2. *Example of $n = 250$ a Multi-Tolerance Vector Tube $\mathcal{T}^{(6)}_{m,s}$ with $\alpha = \lfloor (2, 5, 10, 25, 50)/n \rfloor$. Here the function m consists of a linear combination of 10 localized basis-functions.*

shown to yield a useful lowerbound to the mutual information of the sample by extending Fano's theorem in combination with standard results of learning theory.

REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*. Wiley and Sons, 1998.

[2] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[3] R. Koenker, *Quantile Regression*, ser. Econometric Society Monograph Series. Cambridge University Press, 2005.

[4] I. Takeuchi, Q. Le, T. Sears, and A. Smola, "Nonparametric Quantile Estimation," *The Journal of Machine Learning Research*, vol. 7, pp. 1231–1264, 2006.

[5] A. Christmann and I. Steinwart, "Consistency of kernel-based quantile regression," *Applied Stochastic Models in Business and Industry*, vol. 24, no. 2, pp. 171–183, 2008.

[6] J. Hartigan, "Estimation of a convex density contour in two dimensions," *J. Amer. Statist. Assoc*, vol. 82, no. 397, pp. 267–270, 1987.

[7] T. Cover and J. A. Thomas, *Elements of Information Theory*. Springer, 1991.

[8] M. Feder and N. Merhav, "Relations between entropy and error probability," *Information Theory, IEEE Transactions on*, vol. 40, no. 1, pp. 259–266, 1994.

[9] D. Guo, S. Shamai, and S. Verdu, "Mutual information and minimum mean-square error in gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261– 1282, 2005.

[10] J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[11] L. Paninski, "Estimation of Entropy and Mutual Information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[12] N. Littlestone and M. Warmuth, "Relating data compression and learnability," *Technical Report University of California, Santa-Cruz*, 1986.

[13] S. Floyd and M. Warmuth, "Sample compression, learnability and the VC dimension," *Machine Learning*, vol. 21, no. 3, pp. 269–304, 1995.

[14] U. von Luxburg, O. Bousquet, and B. Schölkopf, "A compression approach to support vector model selection," *Journal of Machine Learning Research*, vol. 5, pp. 293–323, 2004.

[15] S. Wilks, "Determination of sample sizes for setting tolerance limits," *The Annals of Mathematical Statistics*, vol. 12, no. 1, pp. 91–96, 1941.

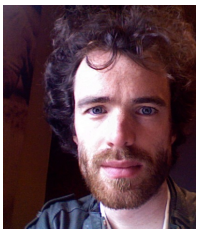
[16] J. Rice, *Mathematical statistics and data analysis*. Pacific Grove, California: Duxbury Press, 1988.

[17] R. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[18] B. Efron, "The convex hull of a random set of points," *Biometrika*, vol. 52, pp. 331–343, 1965.

[19] B. Schölkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.

[20] P. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*. Wiley & sons, 1986.



Kristiaan Pelckmans Kristiaan Pelckmans was born in Turnhout, Belgium in 1978. He received a masters degree in Computer Science in 2000, and a Ph.D. in Applied Science in 2005 in K.U.Leuven. His dissertation entitled 'Primal-Dual Kernel Machines' concerns theory and practice of machine learning, optimization and inference. He has been a postdoctoral researcher at ESAT- SISTA/SCD since 2005, and was granted a FWO Research Fellowship 2007-2009. Under this grant he has been visiting amongst other Darmstadt, Fh.G in 2005, and University College London, UK (2007-2008). He is currently a research Professor (forskarsassistent) in Uppsala University, Sweden, Department of Information Technology, Division of Systems and Control. He is the (co-) author of 10 Journal publications and 33 Peer reviewed conference publications.



Jos De Brabanter Jos De Brabanter was born in Ninove Belgium, January 11 1957. He received the degree in Electronical Engineering in 1990, Safety Engineer in 1992, Master of Environment, Human Ecology in 1993, Master in Artificial Intelligence in 1996, Master of Statistics in 1997 and the Ph.D. degree in Applied Sciences (Katholieke Universiteit Leuven) in 2004. He currently holds an associated docent position at the K.U.Leuven, Belgium. His research mainly focuses on Statistics and Nonlinear systems.



Johan A.K. Suykens Johan A.K. Suykens was born in Willebroek Belgium, May 18 1966. He received the degree in Electro-Mechanical Engineering and the Ph.D. degree in Applied Sciences from the Katholieke Universiteit Leuven, in 1989 and 1995, respectively. In 1996 he has been a Visiting Postdoctoral Researcher at the University of California, Berkeley. He has been a Postdoctoral Researcher with the Fund for Scientific Research FWO Flanders and is currently a Professor (Hoogleraar) with K.U.Leuven. He is author of the books "Artificial Neural Networks for Modelling and Control of Non-linear Systems" (Kluwer Academic Publishers) and "Least Squares Support Vector Machines" (World Scientific), co-author of the book "Cellular Neural Networks, Multi-Scroll Chaos and Synchronization" (World Scientific) and editor of the books "Nonlinear Modeling: Advanced Black-Box Techniques" (Kluwer Academic Publishers) and "Advances in Learning Theory: Methods, Models and Applications" (IOS Press). He is a Senior IEEE member and has served as associate editor for the IEEE Transactions on Circuits and Systems (1997-1999 and 2004-2007) and since 1998 he is serving as associate editor for the IEEE Transactions on Neural Networks. He received an IEEE Signal Processing Society 1999 Best Paper (Senior) Award and several Best Paper Awards at International Conferences. He is a recipient of the International Neural Networks Society INNS 2000 Young Investigator Award for significant contributions in the field of neural networks. He has served as a Director and Organizer of the NATO Advanced Study Institute on Learning Theory and Practice (Leuven 2002), as a program co-chair for the International Joint Conference on Neural Networks 2004 and the International Symposium on Nonlinear Theory and its Applications 2005, and as an organizer of the International Symposium on Synchronization in Complex Networks 2007.



Bart De Moor Bart De Moor was born 1960 in Halle, Belgium. He obtained his Master (Engineering) Degree in Electrical Engineering at the Katholieke Universiteit Leuven, Belgium, and a PhD in Engineering at the same university in 1988. He spent 2 years as a Visiting Research Associate at Stanford University (1988-1990) at the departments of EE (ISL, Prof. Kailath) and CS (Prof.Golub). Currently, he is a full professor at the Department of Electrical Engineering (ESAT) of the K.U.Leuven in the research group SCD. Currently, he is leading a research group of 30 PhD students and 8 postdocs and in the recent past, 55 PhDs were obtained under his guidance. His work has won him several scientific awards (Leybold-Heraeus Prize (1986), Leslie Fox Prize (1989), Guillemin-Cauer best paper Award of the IEEE Transaction on Circuits and Systems (1990), Laureate of the Belgian Royal Academy of Sciences (1992), bi-annual Siemens Award (1994), best paper award of Automatica (IFAC, 1996), IEEE Signal Processing Society Best Paper Award (1999). Since 2004 he is a fellow of the IEEE. He is on the board of 6 spin-off companies(IPCOS, Data4s, TMLeuven, Silicos, Dsquare, Cartagenia), of the Flemish Interuniversity Institute for Biotechnology (VIB), the Study Center for Nuclear Energy (SCK), the Institute for Broad Band Technology (IBBT). He is also the Chairman of the Industrial Research Fund (IOF), Hercules (heavy equipment funding in Flanders,) and several other scientific and cultural organizations. He was a member of the Academic Council of the Katholieke Universiteit Leuven, and of its Research Policy Council. Full details on his CV can be found at www.esat.kuleuven.be/ demoor.