Support and Quantile Tubes

Kristiaan Pelckmans, Jos De Brabanter, Johan A.K. Suykens, and Bart De Moor. *

March 1, 2007

Abstract

This correspondence studies an estimator of the conditional support of a distribution underlying a set of i.i.d. observations. The relation with mutual information is shown via an extension of Fano's theorem in combination with a generalization bound based on a compression argument. Extensions to estimating the conditional quantile interval, and statistical guarantees on the minimal convex hull are given.

Keywords: - Statistical Learning, Fano's inequality, Mutual Information, Support Vector Machines

1 Introduction

Given a set of paired observations $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ which are i.i.d. copies of a random vector (X, Y) possessing a fixed but unknown joint distribution F_{XY}, this letter concerns the question which values the random variable Y can possibly/likely take given a covariate X. This investigation on predictive tolerance intervals is motivated as one is often interested in other characteristics of the joint distribution than the conditional expectation (regression): e.g. in econometrics one is often more interested in the volatility of a market than in its precise prediction. In environmental sciences one is typically concerned with the extremal behavior (i.e. the min or max value) of a magnitude, and its respective conditioning on related environmental variables.

The main contribution of this letter is the extension to Fano's classical inequality (see e.g. [1], p. 38) which gives a lower-bound to the mutual information of two random variables. This classical result is extended towards a setting of learning theory where random variables have an arbitrary

*Pelckmans *et al.* are with KULeuven-ESAT-SCD/sista, Kasteelpark Arenberg 10, Leuven - B-3001, Belgium

fixed distribution. The derivation yields a non-parametric estimator of the mutual information possessing a probabilistic guarantee which is derived using a classical compression argument. The described relationship differs from other results relating estimators and mutual information as e.g. using Fisher's information matrix [1] or based on Gaussian assumptions as e.g. in [2], as a distribution free context is adopted. As an aside, (i) an estimator of the conditional support is derived and is extended to the setting of conditional quantiles, (ii) its theoretical properties are derived, (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.

While studied in the literature e.g. on quantile regression [3], we argue that this question can be approached naturally from a setting of statistical learning theory, pattern recognition and Support Vector Machines (SVM), see [4, 5] for an overview. A main conceptual difference with the existing literature on classical regression and other predictor methods is that no attempt is made whatsoever to reveal an underlying conditional mean (as in regression), conditional quantile (as in quantile regression), or minimal risk point prediction of the dependent variable (as in pattern recognition). Here we target instead (the change of) the rough contour of the conditional distribution. This implies that one becomes interested in (i) to what extent the estimated conditional support of the tube is conservative (i.e. does it overestimate the actual conditional support?), and (ii) what is the probability of covering the actual conditional support (i.e. to what probability a new sample can occur outside the estimated interval).

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.

2 Support and Quantile Tubes

2.1 Support Tubes and Risk

Definition 1 (Support and Quantile 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 $\Gamma(\mathcal{H}_1, \mathcal{H}_2)$*

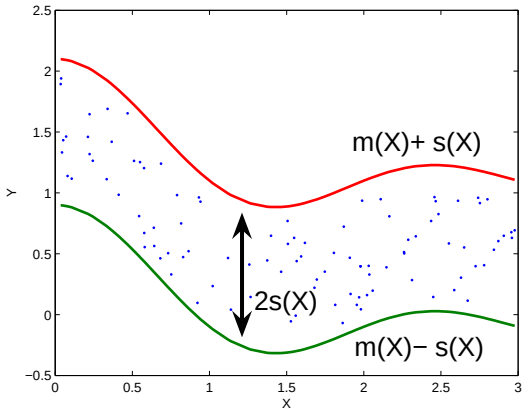

Figure 1: *Example of a support vector tube based on a finite sample of a bivariate random variable* (X, Y). *A tube* $\mathcal{T}_{m,s}$ *is defined as the conditional interval* $\mathcal{T}_{m,s}(X) = [m(X) - s(X),\ \ m(X) + s(X)]$ *with width* $2s(x)$.

is defined as

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

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 quantile tube (QT) 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 $\eta(\cdot; \Gamma(\mathcal{H}_1, \mathcal{H}_2)) : [0,1] \rightarrow [0,1)$ expresses the *confidence level* in the probability of covering.

2.2 Generalization Bound

For now, we focus on the case of the ST, extensions specific to the QT 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 [6], [7], and refined in various publications as e.g. [8].

Theorem 1 (Compression Bound on Risk of a ST) *Let $\mathcal{D}_n$ be i.i.d. sampled 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. Then, with probability exceeding $1-\delta < 1$, the following inequality holds for any $\mathcal{T}_{m,s}$ where $\mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n) = 0$:*

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

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: At first, fix 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 chance that the remaining $n - D$ i.i.d. samples $\{(X_{D+1}Y_{D+1}), \ldots, (X_n, Y_n)\}$ are by chance consistent with $\mathcal{T}^D_{m,s}$, 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. [5]. 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:

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

where the inequality $\binom{n}{D} \leq (\frac{ne}{D})^D$ of the binomial coefficient is used. Combining (6) and (5), and inverting the statement 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 (Tolerance 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 $\widehat{\mathcal{T}_{m,s}} = \hat{w}x \pm \hat{t}$ 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 [9], see e.g. [10].

Corollary 1 (Bound by Order Statistics) *Let $\mathcal{D}_n$ be 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 for any $\mathcal{T}_{m,s}$ where $\mathcal{R}_n(\mathcal{T}_{m,s};\mathcal{D}_n)=0$:*

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

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 [9] for such tolerance intervals gives

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

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 a most 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 Planar Convex Hull) *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 $1-\delta<1$, the probability mass outside the minimal convex hull $\mathrm{CH}(\mathcal{D}_n)$ is bounded as follows*

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

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,

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

Now we proof that $\#\mathrm{CH}(\mathcal{D}_n) = \mathrm{CH}(\mathcal{D}_n)$. Assume at first that $\#\mathrm{CH}(\mathcal{D}_n) \subset \mathrm{CH}(\mathcal{D}_n)$, then a point $(X,Y) \in \mathrm{CH}(\mathcal{D}_n)$ exists where $(X,Y) \notin \#\mathrm{CH}(\mathcal{D}_n)$, but this is in contradiction to the assertion that $\mathrm{CH}(\mathcal{D}_n)$ should be minimal: indeed also $\#\mathrm{CH}(\mathcal{D}_n)$ is convex (an intersection of convex sets), and contains all samples by construction.

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. [11]), 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

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

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 (9) 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 [12]. 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 [4] and Theorem 2 in [12], and the coding interpretation of SVMs, see e.g. [4, 7, 8]. 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 [5]. 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.

2.3 Support Tubes and Mutual Information

At first, a technical Lemma is proven which will play a major role in the main result of the paper stated below.

Lemma 1 (Upper-bound to the Conditional Entropy) *Let $\mathcal{T}_{m,s} : \mathbb{R}^d \rightarrow V \subset \mathbb{R}$ 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: The proof follows from the following inequality, 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 a fixed interval. The conditional distribution is then defined as follows

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

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) *Given an hypothesis class of tubes $\Gamma(\mathcal{H}_1, \mathcal{H}_2)$ and a set of i.i.d. samples $\mathcal{D}_n$. Let $\epsilon(\delta, D, n)$ as in equation (9) for a confidence exceeding $1-\delta < 1$, and assume that the corresponding probability of covering satisfies $\epsilon(\delta, D, n) < 0.5$. The following lower bound on the expected mutual information $I(Y|X)$ holds with probability exceeding $1 - \delta$*

$$H(Y|X) \leq \epsilon(\delta, D, n)H(Y) + (1 - \epsilon)\mathbb{E}[\log(2s(X))] \tag{15}$$

and equivalently

$$I(Y|X) \geq (1 - \epsilon(\delta, D, n))\Big(H(Y) - \mathbb{E}\left[\log(2s(X))\right]\Big) - h(\epsilon(\delta, D, n)), \tag{16}$$

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

Proof: The proof of this inequality follows roughly the derivation of Fano's inequality as in e.g. [1]. 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

$$\begin{aligned} H(U, Y|X) &= H(Y|X) + H(U|X, Y) = H(Y|X) && (17) \\ H(Y, U|X) &= H(U|X) + H(Y|U, X) \\ &\leq H(U) + H(Y|U, X), && (18) \end{aligned}$$

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 $\mathbb{E}[U] = \mathcal{R}(\mathcal{T}_{m,s}; F_{XY}) \leq \epsilon(\delta, D, n)$ with probability higher than $1 - \delta$, such that the quantity $H(U)$ can be bounded with the same probability as

$$H(U) \leq -\epsilon \log(\epsilon) - (1 - \epsilon) \log(1 - \epsilon) \triangleq h(\epsilon), \tag{19}$$

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. [1].

Now, the second term of the rhs of (18) is considered. Note first that since $H(Y) \geq H(Y|X, U = 0)$, it holds for all $0 < a < \epsilon(\delta, D, n) \leq 0.5$ that

$$\begin{aligned} &aH(Y) + (1 - a)H(Y|X, U = 0) \\ &\qquad \leq \epsilon H(Y) + (1 - \epsilon(\delta, D, n))H(Y|X, U = 0). \end{aligned} \tag{20}$$

Hence,

$$\begin{aligned} H(Y|U, X) &= P(U = 1)H(Y|X, U = 1) \\ &\quad + P(U = 0)H(Y|X, U = 0) \\ &\leq P(U = 1)H(Y) + P(U = 0)H(Y|X, U = 0) && (21) \\ &\leq \epsilon(\delta, D, n)H(Y) + (1 - \epsilon(\delta, D, n))H(Y|X, U = 0) \\ &\leq \epsilon(\delta, D, n)H(Y) + (1 - \epsilon(\delta, D, n))\mathbb{E}[\log(2s(X))], && (22) \end{aligned}$$

where the first inequality follows from $H(Y|X, U = 1) \leq H(Y)$, and the second one from (20) and since $P(U = 1) < \epsilon(\delta, D, n)$. The third inequality constitutes the core of the proof, following from the previous Lemma. Combining this inequality with (19) and the definition of mutual information, $I(Y|X) = H(Y) - H(Y|X)$ yields inequality (16).

□

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 $1-\delta < 1$, the following lower-bound holds

$$I(Y|X) \geq (1 - \epsilon(\delta, D, n)) \left(H(Y) - \log(2t)\right) - h(\epsilon(\delta, D, n)), \tag{23}$$

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

$$\mathcal{T}_{m,s}^* = \underset{\mathcal{T}_{m,s} \in \Gamma(\mathcal{H}_1, \mathcal{H}_2)}{\arg\min} \|s\| \text{ s.t. } \mathcal{T}_{m,s} \text{ is a ST to } F_{XY}. \tag{24}$$

where $\|\cdot\|$ denotes a (pseudo-) norm on the hypothesis space $\mathcal{H}_2$, proportional to the term $\mathbb{E}[\log 2s(X)]$ of equation (16). 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

$$\widehat{\mathcal{T}_{m,s}} = \underset{\mathcal{T}_{m,s} \in \Gamma(\mathcal{H}_1, \mathcal{H}_2)}{\arg\min} \|s\|_{\mathcal{H}_1} \text{ s.t. } \mathcal{R}_n(\mathcal{T}_{m,s}; \mathcal{D}_n) = 0. \tag{25}$$

2.4 Quantile Tubes

The discussion can be extended to the case of quantile tubes of a level $0 < \alpha \leq 1$. 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 1 (Deviation Inequality for Quantile Tubes) *When $\mathcal{D}_n$ contains n i.i.d. samples, and any hypothesis $\mathcal{T}_{m,s}$ can be represented by exactly D samples, one has with probability exceeding $1 - \delta < 1$, one has*

$$\mathcal{R}(\widehat{\mathcal{T}_{m,s}}; F_{XY}) - \alpha \leq \mathcal{R}_n(\widehat{\mathcal{T}_{m,s}}; \mathcal{D}_n) + 2\sqrt{\frac{2D \log(\frac{2ne}{D}) - 2\log\left(\frac{8}{\delta}\right)}{n}}. \tag{26}$$

This proof follows straightforwardly from the Vapnik and Chervonenkis inequality with $K_{n,D}(\Gamma) \leq \left(\frac{2ne}{D}\right)^D$ different hypotheses, see e.g. [4] or [5]. It is a straightforward exercise to use this result to derive a bound on the mutual information in the case of quantile tubes as previously.

3 Linear Support/Quantile Vector Tubes

Given the specified methodology, this section elaborates on a practical estimator and shows how to extend results to quantile 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 (25) with $\Gamma(\mathbb{R}^d, \mathbb{R}^+)$ can be casted as a linear programming problem as follows,

$$(\hat{w}, \hat{t}) = \underset{w,t>0}{\arg\min} \quad t \text{ s.t.} \quad -t \leq Y_i - w^T X_i \leq t \ \ \forall i = 1, \ldots, n. \tag{27}$$

The more general case of QT requires an additional step:

Lemma 2 (Quantile Vector Tubes) *The following estimator (strictly) excludes at most C observations (*quantile property*), while the functions $w^T x - t$ and $w^T x + t$ interpolate at least $d+1$ sample points (*interpolation 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.*

$$\begin{aligned}(\widehat{\mathcal{T}_{w,t}}, \xi_i) = \underset{w,t,\xi_i}{\arg\min}\, \mathcal{J}_C(t, \xi_i) = Ct + \sum_{i=1}^{n} \xi_i \\ s.t. \quad -t - \xi_i \leq w^T X_i - Y_i \leq t + \xi_i, \xi_i \geq 0 \ \ \forall i = 1, \ldots, n. \end{aligned} \tag{28}$$

*Moreover, the observations which satisfy the inequality constraints exactly determine the solution completely (*representer property*), hereby justifying the name of Support/Quantile 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 (32) 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

$$\begin{cases} \dfrac{\partial \mathcal{L}_C}{\partial t} = 0 \rightarrow C = \sum_{i=1}^{n} (\alpha_i^+ + \alpha_i^-) & (a) \\ \dfrac{\partial \mathcal{L}_C}{\partial w} = 0 \rightarrow 0_n = \sum_{i=1}^{n} \left(\alpha_i^- - \alpha_i^+\right) X_i & (b) \\ \dfrac{\partial \mathcal{L}_C}{\partial \xi_i} = 0 \rightarrow 1 = (\alpha_i^+ + \alpha_i^-) + \beta_i. & (c) \end{cases} \tag{29}$$

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 (29.a) and (29.c) proofs the quantile property.

The interpolation 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 $x = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times d}$ be a matrix and $y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ be a vector. If the matrix $(1_N, x, y) \in \mathbb{R}^{n \times (1+d+1)}$ is nonsingular (F_{XY} is non-degenerate) the solution to the problem (32) satisfies exactly $n + d + 1$ inequalities, and any two functions $\{w^T x - t, \; w^T x + t\}$ 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

$$w^T X_i \pm t = Y_i, \quad \forall i \in \mathcal{S}, \tag{30}$$

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 principle lies at the heart of the derivation of the ν-SVM [13]. The representer 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 (32) results in the smallest support tube. When $C \rightarrow +\infty$, the robust L_1 norm is obtained [14], 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 quantile intervals at the same time.

Proposition 2 (Multi-Quantile Vector Tubes) *Consider the set of tubes defined as*

$$\mathcal{T}_{m,s}^{(m)} = \left\{ \mathcal{T}_{m,s}^{l} = \left[w^T x - \sum_{k=1}^{l} t_k^-, \quad w^T x + \sum_{k=1}^{l} t_k^+ \right] \right\}_{l=1}^{m} \tag{31}$$

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

$$\min_{w, t^+, t^-, \xi_i^m} \mathcal{J}_C(t^+, t^-, \xi_i^m) = \sum_{l=1}^{m} C_l (t_l^+ + t_l^-) + \sum_{l=1}^{m} \sum_{i=1}^{n} (\xi_i^{l+} + \xi_i^{l-})$$

$$s.t. \quad \begin{cases} -\xi_i^{-l} - t_l^- \leq (w^T X_i - Y_i) \leq t_l^+ + \xi_i^{+l} \\ 0 \leq \xi_i^{l+}, \xi_i^{l-}, \quad 0 \leq t_l^+, t_l^- \end{cases}$$

$$\forall l = 1, \ldots, m, \quad \forall i = 1, \ldots, n. \tag{32}$$

*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.*

Proof: The proof follows exactly the same lines as in Proposition 5, employing the fundamental theorem of linear programming and the first order conditions of optimality. Note that by construction, the different quantiles are properly nested, i.e. not allowed to cross.

□

Figure 2 gives an example of such a multi-quantile 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 $d \to \infty$, or by using reproducing kernels). Hereto, it is

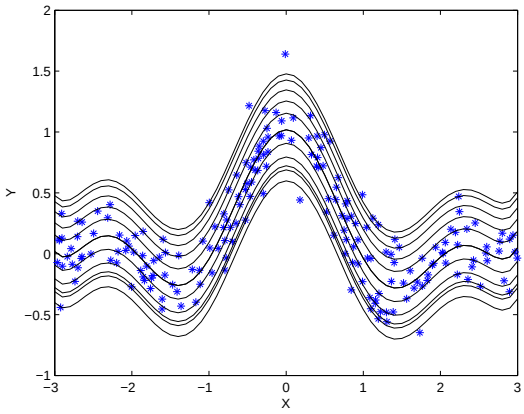

Figure 2: *Example of $n = 250$ a Multi-Quantile Vector Tube $\mathcal{T}_{m,s}^{(6)}$ with $\alpha = (25, 12, 6, 3, 2, 1)$. Here m consists of a linear combination of 10 localized basis-functions.*

easily 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. [4]). The disadvantage in this case is on the one hand that one should should choose the regularization constant in an appropriate way *a priori*. On the other hand, the influence of the regularization term becomes nontrivial in both the theoretical as well as in the computational derivation.

4 Conclusion

This paper [1] studied an intuitive estimator of the conditional support and quantiles of a distribution. The result is shown to be useful to estimate the

[1] **Acknowledgments**

Research supported by BOF PDM/05/161, FWO grant V 4.090.05N, IPSI Fraunhofer FgS, Darmstadt, Germany. (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. JS is a professor and BDM is a full professor at K.U.Leuven Belgium.

mutual information of the sample by extending the reach of Fano's theorem in combination with standard results of learning theory. It is indicated how the theoretical results relate to estimating the minimal convex hull.

References

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

[2] 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.

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

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

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

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

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

[8] 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.

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

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

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

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

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

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