NONPARAMETRIC VIRTUAL SENSORS FOR SEMICONDUCTOR MANUFACTURING

Using Information Theoretic Learning and Kernel Machines

Andrea Schirru[1], Simone Pampuri[1,2], Cristina De Luca[2] and Giuseppe De Nicolao[1]
[1]*Department of Computer Science Engineering, University of Pavia, Pavia, Italy*
[2]*Infineon Technologies Austria, Villach, Austria*

Keywords: Semiconductors, Machine learning, Entropy, Kernel methods.

Abstract: In this paper, a novel learning methodology is presented and discussed with reference to the application of virtual sensors in the semiconductor manufacturing environment. Density estimation techniques are used jointly with Renyi's entropy to define a loss function for the learning problem (relying on Information Theoretic Learning concepts). Furthermore, Reproducing Kernel Hilbert Spaces (RKHS) theory is employed to handle nonlinearities and include regularization capabilities in the model. The proposed algorithm allows to estimate the structure of the predictive model, as well as the associated probabilistic uncertainty, in a nonparametric fashion. The methodology is then validated using simulation studies and process data from the semiconductor manufacturing industry. The proposed approach proves to be especially effective in strongly nongaussian environments and presents notable outlier filtering capabilities.

1 INTRODUCTION

Virtual sensors are employed in many industrial settings to predict the result of an operation (most often a measurement) when the implementation of an actual sensor would be uneconomic or impossible (Rallo et al., 2002). In general, a virtual sensor finds and exploits a relation between some easily collectible variables (input) and one or more *target* (output) variables. Virtual sensor modeling techniques range from purely physics-based approaches (Popovic et al., 2009) to machine learning and statistical methodologies (Wang and Vachtsevanos, 2001). This paper is motivated by a specific class of virtual sensors used in semiconductor manufacturing, namely Virtual Metrology (VM) tools. The measurement operations on processed silicon wafers are particularly time-consuming and cost-intensive: therefore, only a small subset of the production is actually evaluated (Weber, 2007). Conversely, Virtual Metrology tools are able to predict metrology results at process time for every wafer, relying only on process data: such predictions are expected to reduce the need for actual measurement operations and, at the same time, establish positive interactions with metrology-related equipment tools (such as Run-to-Run controllers and decision aiding tools).

A Virtual Metrology tool is expected to **(i)** find and exploit complex, nonlinear relations between process data and metrology results, and **(ii)** assess prediction uncertainty in a meaningful way; in order to achieve such goals, it is key to make the right assumptions on the observed data. Remarkably, a precise characterization of the process variability is in general hard to obtain: for instance, the observed data might be distributed according to fat-tailed or strongly non-Gaussian distributions, be affected by outliers or present signs of multimodality; it is to note that such difficulties are shared among many disciplines (Ackerman et al., 2010). It is intuitive that suboptimal assumptions are likely to result in ineffective predictive models. In this paper, we present a novel methodology, inspired by Information Theoretic Learning theory (Principe, 2010), to tackle such an issue employing a regularized Reproducing Kernel Hilbert Space (RKHS) framework jointly with nonparametric density estimation techniques. The proposed approach is able to simultaneously estimate nonlinear predictive models and the associated prediction uncertainty, enabling the delivery of probabilistic predictions. The paper is structured as follows:

- Section 2 introduces the needed elements of machine learning and Kernel methods.

Schirru A., Pampuri S., De Luca C. and De Nicolao G..
NONPARAMETRIC VIRTUAL SENSORS FOR SEMICONDUCTOR MANUFACTURING - Using Information Theoretic Learning and Kernel Machines.
DOI: 10.5220/0003520403490358
In *Proceedings of the 8th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2011), pages 349-358
ISBN: 978-989-8425-75-1

- Section 3 presents and justifies the proposed approach from a theoretical point of view.
- Section 4 tests the proposed methodology against simulation studies and data from the semiconductor manufacturing environment.

Appendix A is devoted to numerical techniques used to solve the proposed problem, while Appendix B contains mathematical proofs.

2 MACHINE LEARNING AND KERNEL METHODS

The goal of a learning task is to estimate, from data, a relationship between an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. In order to achieve such result, it is necessary to rely on a set of observations $\mathcal{S} = \{x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1,\dots,N\}$. In other words, the goal is to find a map $f : \mathcal{X} \to \mathcal{Y}$ such that, given a new observation $\{x_{new} \in \mathcal{X}, y_{new} \in \mathcal{Y}\}$, $f(x_{new})$ will adequately predict y_{new}. In this framework, $\mathcal{S}$ is called a *training set* and the function f is an *estimator*. In the following, let f depend on a set of parameters θ, such that $f(x) := f(x;\theta)$; the optimization of θ with respect to some suitable criterion (function of $\mathcal{S}$ and θ) leads to the creation of a predictive model.

2.1 Regularized Machine Learning

In this paper, a regularized machine learning setting is employed to introduce and test the proposed methodology: the estimator f is found by minimizing some *loss function* $\mathcal{J}(\theta)$ with respect to θ. Such loss function is usually the sum of a *loss term* $\mathcal{L}$ and a *regularization term* $\mathcal{R}$, so that

$$\mathcal{J}(\theta) = \mathcal{L}(\theta) + \lambda\mathcal{R}(\theta) \tag{1}$$

In this framework, given a model specified by θ, $\mathcal{L}$ measures the quality of approximation on the training set $\mathcal{S}$ and $\mathcal{R}$ is a measure of the complexity of the model. Intuitively, the coexistence of $\mathcal{L}$ and $\mathcal{R}$ relates to a tradeoff between model regularity and performances on $\mathcal{S}$. The *regularization parameter* $\lambda \in \mathbb{R}^+$ acts as a tuning knob for such tradeoff: as λ grows, the order of the selected model gets lower and lower. In this paradigm, a learning algorithm is entirely specified by **(i)** the loss term $\mathcal{L}(\theta)$, **(ii)** the regularization term $\mathcal{R}(\theta)$ and **(iii)** the estimator structure $f(x;\theta)$. Remarkably, this structure assumes that the prediction of a generic y_i can be obtained, at best, up to a random uncertainty (depending on $\mathcal{L}$). In other words, adopting an additive error paradigm, it is implied that

$$y_i = f(x_i) + \varepsilon_i$$

where ε_i is a random variable whose distribution depends on $\mathcal{L}$.

In the following, let $\mathcal{X} \equiv \mathbb{R}^p$ and, with no loss of generality, $\mathcal{Y} \equiv \mathbb{R}$. The goal is to build a map $f : \mathbb{R}^p \to \mathbb{R}$ of the relationship between an input dataset $X \in \mathbb{R}^{N\times p}$ and an array of target observations $Y \in \mathbb{R}^N$. Furthermore, let x_i be the i-th row of X, and y_i be the i-th entry of Y.

2.2 Linear Predictive Models

Perhaps the most notable example of estimation technique is the method of Least Squares, that can be traced back to Gauss and Legendre. Such methodology assumes a linear relationship between the input and output spaces, so that

$$f(x_i) := f(x_i; w) = x_i w \tag{2}$$

where w is a p-variate vector of parameters. Furthermore, let $\mathcal{L}$ be the sum of squared residuals

$$\mathcal{L}(w) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{3}$$

and let $\mathcal{R}(w) \equiv 0$. The optimal w^* (minimizer of $\mathcal{L}(w)$) is then

$$w^* = (X'X)^{-1}X'Y$$

When a new input observation x_{new} is available, the optimal least squares prediction of y_{new} is

$$\hat{y}_{new} = E[y_{new}|x_{new}] = x_{new}w^*$$

Equation (3) implies a Gaussian distributed ε_i with

$$\varepsilon_i \sim N(0, \sigma^2)$$

where σ^2 is the variance of the observation uncertainty. Notably, $\hat{y}_{new}$ is independent of σ^2: it is necessary to tune the variance term only if a probabilistic output is needed (such as prediction confidence intervals). Least squares is a simple yet powerful method that suffers from two main drawbacks, namely **(i)** overfitting in high-dimensional spaces (p close to N) and **(ii)** possible ill-conditioning of the matrix $X'X$. In order to overcome such issues, a regularization term is employed: by using (2) and (3), and letting

$$\mathcal{R}(w) = \sum_{i=1}^{N} w_i^2$$

Ridge Regression is obtained. More and more stable (low sum of squared coefficients) models are selected as λ grows, at the cost of worsening the performances on the training set. The idea behind Ridge

Regression is that the optimal λ allows to build a predictor that includes all and only the relevant information. The optimal Ridge Regression coefficient vector is

$$w^* = (X'X + \lambda I)^{-1} X'Y$$

Similarly to least squares, it is not necessary to explicitly address the tuning of the error variance σ^2 unless a probabilistic output is needed.

2.3 Nonlinear Predictive Models

It is apparent that (2) defines a linear relationship between the *p*-variate input space and the output space. In a wide variety of applications, however, a linear model is not complex enough to obtain the desired prediction performances. An unsophisticated approach would be to adopt an *expanded basis* (augmenting the input set X with nonlinear functions of its columns - for instance, polynomials) to tackle such issue. It is to note, however, that this simple approach would yield computationally intractable problems also for a relatively small values of p (Hastie et al., 2005): nonlinearities are more efficiently handled using kernel-based methodologies. In the case of Ridge Regression, consider a symmetric positive definite matrix $K \in \mathbb{R}^{N \times N}$ whose entries arise from a suitable positive definite inner product $\mathcal{K}$ (kernel function), such that

$$K_{ij} = \mathcal{K}(x_i, x_j) \tag{4}$$

Furthermore, consider the model structure

$$f(x_i) = K_i c \tag{5}$$

where K_i is the i-th row of K, and the regularization term

$$\mathcal{R}(c) = c'Kc$$

In this framework, $c \in \mathbb{R}^N$ is the coefficient vector of the so-called *dual form* of the learning problem, and $\mathcal{R}$ is the norm of f in a nonlinear Hilbert space. The resulting model f exploits a nonlinear relationship (specified by $\mathcal{K}$) between X and Y. This result arises from RKHS (Reproducing Kernel Hilbert Spaces) theory and Riesz Representation Theorem: the kernel function $\mathcal{K}$ is used to establish a relationship between the p features and the N examples. Among the most popular kernel functions, the inhomogeneous polynomial kernel

$$\mathcal{K}(x_i, x_j; d) = (x_i x_j' + 1)^d$$

incorporates the polynomial span of X up to the d-th grade, and the exponential kernel

$$\mathcal{K}(x_i, x_j; \xi) = e^{-\frac{||x_i - x_j||^2}{\xi^2}}$$

relates to an infinite-dimensional feature space whose bandwidth is controlled by ξ^2. The optimal Kernel Ridge Regression coefficient vector is

$$c^* = (K + \lambda I)^{-1} Y$$

and the predictor is

$$\hat{y}_{new} = k_{new} c^*$$

with $k_{new} = [\mathcal{K}(x_{new}, x_1) \dots \mathcal{K}(x_{new}, x_N)]$. A thorough review of kernel-based methodologies is beyond the scope of this section: the interested reader can find more information in (Scholkopf and Smola, 2002).

2.4 Learning in Nongaussian Settings

It is to note that the methodologies reviewed in this section rely on Gaussian assumptions: the probabilistic interpretation of the loss function (3) is that y_{new} can be predicted, at best, with an additive Gaussian-distributed uncertainty with fixed variance. The reasons for adopting such an assumption are both historical (linking to the concept of Least Squares estimation) and methodological (Central Limit Theorem and closed form solution), but other choices are possible. For instance, the Huber loss function (used in robust statistics) implies a Gaussian distribution near the origin with Laplace tails and allows to reduce the weight of outliers in the learning process. Another notable example is the ε-insensitive loss function, that relates to a uniform distribution between $[-\varepsilon, \varepsilon]$ with Laplace tails, and is mainly used in Support Vector Machines (SVM).

Remarkably, all the loss terms described in this section rely on parametric assumptions: the uncertainty is assumed to follow a specific (known) distribution depending on a set of unknown parameters. In a real setting, however, it is often not possible (and sometimes not even desirable) to identify the uncertainty distribution in a parametric way: in such situations, a more flexible characterization is needed to achieve the best performances. In the next section, a learning method that achieves such flexibility is presented using Density Estimation techniques jointly with Entropy-related criteria.

3 REGULARIZED ENTROPY LEARNING

In this section, an entropy-based learning technique that makes no assumptions about the uncertainty distribution is presented and discussed. The novel methodology will be referred to as "Regularized Entropy Learning".

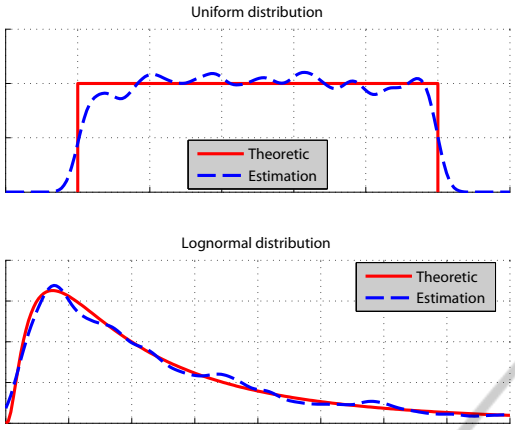

Figure 1: Density estimation of uniform and lognormal distributions, using Gaussian densities.

3.1 Density Estimation and Learning

Consider a real-valued array $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_N]'$, where every ε_i is assumed to be independently drawn from the same unknown distribution. In order to obtain a nonparametric estimate of the probability density of ε, it is convenient to resort to *density estimation* techniques (Parzen, 1962).

> Remark: it would be more correct to use the term "*kernel* density estimation" (KDE). In order to avoid confusion (the word "kernel" has different meaning in KDE and Kernel Methods), KDE will be referred to as *density estimation* (DE).

DE techniques are able to estimate a probability density from a set of observations, using a mixture of predetermined distributions. Given the vector ε, the underlying distribution is estimated as

$$p_\varepsilon(x) = \frac{1}{N}\sum_{i=1}^{N} g(\varepsilon_i; x) \tag{6}$$

where $g(\cdot; x)$ is a nonnegative function such that

$$\int_{-\infty}^{+\infty} g(\cdot; x)dx = 1$$

It is immediate to prove that (6) is a probability distribution. In this paper, we employ the Gaussian density

$$G(\mu, \sigma^2; x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{7}$$

so that $g(z; x) := G(z, \sigma^2; x)$. Hereby σ^2 is the *bandwidth* of the estimator, related to the smoothness of the estimated density: its tuning will be discussed in a later subsection. The density p_ε is rewritten as

$$p_\varepsilon(x) = \frac{1}{N}\sum_{i=1}^{N} G(\varepsilon_i, \sigma^2; x) \tag{8}$$

that is, a Gaussian density of variance σ^2 is centered on every observation ε_i. With reference to the learning setting presented in the previous section, let ε_i be the estimation error (residual) on the i-th sample of $\mathcal{S}$, for some value of c:

$$\varepsilon_i := \varepsilon_i(c) = y_i - K_i c \tag{9}$$

where K_i is the i-th row of the kernel matrix K. In the next subsection, (8) and (9) are used to define a loss term related to the concept of information entropy.

3.2 Entropy-based Loss Term

In information theory, entropy is a measure of the uncertainty of a random variable: while an high entropy is associated to chaos and disorder, a quiet and predictable random variable is characterized by low entropy (Gray, 2010). Notably, by minimizing the entropy of a random variable, a constraint is imposed on all its moments (Erdogmus and Principe, 2002). For this reason, the definition of an entropy-based loss term is desirable with respect to the Least Squares loss term, that involves only the second moment (variance). More interesting properties of such a loss term are investigated in (Principe et al., 2000).

Shannon's entropy, perhaps the most notable entropy measure, is defined as the expected value of the information contained in a message. Renyi's entropy generalizes this concept to a family of functions depending on a parameter $\alpha \geq 0$. Consider a continuous random variable ε; its Renyi's entropy $H_\alpha(\varepsilon)$ is

$$H_\alpha(\varepsilon) = \frac{1}{1-\alpha}\log \int_{-\infty}^{+\infty} p_\varepsilon(x)^\alpha dx \tag{10}$$

We consider the quadratic Renyi's entropy $H_2(\cdot)$ of the random variable $\varepsilon|c$, as

$$H_2(\varepsilon(c)) = -\log \int_{-\infty}^{+\infty} p_{\varepsilon|c}(x)^2 dx \tag{11}$$

It is easily noted that $H_2(\varepsilon)$ reaches its infimum when $p_\varepsilon(x)$ is a Dirac Delta (complete predictability), and its supremum when $p_\varepsilon(x)$ is flat over $\mathbb{R}$ (complete uncertainty). In order to define the desired loss term, we consider the following

Theorem 1. Let $A \in \mathbb{R}^{s\times s}$, $a \in \mathbb{R}^s$, $B \in \mathbb{R}^{t\times t}$, $b \in \mathbb{R}^t$ and $Q \in \mathbb{R}^{s\times t}$. Let $\mathbf{x} \in \mathbb{R}^t$ be an input variable. It holds that

$$G(a, A; Q\mathbf{x})G(b, B; \mathbf{x}) = G(a, A + QBQ'; b)G(d, D; \mathbf{x})$$

with

$$\begin{aligned} D &= (Q'A^{-1}Q+B^{-1})^{-1} \\ d &= b+DQ'A^{-1}(a-Qb) \end{aligned}$$

□

It is possible to express $H_2(\varepsilon)$ in function of a weighted sum of Gaussian densities: this result is summarized in the following

Proposition 1. Applying Theorem 1 (Miller, 1964) and using (8) and (11), it holds that

$$H_2(\varepsilon(c)) = -\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(y_i-y_j,2\sigma^2;(K_i-K_j)c)\right)$$

Exploiting the symmetry of the Gaussian density, we define

$$\mathcal{H}(c) := \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=i+1}^{N} G(y_i-y_j,2\sigma^2;(K_i-K_j)c)$$

and observe that $\mathcal{H}(c)$ is equal to $e^{-H_2(\varepsilon(c))}$ up to an additive constant. Since the exponential transformation is monotonic,

$$\arg\max_c \mathcal{H}(c) = \arg\min_c H_2(\varepsilon(c)) \qquad (12)$$

□

Equation (12) states that a minimum entropy estimator can be obtained by maximizing a mixture of Gaussian densities with respect to the parameters vector c. In the following, since ε is entirely specified by c, we let $H_2(c) := H_2(\varepsilon(c))$.

3.3 Regularized Entropy Learning

In this section, we consider the properties of the learning algorithm for which

$$\begin{aligned} \mathcal{L}(c) &= H_2(c) \\ \mathcal{R}(c) &= c'Kc \\ f(y_i) &= k_i c \end{aligned}$$

The novelty of the proposed approach lies in the RKHS regularization of an entropy-related loss term. Consider the following

Proposition 2. Given the loss function

$$\mathcal{J}(c) = H_2(c)+\lambda c'Kc \qquad (13)$$

it holds that

$$e^{-\mathcal{J}(c)} \propto \mathcal{H}(c)G\left(0_N,\frac{K^{-1}}{\lambda};c\right) \qquad (14)$$

that is, applying an exponential transformation to $\mathcal{J}(c)$, it is possible to write it as the product between a weighted sum of Gaussian densities ($\mathcal{H}(c)$) and a Gaussian density dependent on λ. □

Furthermore, it has to be considered that $H_2(c)$ is shift-invariant: this result is discussed in the following

Proposition 3. Let $\varepsilon(c)$ be a real valued vector of residuals associated to a coefficient vector c, and let $\varepsilon(c^*) = \varepsilon(c)+z$, where z is a real constant. It holds that

$$H_2(c) \equiv H_2(c^*)$$

□

Following Proposition 3, the expected value of the residuals represents an additional degree of freedom to be set in advance. Without loss of generality we choose to ensure that, given a random variable γ,

$$(p(\gamma|c) = p_\varepsilon(x)) \rightarrow (E[\gamma] = 0) \qquad (15)$$

According to Proposition 2, it is possible to write $\mathcal{J}(c)$, upon a monotonic transformation, as a sum of products of Gaussian densities. In order to define an efficient minimization strategy for $\mathcal{J}$, we consider the following

Proposition 4. It holds that

$$e^{-\mathcal{J}(c)} \propto \sum_{i=1}^{N}\sum_{j=i+1}^{N} \alpha_{ij}G(d_{ij},D_{ij};c) \qquad (16)$$

with

$$\begin{aligned} \alpha_{ij} &= G(y_i,2\sigma^2+\frac{K_{ii}-2K_{ij}+K_{jj}}{\lambda};y_j) \\ D_{ij} &= \left(\frac{(K_i-K_j)'(K_i-K_j)}{2\sigma^2}+\lambda K\right)^{-1} \\ d_{ij} &= D_{ij}\frac{(K_i-K_j)'(y_i-y_j)}{2\sigma^2} \end{aligned}$$

where K_{st} is the $\{s,t\}$ entry of K, and therefore

$$\mathcal{J}(c) = -\log\left(\sum_{i=1}^{N}\sum_{j=i+1}^{N}\alpha_{ij}G(d_{ij},D_{ij};c)\right)$$

up to an additive constant. □

Proposition 4 straightforwardly applies Theorem 1 to state that it is possible to write $\mathcal{J}(c)$ as the logarithm of a weighted sum of Gaussian densities. The

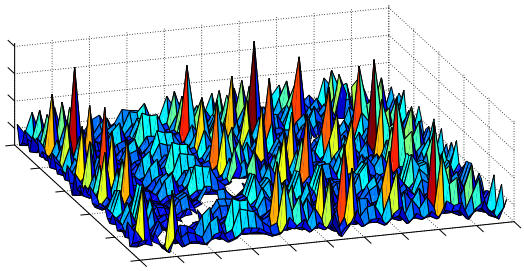

Figure 2: Graphical representation of the matrix $\{\alpha_{ij}\}$ (surface plot).

multiplicative coefficients α_{ij} admit an interesting interpretation: α_{ij}, for which

$$0 \leq \alpha_{ij} \leq G(0, 2\sigma^2; 0) \tag{17}$$

gets monotonically closer to its supremum as two conditions are met: **(i)** y_i is close to y_j and **(ii)** $K_{ii} + K_{ij}$ is close to $2K_{ij}$. Using the definition of $K_{ij} = \mathcal{K}(x_i, x_j)$, it is immediately verified that condition **(ii)** occurs when x_i is close to x_j. Therefore, the multiplicative coefficient α_{ij} relates to the *information consistency* between the i-th and j-th sample: in other words, it is a measure of the similarity between the i-th and j-th observations. This allows for two interesting properties: **(i)** given a training set $\mathcal{S}$, it is possible to identify the most consistently informative pairs of examples (Figure 2). This information can be subsequently used, for instance, as a pruning criterion to obtain a minimal representative dataset. Furthermore, **(ii)** it is possible to use $\{\alpha_{ij}\}$ to discover mixtures in $\mathcal{S}$: indeed, if it is possible to identify two sets $\mathcal{S}_1 \subset \mathcal{S}$ and $\mathcal{S}_2 \subset \mathcal{S}$ such that, for all $(i, j \in \mathcal{S}_1)$ and $(k, z \in \mathcal{S}_2)$,

$$\begin{aligned} \alpha_{ij} &\gg \alpha_{ik} \\ \alpha_{kz} &\gg \alpha_{ik} \end{aligned}$$

the information conveyed by $\mathcal{S}_1$ and $\mathcal{S}_2$ are significantly decoupled. Figure 3 depicts a colormap of $\{\alpha_{ij}\}$ for a toy dataset with $N = 30$, obtained by concatenating two decoupled sets of observations. As expected, the upper-left and lower-right $15x15$ submatrices show the highest values of α_{ij}.

3.4 Model Estimation

In the previous subsection, we have shown that the loss function of the proposed method is monotonically related to a weighted sum of N-variate Gaussian densities. In this subsection, an optimal (entropy-wise) regularized estimator of c is derived and employed to build a predictor for new observations. In order to obtain c^*, it is necessary to solve the following

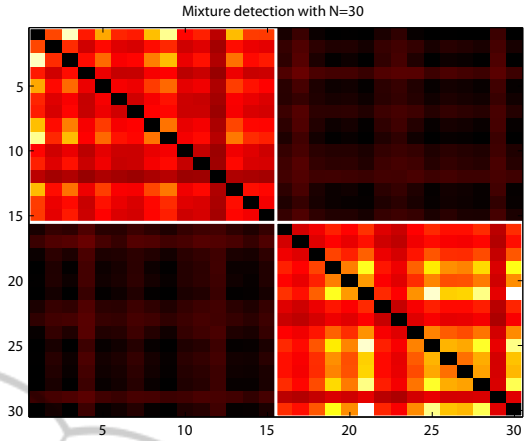

Figure 3: Mixture recognition capabilities of the coefficients α_{ij}: the bright areas are self-consistent groups of homogeneous observations.

Problem 1. (Minimization of $\mathcal{J}(c)$): find

$$c^* = \arg\min_c e^{\mathcal{J}(c)}$$

with

$$\mathcal{J}(c) = -\log\left(\sum_{i=1}^{N}\sum_{j=i+1}^{N} \alpha_{ij} G(d_{ij}, D_{ij}; c)\right)$$

Since the exponential transformation is monotonic, the minimizer of $e^{\mathcal{J}(c)}$ minimizes also $\mathcal{J}(c)$; the exponential formulation yields, however, simpler derivatives. Implementation details about the solution of Problem 1 are reported in Appendix A. The estimate c^* represents a compromise between the RKHS norm of f, $\mathcal{R}(c)$, and Renyi's second order entropy of the estimation errors, $H_2(c)$. Additionally, $H_2(c^*)$ is the minimum reachable entropy configuration for the dataset $\mathcal{S}$ for a given value of λ. As c^* is obtained by solving Problem 1, it is necessary to set the additional degree of freedom discussed in Proposition 3: the bias term $\mathcal{B}$ is computed as

$$\mathcal{B} := \text{mean}(\{\varepsilon_i\}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - K_i c^*)$$

It is then possible to compute predictions whenever a new observation x_{new} is available. Using the real-valued array

$$k_{new} = [\mathcal{K}(\tilde{x}_{new}, x_1) \ldots \mathcal{K}(\tilde{x}_{new}, x_N)]$$

the estimator $\hat{y}_{new}$ is

$$\hat{y}_{new} = E[y_{new}|x_{new}] = k_{new} c^* \tag{18}$$

Furthermore, the prediction uncertainty γ can be estimated using a Leave-One-Out approach: let

$$p(\gamma) = \frac{1}{N}\sum_{i=1}^{N} G(y_i - K_i c^*_{(i)} - \mathcal{B}_{(i)}, \sigma^2; x) \qquad (19)$$

where $c^*_{(i)}$ and $\mathcal{B}_{(i)}$ solve Problem 1 using the reduced dataset $\mathcal{S}_{(i)} = \mathcal{S} \setminus \{x_i, y_i\}$. The probabilistic form of the predictor is then

$$y_{new} = \hat{y}_{new} + \gamma \qquad (20)$$

where $\hat{y}_{new}$ is deterministic and γ is a random variable.

It is to be noted that the solution of Problem 1 was obtained for fixed bandwidth σ^2 and regularization parameter λ. In practice, however, it is generally needed to estimate such parameters from data as well. In the presented univariate case, the tuning of such parameters was efficiently performed by means of a Generalized Cross Validation (GCV) approach. It is also possible, if there is not enough data to perform GCV, is to tune σ^2 relying on some theoretical criterion, such as Silverman's rule (Silverman, 1986).

4 RESULTS

In this section, Regularized Entropy Learning (REL) is tested against simulated datasets and data coming from the semiconductor manufacturing industry. In order to understand the potential of the proposed approach, its performances are compared with Kernel Ridge Regression (KRR). The performance gap between Regularized Entropy Learning and Kernel Ridge Regression is expected to vary: intuitively, if the real uncertainty distribution is strongly nongaussian, REL is expected to outperform KRR. For all the experiments, RMSE was used as metric; the parameters of both methods were tuned by means of GCV.

4.1 Simulated Datasets

Two datasets (100 training, 50 validation and 50 test observations) were generated using strongly nongaussian uncertainty distributions. In the first case, a uniform random variable was employed; in the second case, the uncertainty was modeled as a Laplace distribution. Multiplicative (5x) outliers were inserted (with 10% probability) in each dataset to test the robustness of the proposed methodology. Table 1 presents the RMSE ratios for the four simulated experiments: as expected, the best results are obtained with the Laplace distribution (whose power law is hardly approximated by a Gaussian). Figures 4 and 5 show the uncertainty distributions considered by REL and KRR: the advantage of the proposed approach in such situations is clear.

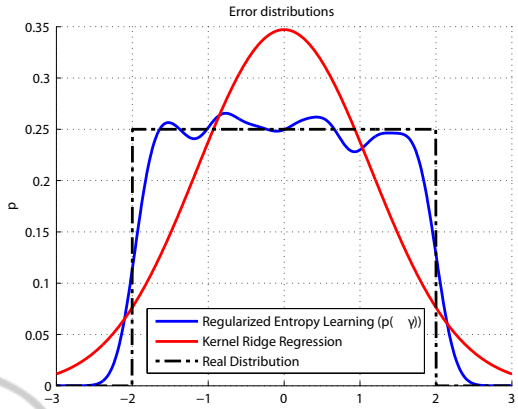

Figure 4: Error distributions for the prediction of a simulated dataset with uniform uncertainty: the nonparametric estimation of $p(\gamma)$ allows the proposed methodology to outperform KRR.

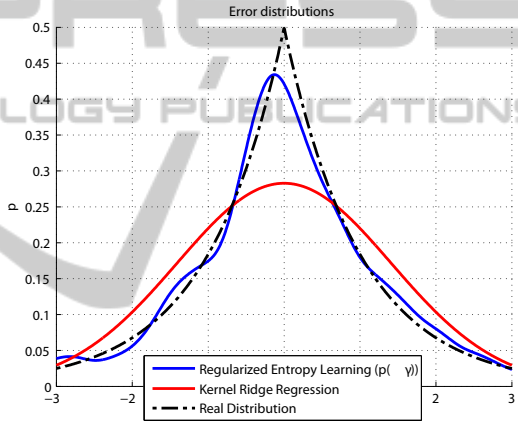

Figure 5: Error distributions for the prediction of a simulated dataset with Laplace-distributed uncertainty: the fat tails of the Laplace distribution are correctly recognized by REL, while the Gaussian assumptions of KRR fail to achieve the best prediction results.

4.2 Semiconductor Dataset

REL was tested against a set of homogeneous observations from the semiconductor manufacturing environment (courtesy of the Infineon Austria facility in Villach). Specifically, a homogeneous dataset consisting of 239 measured wafers was collected from a Chemical Vapor Deposition (CVD) equipment. Every wafer is characterized by 30 input variables and 9 thickness measurements (associated to different sites on the wafer). Four experiments were conducted:

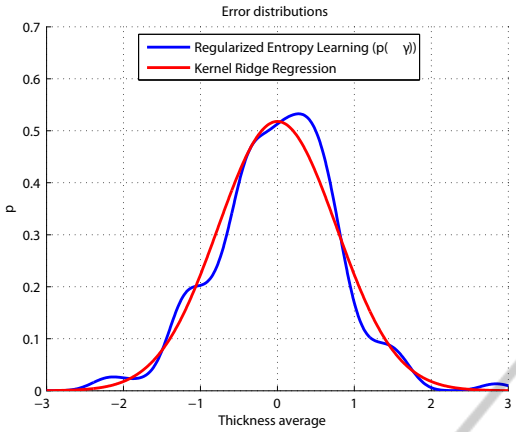

Figure 6: Error distributions for the prediction of average thickness with no outliers: notably, in this case the Gaussian assumptions are verified, and KRR performs better than the proposed methodology.

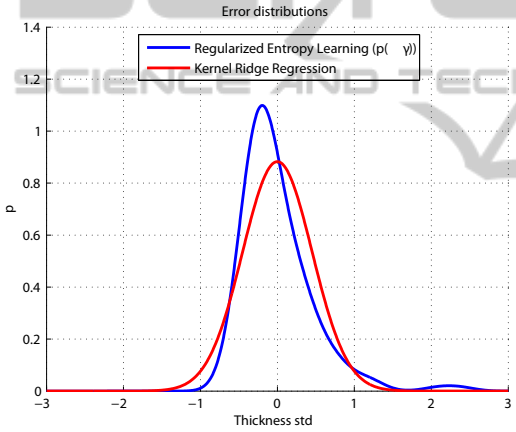

Figure 7: In the prediction of standard deviation, the density estimated distribution is skewed and strongly nongaussian (both the distributions in this Figure has expected value 0). Thanks to this better understanding of the uncertainty distribution, the proposed methodology performs well in this setting.

- Predict the average layer thickness (with and without outliers).
- Predict the standard deviation (9 points) of layer thickness (with and without an outliers).

In order to produce outlier-free experiments, a standard outlier elimination technique based on Mahalanobis distance was employed. The original dataset was split in 3 consecutive groups: a training dataset of 150 wafers, a validation dataset of 50 wafers and a test dataset of 39 wafers. The validation dataset was used to tune the hyperparameters, while the algorithms were compared on the test dataset.

The results of this experiment are reported in Table 1. Notably, the Gaussian assumptions are verified for the average thickness without outliers (Figure 6): in this case, KRR performs better than REL. On the other hand, the presence of outliers in the dataset significantly degrades the performances of KRR, while the proposed methodology proves to be naturally robust. Perhaps the most interesting result comes from the prediction of standard deviation: thanks to the skewed uncertainty distribution associated to the standard deviation measurements (Figure 7), REL outperforms KRR.

Table 1: RMSE ratios REL/KRR: the proposed methodology shows natural robustness with respect to outliers, and performs better than KRR when Gaussianity assumptions are less realistic. The best result is in bold, while the worst result is in red.

	With outliers	Without outliers
Uniform	0.61	0.67
Laplace	**0.52**	0.56
Thickness Avg.	0.98	1.17
Thickness Std.	**0.73**	0.89

5 CONCLUSIONS

In this paper, a novel learning methodology is proposed relying on information theory concepts in a regularized machine learning framework. This study is motivated by the application of Virtual Metrology in semiconductor manufacturing. The estimation of a nonlinear predictive model, jointly with the associated uncertainty distribution, is achieved in a nonparametric way using a metric based on Renyi's entropy and regularized with a RKHS norm. The proposed methodology, namely Regularized Entropy Learning (REL), has been tested with promising results on simulated datasets and process data from the semiconductor manufacturing environment. Specifically, the proposed approach presents a clear advantage in outlier-intensive and strongly nongaussian environments: for this reasons, REL is a strong candidate for the use in an industrial setting, where a flexible assessment of data variability is key to achieve good predictive performances.

REFERENCES

Ackerman, F., Stanton, E., and Bueno, R. (2010). Fat tails, exponents, extreme uncertainty: Simulating catastrophe in DICE. *Ecological Economics*, 69(8):1657–1665.

Carreira-Perpiñán, M. (2002). Mode-finding for mixtures of Gaussian distributions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1318–1323.

Erdogmus, D. and Principe, J. (2002). An error-entropy minimization algorithm for supervised training of nonlinear adaptive systems. *Signal Processing, IEEE Transactions on*, 50(7):1780–1786.

Gray, R. (2010). *Entropy and information theory*. Springer Verlag.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Miller, K. (1964). *Multidimensional gaussian distributions*. Wiley New York.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Popovic, D., Milosavljevic, V., Zekic, A., Macgearailt, N., and Daniels, S. (2009). Impact of low pressure plasma discharge on etch rate of SiO2 wafer. In *APS Meeting Abstracts*, volume 1, page 8037P.

Principe, J. (2010). *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Verlag.

Principe, J., Xu, D., Zhao, Q., and Fisher, J. (2000). Learning from examples with information theoretic criteria. *The Journal of VLSI Signal Processing*, 26(1):61–77.

Rallo, R., Ferre-Giné, J., Arenas, A., and Giralt, F. (2002). Neural virtual sensor for the inferential prediction of product quality from process variables. *Computers & Chemical Engineering*, 26(12):1735–1754.

Scholkopf, B. and Smola, A. (2002). *Learning with kernels*, volume 64. Citeseer.

Silverman, B. (1986). Density Estimation for Statistics and Data Analysis. Number 26 in Monographs on statistics and applied probability.

Wang, P. and Vachtsevanos, G. (2001). Fault prognostics using dynamic wavelet neural networks. *AI EDAM*, 15(04):349–365.

Weber, A. (2007). Virtual metrology and your technology watch list: ten things you should know about this emerging technology. *Future Fab International*, 22:52–54.

APPENDIX A

Problem 1 is an unconstrained global optimization problem. In order to derive a solution, we consider the features of c^* in the following

Proposition 5. Let c^* be the global minimum of a negatively weighted sum of Gaussian densities $J(c)$. Thanks to the properties of $J(c)$, there exists a real m such that

$$||c^* - d_{ij}||^2_{D_{ij}^{-1}} \leq m$$

for at least one mean vector d_{ij}. Furthermore,

$$m \leq \log\left(\frac{C_0}{\tilde{J}}\right)$$

where C_0 is a negative constant and

$$\tilde{J} = \min_{i,j} J(d_{ij})$$

That is, m is superiorly limited by a decreasing function of the minimum value of J evaluated in the mean vectors d_{ij}. □

Proposition 5 has two notable implications: **(i)** there is at least one mean vector d_{ij} that serves as suitable starting point for a local optimization procedure, and **(ii)** the global minimum gets closer to one of the mean vectors as the computable quantity $\tilde{J}$ increases. Using these results, c^* is found by means of the following

Algorithm 1: solution of Problem 1.

1. Set $c^* = 0_N$
2. For $i = 1, \ldots, N$
   (a) For $j = i+1, \ldots, N$
   - Use a Newton-Raphson algorithm to solve the local optimization problem
     $$c^*_{ij} = \arg\min_c J(c)$$
     using d_{ij} as starting point.
   - If $J(c^*_{ij}) < J(c^*)$
   - $c^* = c^*_{ij}$
   - End if

   (b) End for
3. End for

Algorithm 1 was originally proposed in (Carreira-Perpiñán, 2002), and guarantees to find all the modes of a mixture of Gaussian distributions. It is to be noted, however, that the exhaustive search performed by Algorithm 1 might be computationally demanding, and only the global minimum of J is of interest in the presented case. It is convenient, if an approximate optimal solution is acceptable, to perform convex optimization using a reduced number of starting points. In our experiments, the best performances were obtained using the d_{ij} associated to the least (1% to 5%) values of $\{J(d_{ij})\}$. It is to be noted that this reduced version of Algorithm 1 does not guarantee to reach the global minimum (although it has been verified, via simulation studies, that the global optimum is found with a very high success rate). In order to set up the Newton-Raphson algorithm used in Algorithm 1, it is

necessary to know the Jacobian and Hessian matrices associated to $J(c)$. Letting

$$\begin{aligned} Q_{ij}(c) &= -\frac{1}{2}(c-d_{ij})'D_{ij}^{-1}(c-d_{ij}) \\ \overline{\alpha}_{ij} &= \frac{\alpha_{ij}|D_{ij}^{-1}|^{1/2}}{(2\pi)^{N/2}} \end{aligned}$$

it is possible to write

$$J(c) = -\sum_{i=1}^{N}\sum_{j=i+1}^{N}\overline{\alpha}_{ij}e^{Q_{ij}(c)}$$

Therefore, the Jacobian of $J(c)$ is

$$\frac{\partial J}{\partial c} = -\sum_{i=1}^{N}\sum_{j=i+1}^{N}\overline{\alpha}_{ij}e^{Q_{ij}}\frac{\partial Q_{ij}}{\partial c}$$

and the Hessian is

$$\frac{\partial^2 J}{\partial^2 c} = -\sum_{i=1}^{N}\sum_{j=i+1}^{N}\overline{\alpha}_{ij}e^{Q_{ij}}\left[\frac{\partial Q_{ij}}{\partial c}\frac{\partial Q_{ij}}{\partial c'}+\frac{\partial^2 Q_{ij}}{\partial^2 c}\right]$$

with

$$\begin{aligned} \frac{\partial Q_{ij}}{\partial c} &= D_{ij}^{-1}(d_{ij}-c) \\ \frac{\partial^2 Q_{ij}}{\partial^2 c} &= -D_{ij}^{-1} \end{aligned}$$

Considering the Taylor series of $J(c)|_{c=c_k}$ truncated to the second order

$$\begin{aligned} J(c)|_{c=c_k} \approx J(c_k) &+ (c-c_k)'\frac{\partial J}{\partial c}|_{c=c_k} \\ &+\frac{1}{2}(c-c_k)'\frac{\partial^2 J}{\partial^2 c}|_{c=c_k}(c-c_k) \end{aligned}$$

the next c_{k+1} is

$$c_{k+1} = x_k - \left(\frac{\partial^2 J}{\partial^2 c}|_{c=c_k}\right)^{-1}\frac{\partial J}{\partial c}|_{c=c_k}$$

Remark: in order to evaluate $J(c)$, as well as its Jacobian and Hessian, it is not necessary to explicitly compute any matrix inversion: indeed,

$$D_{ij}^{-1} = \frac{(K_i-K_j)'(K_i-K_j)}{2\sigma^2}+\lambda K$$

APPENDIX B

Proof of Proposition 3. It is apparent that a global shift of x does not influence the value of H_2, thanks to the infinite integration interval:

$$\int_{-\infty}^{+\infty} p_\varepsilon(x)^2 dx \equiv \int_{-\infty}^{+\infty} p_\varepsilon(x+\tau)^2 dx$$

The result follows. □

Proof (by contradiction) of Proposition 5. Consider, without loss of generality, the problem of finding the global maximum x^* of a weighted sum of Gaussian densities $L(x)$, such that

$$L(x) = \sum_{i=1}^{N}\alpha_i G(\mu_i,\Sigma_i;\, x)$$

Letting

$$\begin{aligned} \overline{\alpha}_i &= \frac{\alpha_i}{(2\pi)^{N/2}|\Sigma_i|^{1/2}} \\ ||x-\mu_i||^2_{\Sigma_i^{-1}} &= \frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) \end{aligned}$$

we write $L(x)$ as

$$L(x) = \sum_{i=1}^{N}\overline{\alpha}_i e^{-||x-\mu_i||^2_{\Sigma_i^{-1}}}$$

Consider then the best function value among the $\{\mu_i\}$ as

$$\tilde{L} = \max_i L(\mu_i)$$

and let x^* be far from every μ_i so that

$$||x^*-\mu_i||^2_{\Sigma_i^{-1}} \geq m \;\forall i$$

where m is a lower bound. It is apparent that

$$\tilde{L} \leq L(x^*) \leq Ne^{-m}\sum_{i=1}^{N}\overline{\alpha}_i \qquad (21)$$

and it is immediately verified that if

$$m > \log\left(\frac{N\sum_{i=1}^{N}\alpha_i}{\tilde{L}}\right)$$

inequality (21) does not hold: by contradiction, x^* is not the global maximum. Therefore, there exists at least one mean vector μ_i such that

$$||x^*-\mu_i||^2_{\Sigma_i^{-1}} < \log\left(\frac{N\sum_{i=1}^{N}\alpha_i}{\tilde{J}}\right)$$