CENTRE DE RECERCA MATEMÀTICA

Preprint núm. 1151
April 2013

Sepsis mortality prediction with the quotient basis kernel

V. J. Ribas, A. Vellido, E. Romero, J. C. Ruiz-Rodríguez

SEPSIS MORTALITY PREDICTION WITH THE QUOTIENT BASIS KERNEL

VICENT J. RIBAS RIPOLL, ALFREDO VELLIDO, ENRIQUE ROMERO,
JUAN CARLOS RUIZ-RODRÍGUEZ *MD*

Abstract. Sepsis is a common clinical syndrome at the intensive care unit (ICU). This condition may lead to severe sepsis, or to a more severe state of septic shock and multiorganic failure, which entails a substantial risk of death. The extreme realtime demands of the ICU require informed decision making on the basis of the available evidence, and such decision making can be supported by semi-automated methods for quantitative mortality prediction. These methods should be robust and feasible within the constraints of the domain. In this paper, we describe a novel sepsis mortality prediction method that first embeds the available data in a suitable feature space, and then uses algorithms based on linear algebra, geometry and statistics for inference. A simplified version of the Fisher kernel (practical Fisher kernel for multinomial distributions) as well as a novel kernel, namely the Quotient Basis Kernel (QBK), are defined and used as the basis for mortality prediction using soft-margin support vector machines. The results compare favorably with those obtained using alternative kernels and the standard clinical prediction method based on the basal SAPS score.

1. Introduction

Sepsis is a clinical syndrome defined by the presence of both infection and Systemic Inflammatory Response Syndrome (SIRS). This condition may lead to severe sepsis, which implies organ dysfunction, or to an even more severe state: septic shock and multiorganic failure [Ame92, Lev03].

This pathology has followed a clear upwards trend over the last 20 years, reaching 300,000 cases per year only in the United States of America. This figure is likely to grow as the population ages and treatment becomes more aggressive [Ang01, Mar03]. In western health systems, septic patients account for as much as 25% of bed utilization at the Intensive Care Unit (ICU) and the pathology occurs in 1% - 2% of all hospitalizations. The mortality rates of sepsis are very high, ranging from 12.8% for sepsis to 45.7% for the most severe septic shock [Est07].

These figures justify by themselves the need for a quantitative approach to mortality prediction due to sepsis at the ICU. The extreme demands of this

Key words and phrases. Sepsis, mortality prediction, kernels, support vector machines, critical care.

clinical environment further require prediction methods that are both robust and feasible within its constraints.

In this paper, we describe a novel sepsis mortality prediction method that first embeds the available data in a suitable feature space, and then uses algorithms based on linear algebra, geometry and statistics for inference. More specifically, we present a novel kernel for multinomial distributions, namely the Quotient Basis Kernel (QBK), which is based on the reparametrization of the input space through algebraic geometry and algebraic statistics. This kernel can be efficiently modelled algebraically by means of the regular exponential family. In addition, we present a generative approach that exploits the inner structure of our data in order to build a set of efficient closed-form kernels that are best suited for the multinomial distribution.

The QBK is the result of calculating the covariance of the design matrix of a Gröbner basis. In this paper, we hipothesize that the QBK is particularly well suited for the problem at hand because not only does it exploit the inner structure of the data (i.e., it is generative), but it also provides a geometric representation framework that accounts for the inner dependencies between its inputs [Gig00] (maximum/minimum Sequential Organ Failure Assessment (SOFA) and Simplified Acute Physiology Score (SAPS) scores) and represents them as polynomial terms. It has also been shown that such a representation is very closely related to graphical models [Pac05] in such a way that these kernels could be considered as *open-box* methods.

The performance of the proposed QBK method in the prediction of mortality due to sepsis is compared, using soft-margin Support Vector Machines (SVM), to those obtained with a number of alternative kernels. It is also compared to a standard method used in clinical practice that is based on the basal SAPS score [Le 84] (i.e. through the automatic selection of a threshold).

The remaining of the paper is organized as follows. Section 2 presents the database used in this study along with the two main indices that shall be used for mortality prediction: the SOFA and SAPS scores. In section 2, we describe a simplified version of the Fisher kernel for multinomial families. This section also provides an overview of the kernels based on the Jensen-Shannon metric [Aga11] with a special emphasis on a reparametrization of the log-Laplace transform term of a regular exponential family. This section closes with the formal definition of the novel QBK and a short overview of SVMs. In section IV, we show the experimental prediction results for each different kernel as well as their comparison with standard mortality prediction from the basal SAPS score.

In normal clinical practice, while treating sepsis and its more severe manifestations, clinicians are often forced to catch up with the pathology. They are treating severely ill patients at later stages of illness. Many of these may be suffering from a combination of chronic and acute disease.

Illness scoring systems are commonplace at the ICU. The rationale for using these systems in a clinical environment is to ensure that the increased complexity of disease in patients currently being treated is consistently represented and assessed. A specific goal of severity scoring systems is to use the representative attributes of these patient to describe the relative risks they face and to identify where the patient can be located along the continuum of illness severity.

It is increasingly being recognized that the ultimate goal of severity scoring can be more than just obtaining an overall figure representing the degree of physiologic disturbance. Severity scoring can be used in conjunction with other risk factors such as disease aetiology to anticipate and estimate outcomes such as ICU mortality. These estimates can be calculated at the time a patient presents for care or for entry into a clinical trial. Therefore, they can serve as a pre-treatment protocol. They can also be updated during the course of therapy, thereby describing the course of illness and providing an alternative for the evaluation of response.

This study uses the MIMIC database [SLRM02] [GAG^{+}00] for studying sepsis with different generative kernel methods. The MIMIC study was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived as the study did not impact clinical care and all data were de-identified.

The database was queried for septic patients and yielded 2,002 entries. In the experimental work reported next, we investigate the prognosis of sepsis from the Simplified Acute Physiology Score for ICU Patients (SAPS) and Sequential Organ Failure Assessment (SOFA) scores at admittance, as well as from its worst and best values (i.e. maximum and minimum) during ICU stay. The mortality rate for this study was 21.51%. Table 1 shows summary values for each of the variables considered in this work.

1.1. **Sequential Organ Failure Assessment Score (SOFA).** In 1994, the European Society of Intensive Care Medicine (ESICM) [Vin96] organized a consensus meeting in Paris to create the SOFA Score, with the aim of objectively and quantitatively describing the degree of organ dysfunction/failure over time in groups of patients or even individuals. Its two main applications are:

(1) Improving the understanding of the natural history of organ dysfunction/failure and the interrelation between the failure of various organs / systems.
(2) Assessing the effect of new therapies on the course of organ dysfunction/failure. This could be used to characterize patients at admission in the ICU (and even serve as an ICU entry criterion[1]), or to evaluate treatment efficacy.

[1] In this regard, during the 2010 flu pandemic in Australia, patients were admitted in the ICU with a maximum SOFA score of 7.

SOFA/SAPS	median (IQR)
SOFA admiss.	8 (7)
SOFA min.	3 (4)
SOFA max.	9 (8)
SAPS I admiss.	16 (8)
SAPS I min.	11 (6)
SAPS I max.	18 (9)

Table 1. SOFA/SAPS I at admission in the ICU and its maximum and minimum values during ICU stay presented as median (interquartile range)

Originally, the SOFA score was not designed to predict outcome (mortality) but to describe a series of complications on the critically ill. Although any assessment of morbidity is related to mortality to some extent, the SOFA score was not designed just to describe organ dysfunction/failure according to mortality. However, and as described elsewhere [Rib12], SOFA scores greater than 7 have important ICU outcome prediction capabilities. Moreover, when combined with additional parameters, it provides a very powerful set of predictors not only for outcome assessment but also for the study of the evolution of sepsis into its more severe states.

The SOFA limits the number of organs/systems under study to six, namely: Respiratory (inspiration air pressure), Coagulation (Platelet Count), Liver (Bilirrubine), Cardiovascular (Hypotension), Central Nervous System (Glasgow Comma Score), Renal (Creatinine or Urine Output). The scoring for each organ/system ranges from 0 for *normal function* to 4 for *maximum failure/ dysfunction*. The final SOFA score is the addition of the dysfunction indexes for all organs/systems. Therefore, the maximum possible SOFA score is 24, corresponding to maximum failure for all of the six organs/systems considered.

From a clinical perspective, a SOFA score greater than 1 corresponds to Multiple Organ Dysfunction Syndrome (MODS), while Cardiovascular SOFA scores greater than 2 correspond to Septic Shock. Normally, SOFA scores are calculated at ICU admission. However, daily calculations of SOFA scores (Dynamic SOFA) [Lev05] [Kaj05] provide valuable information about organ dysfunction evolution and prognosis. In order to expedite the calculation of the Gröbner bases presented below, the input values for SOFA have been transformed into deciles before calculating all kernels.

1.2. **Simplified Acute Physiology Score for ICU Patients (SAPS).** The SAPS uses 14 routinely measured biologic and clinical variables [Le 84] to develop a simple scoring system for the calculation of the risk-of-death (ROD) in ICU patients. Each variable is assigned a range from 0 to 4 (i.e. the score ranges from 0 to $14 \times 4 = 56$).

It has been reported that SAPS presents sensitivity and specificity of 0.69 for a cutoff value of 12 [Le 84] and a general population (i.e. for a population with a broader pathology base than sepsis). This score has been recently updated [Le 05], [MMA$^+$05a] and [MMA$^+$05b], but the version used in this paper is the only publicly available one for research purposes. This is also the version made available by the MIT team that built the MIMIC II database [SLRM02]. As for the SOFA scores, the input values for SAPS have been transformed into deciles before calculating all kernels.

2. Methods

In this section, we start by definining a regular exponential family. This representation is particularly useful since its properties result in the simplification of the Fisher kernel for multinomial families. It is just the distance to the mean between the sufficient statistics of the underlying regular exponential family. The likelihood function of regular exponential family can be reparameterized through algebraic geometry. Another interesting propery of regular exponential families is that they allow a convex dual in the Legendre sense, which corresponds to the negative Entropy. This convex-dual shall be used later as the building block of the Jensen-Shannon metric taht will be used in the kernels derived from this metric. Finally, we present the QBK, which is the main contribution of this paper. This kernel is defined as the covariance of the design matrices obtained from a Gröbner basis, which are a generative basis of a polynomial ideal.

2.1. Fisher kernel for Exponential Families.

Consider the sample space $\mathcal{X}$ with σ-algebra $\mathcal{A}$ on which a σ-finite measure υ is defined. Let $T\colon \mathcal{X} \to \mathbb{R}^k$ be a measurable map [Drt07]. Define the natural parameter space:

$$N = \left\{ \eta \in \mathbb{R}^k : \int_{\mathcal{X}} e^{\eta^t T(x)} d\upsilon(x) < \infty \right\}. \tag{1}$$

For $\eta \in N^k$, we can define a probability density p_η on $\mathcal{X}$ as

$$p_\eta(x) = e^{\eta^t T(x) - \phi(\eta)}, \tag{2}$$

where

$$\phi(\eta) = \log \int_{\mathcal{X}} e^{\eta^t T(x)} d\upsilon(x) \tag{3}$$

is the logarithm of the Laplace transform on υ^t. Here t denotes matrix/vector transpose. Let P_θ be the probability measure on $(\mathcal{X}, \mathcal{A})$ that has υ-density p_η. Define $\upsilon^t = \upsilon \circ T^{-1}$ to be the measure that the statistic T induces on the Borel σ-algebra of $\mathbb{R}^k$. The support of υ^t is the intersection of all closed sets $A \subseteq \mathbb{R}^k$ that satisfy $\upsilon^t(\mathbb{R}^k \setminus A) = 0$.

Definition 1. *Let k be a positive integer. The probability distributions $(P_\eta | \eta \in N)$ form a regular exponential family of order k if N is an open set in $\mathbb{R}^k$ and the affine dimension of the support υ^t is equal to k. The statistic $T(x)$ that induces the regular exponential family is called a canonical sufficient statistic.*

Let $\mathcal{P} = (P|\eta \in N)$ be a regular exponential family with canonical sufficient statistic T. If we draw a sample $X_1, \ldots, X_n$ of independent random vectors from P_η, then, the canonical statistic becomes $\sum_{i=1}^n T(X_i) = n\bar{T}_x$ and the log likelihood function takes the form

$$l(\eta|\bar{T}) = n\big(\eta^t \bar{T} - \phi(\eta)\big) \tag{4}$$

Definition 2 (Score Function)**.** *The Score Function is the gradient*

$$U\big(\bar{T}, \eta\big) = \frac{\partial l(\eta|\bar{T})}{\partial \eta} = n\bar{T} - \frac{\partial}{\partial \eta}\phi(\eta) \tag{5}$$

By construction of the cumulant generative function $\phi(\eta)$, we have $\zeta(\eta) = \frac{\partial}{\partial \eta}\phi(\eta)$, which is the expectation of our regular exponential family.

The information matrix is (minus) the Hessian of the log-likelihood. In this case, it is also the Fisher or expected information, since it does not depend on X:

$$\begin{aligned} \operatorname{cov}\big(U(\bar{T}, \eta)\big) &= n\frac{\partial^2}{\partial \eta^2}\phi(\eta) \\ &= E_\eta \left\{ \big(n\bar{T} - \zeta(\eta)\big)\big(n\bar{T} - \zeta(\eta)\big)^t \right\} \end{aligned} \tag{6}$$

Definition 3 ([Cri00] Fisher kernel)**.** *The Fisher kernel for a regular exponential family is defined as:*

$$k(x, z) = U(\bar{T}_x, \eta) cov(U(\bar{T}, \eta))^{-1} U(\bar{T}_z, \eta) \tag{7}$$

where T_x and T_z are the sufficient statistics estimated on x and z.

In most cases, the implementation of the Fisher kernel is computationally expensive so that, often, the following simplified (practical) Fisher kernel is implemented

Definition 4 ([Cri00] Practical Fisher kernel)**.**

$$k(x, z) = U(\bar{T}_x, \eta) U(\bar{T}_z, \eta)^t \tag{8}$$

where T_x and T_z are the sufficient statistics estimated on x and z.

Intuitively, the Fisher kernel is a function that measures the similarity of two objects on the basis of sets of measurements for each object and a statistical model. In a classification procedure, the class for a new object (whose real class is unknown) can be estimated by minimising, across classes, an average of the Fisher kernel distance from the new object to each known member of the given class. For multinomial families, the Fisher kernel for exponential famililies is quite simple since it only requires the calculation of the distance towards the mean as shown in algorithm 1.

Algorithm 1 Pseudocode of the Practical Fisher kernel for Multinomial Distributions

```
Input: $x$ and $z$
Output: Fisher kernel $k(x, z)$
  $\mu_X \leftarrow$ mean(x)
  $\mu_Z \leftarrow$ mean(z)
  for $i = 1 \cdots N_x$ do
    for $j = 1 \cdots N_x$ do
      $k(i, j) \leftarrow (T_{x_i} - \mu_x)\left(T_{z_j} - \mu_z\right)^t$ {Product of distances of each point to their mean}
    end for
  end for
```

2.2. **Kernels based on the Jensen-Shannon metric.** For maximum likelihood estimation on a regular exponential family $P_M = (P_\eta, \eta \in M)$, $M \subseteq N$, we need to maximize $l(\eta|\bar{T})$ over the set M. Let A and g be the semi-algebraic set and the diffeomorfism that define the parameter space M. Let $I(A) = (f_1, \ldots, f_m)$ be the ideal of model invariants and let $\gamma = g(\eta)$ the parameters after re-parametrization by g [Drt07]. Then, the maximization problem can be relaxed to

$$\begin{array}{ll} \max & l(\gamma|\bar{T}) \\ \text{s.t.} & f_i = 0 \quad i = 1, \ldots, m, \end{array} \tag{9}$$

where $l(\gamma|\bar{T}) = g^{-1}(\gamma)^t\bar{T} - \phi(g(\gamma)^{-1})$. In our case, we work with the probability simplex as a semi-algebraic set [Drt07] for discrete random variables, which is a convex polyhedron in any dimension. Therefore, the optimization problem (9) is convex. It is important to note that this algebraic representation agrees with the standard theory and it can be represented as a Bregman divergence as we will show below.

Let F be the convex-dual in the Legendre sense of the partition function G. A Bregman divergence is defined as:

Definition 5. [Aga11]*Bregman divergence*

$$\begin{aligned} B_F(\bar{T}||\nabla\phi(g^{-1}(\gamma_i))) = F(\bar{T}) - F(\nabla\phi(g^{-1}(\gamma_i)) \\ - \nabla F(\nabla\phi(g^{-1}(\gamma_i))) \cdot (\bar{T} - \nabla\phi(g^{-1}(\gamma_i))). \end{aligned} \tag{10}$$

By the Legendre dual we have

$$F\big(\nabla\phi(g^{-1}(\gamma)\big) = \nabla\phi\big(g^{-1}(\gamma)\big)g^{-1}(\gamma) - \phi\big(g^{-1}(\gamma)\big) \tag{11}$$

Also, F and G are Legendre functions if their derivatives are inverse functions of each other (i.e. $\nabla F(\nabla\phi(g^{-1}(\gamma)) = g^{-1}(\gamma))$. Since $F(\bar{T})$ does not depend on the parametrization, our optimization problem becomes:

$$\begin{aligned}\max l\big(\gamma\,|\,\bar{T}\big) &= \max\left\{F(\bar{T}) - \sum_{i=1}^{m} B_F(\bar{T}||\nabla\phi(g^{-1}(\gamma_i))\right\}\\ &= \min\left\{\sum_{i=1}^{m} B_F\big(\bar{T}\,||\,\nabla\phi\big(g^{-1}(\gamma_i)\big)\right\}\\ &\quad \text{s.t. } f_i = 0 \; i = 1,\dots,m\end{aligned} \tag{12}$$

In this respect, we can apply the idea that given new facts x_k, a new distribution parametrized by η_i should be chosen which is as hard to discriminate from the original parametrization η as possible, so that the new data produces as small an information gain in $KL(\eta_i\|\eta)$ or $B_F(\bar{T}||\nabla\phi(g^{-1}(\gamma_i))$ as possible [2]. In other words, what we want to achieve is the minimum of the cross-entropy. This approach was already exploited by Kullback and Leibler in [Kul51] and termed it *Principle of Minimum Discrimination Information* (MDI).

Therefore, it is now natural to use the Jensen-Shannon divergence[3] as a metric in order to build kernels that exploit the generative properties of the data. As opposed to [Aga11], the main contribution here is that we are bridging together the use of semi-algebraic sets (which are needed for the parametrization) and the dual structure induced by the diffeomorfism g that re-parametrizes the optimization problem.

Now, we only have to apply the Jensen-Shannon metric over the dual space. More particularly,

Definition 6. [Aga11] [Ber84] [I.J38] *Let* $\gamma_1, \gamma_2 \in M$*:*

$$JS(\gamma_1,\gamma_2) = \frac{F(\gamma_1)+F(\gamma_2)}{2} - F\left(\frac{\gamma_1+\gamma_2}{2}\right). \tag{13}$$

Proposition 1. [Aga11] [Ber84] [I.J38] *centred kernel let* $x_0 \in X$ *define the centred kernel as* $\psi\colon X\times X \to \mathbb{R}$

$$\begin{aligned}\psi(x,y) = JS(x,x_0) + JS(y,x_0)\\ - JS(x,y) - JS(x_0,x_0).\end{aligned} \tag{14}$$

Proposition 2. [Aga11] [Ber84] [I.J38] *exponentiated kernel we define the exponentiated kernel as* $\psi\colon X\times X \to \mathbb{R}$

$$\psi(x,y) = \exp\big(-tJS(x,y)\big) \tag{15}$$

$\forall t > 0.$

[2] KL is a Bregman divergence

[3] remember that the KL divergence is not a metric

Proposition 3. [Aga11] [Ber84] [I.J38] *inverse kernel we define the inverse kernel as* ψ: $X \times X \to \mathbb{R}$

$$\psi(x,y) = \frac{1}{t + JS(x,y)} \tag{16}$$

$\forall t > 0$.

It is obvious that the most important part to calculate the kernels outlined above is the calculation of the Jensen-Shannon metric in dual-space. The pseudocode to implement this metric is provided in algorithms 2 and 3.

Algorithm 2 Pseudocode for the computation of the Jensen-Shannon Metric for Multinomial Distributions

```
Input: $x$ and $z$
Output: Dual $JS(\gamma_i, \gamma_j)$
  for $i = 1 \cdots N_x$ do
    for $j = 1 \cdots N_z$ do
      $\gamma_1 \leftarrow x(i,:)$
      $\gamma_2 \leftarrow z(j,:)$
      Compute the duals F (see algorithm 3)
      $JS(\gamma_i, \gamma_j) \leftarrow \frac{F(\gamma_i)+F(\gamma_j)}{2} - F\left(\frac{\gamma_i+\gamma_j}{2}\right)$ {Compute JS from Duals}
    end for
  end for
```

Algorithm 3 Pseudocode to Compute Duals for Multinomial Distributions

```
Input: Vector $\gamma_x$
Output: Dual $F(\gamma_x)$
  $N = \sum \gamma_x$
  $F \leftarrow \gamma_x \log(\frac{\gamma_x}{N})$
```

2.3. **Quotient Basis Kernel.** In this section we present the definition of algebraic models as given in [PRW01], where inputs are denoted by x, responses or outputs are denoted by y, parametric functions denoted by η or functions of η. These are related by polynomial algebraic relations, possibly implicit. Another feature of this definition is that constraints of polynomial type can be included in the specification of the model. Implicit models and the introduction of constraints can lead to the use of dummy variables.

The parameters of the model as interpreted in statistics are functions of any form with the restriction that they belong to a specified field. For example, $\mathbb{Q}\left(\eta_1, \ldots, \eta_p\right)$ is the set of all rational functions in $\eta_1, \ldots, \eta_p$ with rational coefficients. Another example is $\mathbb{Q}\left(e_1^{\eta}, \ldots, e_p^{\eta}\right)$ the set of all exponential rational functions. Parameters are treated as unknown quantities and in most cases appear in linear form. The algebraic space used is the commutative ring of all polynomials $\mathbb{K}[x_1, \ldots, x_s]$ in the indeterminates $x_1, \ldots, x_s$ and with coefficients in the field $\mathbb{K}$ (in our case $\mathbb{R}$).

For a given initial ordering, a term is specified by the vector of length s of its exponents. Therefore *Term*$\{s\}$ is coded by $\mathbb{Z}_+^s$ [PRW01] (set of positive integers).

When the indeterminates are indexed from 1 to s so that $x_1, \ldots, x_s$, it is convention to consider an initial ordering $x_i \succ x_{i+1}$ $\forall i = 1 \ldots s-1$.

Definition 7 (Polynomial Ideal). (1) *A polynomial ideal I is a subset of a polynomial ring $\mathbb{K}[x]$ closed under sum and product by elements of $\mathbb{K}[x]$. Specifically the set $I \subset \mathbb{K}$ is an ideal if $\forall f, g \in I$ and $s \in \mathbb{K}$ the polynomials $f + g$ and sf are in I.*

(2) *Let F be a set of polynomials. The ideal generated by F is the smallest ideal containing F. It is denoted $\langle F \rangle$.*

(3) *An ideal I is radical if $f \in I$ whenever a positive integer m exists such that $f^m \in I$.*

(4) *The radical of an ideal I is the radical ideal defined as* $\sqrt{I} = \{f \in \mathbb{K} : \exists m \,|\, f^m \in I\}$

The Hilbert basis theorem ([PRW01]) shows that every ideal has a finite basis. This provides a very powerful result since it means that any ideal is finitely generated (even if the generating set is not necessarily unique). Another powerful result is that this generation basis is of a special type called Gröbner Basis, which we define below. This bases will become essential in the derivation of regression/interpolation polynomials and also for the algebraic derivation of the Fisher and the proposed QBK kernels.

Definition 8. [PRW01] *Let τ be a term ordering on $\mathbb{K}[x]$ and f a polynomial in $\mathbb{K}[x]$. The leading term of f, $LT_\tau(f)$ is the largest term with respect to τ among the terms in f.*

Definition 9. [PRW01] *Gröbner Basis: Let τ be a term ordering on $\mathbb{K}[x]$. A subset $G = g_1, \ldots, g_t$ of an ideal I is a Gröbner basis of I with respect to τ iff*

$$\big\langle LT_\tau(g_1), \ldots, LT_\tau(g_t) \big\rangle = \big\langle LT_\tau(I) \big\rangle \tag{17}$$

where $LT_\tau(I) = \{LT_\tau(f) : f \in I\}$.

Theorem 1. [Gig00] [PRW01] *Given a term ordering, every ideal I except* $\{0\}$ *has a Gröbner basis and any Gröbner basis is a basis of I.*

Definition 10. [PRW01] *Ideal of a set of support points: Let A be a set of unique support points $A = \{\mathbf{a_1}, \ldots, \mathbf{a_n}\}$. The set $I(A)$ is the set of all polynomials whose zeros include the points in A.*

Definition 11. *Gröbner basis of unique points* [PRW01], [Gig00]*: Let A be a set of n unique points $A = \{\mathbf{a_1}, \ldots, \mathbf{a_n}\}$ and τ a term ordering. A Gröbner basis of*

A, $G = g_1, \dots, g_t$, is a Gröbner basis of $I(A)$. Therefore, the points in A can be presented as the set of solutions of

$$\left\{\begin{array}{c} g_1(\mathbf{a}) = 0 \\ g_2(\mathbf{a}) = 0 \\ \cdots \\ g_t(\mathbf{a}) = 0 \end{array}\right. \tag{18}$$

Let us formally define the Quotient Basis EST_τ that shall be used in the algorithm below.

Definition 12 ([PRW01] Quotient Basis)**.** *Let A, be a set of $n \times s$ unique support points $A = \{\mathbf{a_1}, \dots, \mathbf{a_n}\}$ and τ a term ordering. A monomial basis of the set of polynomial functions over A is*

$$EST_\tau = \left\{x^\alpha : x^\alpha \notin \left\langle LT(g) : g \in I(A)\right\rangle\right\} \tag{19}$$

This definition states that EST_τ comprises the elements x^α that are not divisible by any of the leading terms of the elements of the Gröbner basis of $I(A)$.

Theorem 2 ([PRW01])**.** *The set EST_τ has as many elements as there are support points.*

Definition 13 (Design Matrix)**.** *Let τ be a term ordering and let us consider an ordering over the support points $A = \{\mathbf{a_1}, \dots, \mathbf{a_n}\}$. We call design matrix (i.e. EST_τ evaluated in A) the following $n \times c$ matrix*

$$Z = \left[EST_\tau\right]\Big|_A \tag{20}$$

where c is the cardinality of EST_τ and n is the number of support points.

Proposition 4. *The covariance of Z,*

$$cov(Z) = E\left\{\left(Z - E(Z)\right)\left(Z - E(Z)\right)^t\right\}$$

is a kernel.

Definition 14 (Quotient Basis Kernel (QBK))**.** *The covariance of the design matrix of EST_τ, which is a kernel, is the QBK.*

The algorithm for the calculation of EST_τ, which shall be used to calculate our QBK from the design matrix Z is as follows:

1. Input: matrix with unique points A and relative frequencies q. Without loss of generality this matrix could also be a transformed version of A by means of a kernel.
2. Define a term ordering τ (for example lexicographic).
3. Calculate the ideal of matrix A (definition 12). In our case, this is done with ApCoCoA [CoC].
4. Calculate the reduced Gröbner Basis G (this can be also calculated with the function IdealOfPoints [ABKR00] in ApCoCoA).

(5) Identify the subset EST_τ (i.e. identify the sub-set of monomials not divided by G).

(6) Let L be the logarithm of the monomials of EST_τ (i.e. exponents). Write the design matrix $Z = [\mathrm{EST}_\tau]\big|_A$ with the terms of EST_τ.

This algorithm was originally developed for the derivation of interpolation/ regression polynomials in [Gig00].

Algorithm 4 Pseudocode of the Quotient Basis Kernel

```
Input: $A$ Input dataset, $x$, $y$ and $EST_\tau$
Output: Quotient Basis Kernel $k(x, y)$
  $\mu \leftarrow \mathrm{mean(A)}$
  $Z_x \leftarrow [\mathrm{EST}_\tau]\big|_x$ {Subs. $x$ in the design matrix calculated from $EST_\tau$}
  $Z_y \leftarrow [\mathrm{EST}_\tau]\big|_y$ {Subs. $y$ in the design matrix calculated from $EST_\tau$}
  $k(x, y) \leftarrow (Z_x - \mu)(Z_y - \mu)^t$
```

2.4. **Overview of Support Vector Machines.** In this paper, the performance of the different kernels described in the previous sections (namely Fisher, exponential, inverse, centred, Gaussian, polynomial of order 2, linear and the proposed QBK) is compared in the task of mortality prediction using soft-margin SVM [Cri00]. This technique is well-suited to the structure of the problem at hand, given that the relation of SAPS and SOFA with mortality is not linear. Also, high values of SAPS and SOFA result in a higher ROD probability which means that some patients will survive despite the severity of their illness (i.e. some points will fall in the 'wrong' side of the margin boundary). At this stage it is also worth noting that SAPS and SOFA overlap in two variables: blood pressure and central nervous system assessment.

In this approach, the objective is to obtain a hyper-surface separating the training points $\mathbf{x_1}, \ldots, \mathbf{x_N}$ into two disjoint sets. Soft-Margin SVMs [Cri00] let some points fall on the incorrect side of the margin boundary by introducing a penalty that increases with the distance from this margin (i.e. the greater the misclassification, the bigger the error). This is achieved by solving the following quadratic problem:

$$\begin{gathered} \arg\max_\alpha \left(\sum_{i=1}^{N} \alpha_i - \tfrac{1}{2}\alpha^t \mathbf{H} \alpha \right) \\ \text{s.t. } 0 < \alpha_i < C \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0. \end{gathered} \tag{21}$$

where $H_{ij} = y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function. The parameter C controls the trade-off between the penalty and the size of the margin. Therefore, it can also be interpreted as a factor controlling the number of support vectors.

3. Results

The performance of the soft-margin SVM with the different kernels listed in the previous section was tested using 10-fold cross-validation to obtain the classifiers, which were evaluated over a stratified test dataset. The latter was obtained by taking 10% out-of-sample data before the cross-validation and was used to evaluate all models. Table 2 shows the results over this test dataset.

kernel	Correct Rate	Sens.	Spec	AUC
Quotient	80.12%	79.20%	83.15%	81.00%
Fisher	77.84%	73.09%	81.74%	78.10%
Exponential	61.33%	57.36%	75.84%	66.69%
Inverse	61.13%	51.96%	76.33%	67.05%
Centred	74.29%	72.05%	82.44%	66.99%
Gaussian	75.33%	73.22%	83.02%	80.27%
Poly (order 2)	75.11%	72.90%	83.15%	80.60%
Linear	75.05%	72.87%	82.99%	81.26%

Table 2. Results for SVM with Generative kernels

The QBK is calculated with algorithm 4 and the kernels based on the Jensen Shannon metric are calculated with definitions 1, 2 and 3 from algorithm 2.2. The simplified Fisher kernel is calculated with algorithm 1. The standard linear, Gaussian and polynomial kernels have also been tested. We used Matlab® SVM QP solver as implemented in the BioInformatics and Optimization Toolboxes. The cross-validation experiment (during train and validation) yielded the appropriate values for C parameters for each kernel. More in particular,

- for the Quotient Basis kernel $C = 1$.
- for the Fisher kernel, $C = 1$.
- for the kernels based on Jensen-Shannon metric, $C = 10$. Besides, the t parameter for the exponential and inverse kernels was set to a value of 0.2.
- for the Gaussian, linear and polynomial kernels, $C = 10$.

Table 2 shows that the QBK consistently yielded the best results in terms of all parameters considered: accuracy, sensitivity, specificity and AUC(which, in a way, summarizes sensitivity and specificity). The Fisher kernel yielded a similar accuracy to the QBK but with lower sensitivity, specificity and AUC. In our study we have not found significant differences in accuracy for the centred, linear, gaussian and polynomial kernels. Finally, the other two kernels based on the Jensen-Shannon metric (inverse and exponential) yielded the worst results for all parameters considered in this study.

Since we have a significant number of patients, it is possible to evaluate the basal SAPS as a prognostic factor for sepsis. This has been done following the methodology presented in [Le 84] where a threhold is selected to assess whether the patient survives. The work from Le Gall et at. [Le 84] selects a threshold of 12, which yields a sensitivity and specificity of 0.69. In our case, we have automatically selected the threshold that resulted in the maximum accuracy in assessing the risk of death. At this stage, it is important to note that our population is not as general as that presented in [Le 84] since it only comprises septic patients. In our case, the threshold that we have obtained is 19.5 for the basal SAPS score (i.e. SAPS at ICU admittance). This threshold yielded an accuracy of .75, a sensitivity of .51 and a specificity of .81. Despite the fact that the accuracy and specificity are similiar to those obtained with our kernels, the resulting sensitivity is quite low. Its poor sensitivity may be the result of the SAPS score including non-sepsis specific clinical traits (for example, the performance of haemocultures, antibiotic administration or vasoactive drug administration [Rib11]. Therefore, a combined approach between SOFA and SAPS makes perfect sense for our population and the problem of assessing risk of death.

4. Discussion

Sensitivity and specificity are important measures of performance for the task of mortality prediction for septic patients. This is due to the fact that more aggressive treatment and therapeutic actions may result in better outcomes for high-risk patients.

The SVMs in the experiments reported in this paper were trained with eight different kernels, out of which five were generative and the other three were kernels that were considered well-suited to the problem at hand. The investigated kernels have been, overall, shown to provide accurate and medically actionable results, whilst keeping an acceptable balance between the different parameters of interest (accuracy rate, sensitivity, specificity and AUC).

The new kernels proposed in the study, QBK and simplified Fisher kernel, respectively defined through the Gröbner basis of an algebraic ideal and the sufficient statistics of a regular exponential family have been shown to outperform not only the alternative kernels, but also the clinical standard method based on the SAPS score in the problem of mortality prediction for septic patients.

The Fisher kernel has been derived by means of a combination of Algebraic Models and well established properties from the regular exponential families. To the best of our knowledge the simplified Fisher kernel presented in this paper is a non-obvious result of the application of regular exponential families for the definition of kernels.

Even though the QBK yields the best results in terms of accuracy and balance between sensitivity and specificity a word of caution must be given regarding the computation time of this kernel. For high-dimensional datasets or very big input matrices, the calculation of a Gröbner basis can be very time-consuming. Therefore, for large datasets we propose to use the simplified Fisher kernel proposed here.

References

[ABKR00] J. Abbott, A. Bigatti, M. Kreuzer, and L. Robbiano. Computing ideals of points. *JSYMC*, 30(4):341–356, 2000.

[Aga11] Agarwal A., Daum III H. Generative kernels for exponential families. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[Ame92] American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Crit Care Med.*, 20:864–874, 1992.

[Ang01] Angus D.C., Linde-Zwirble W.T, Lidicker J., Clermont Gl, Carcillo J., Pinski M.R. Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Crit Care Med*, 29(7):1303–1310, 2001.

[Ber84] Berg C. and Christensen J.P.R and Ressel P. *Harmonic analysis on semigroups.* Springer-Verlag, New-York, 1984.

[CoC] CoCoATeam. CoCoA: a system for doing Computations in Commutative Algebra. Available at `http://cocoa.dima.unige.it`.

[Cri00] Cristianini N., Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge, U.K., 2000.

[Drt07] Drton M., Sullivant S. Algebraic statistical models. *Statist. Sinica.*, 17:1273–1297, 2007.

[Est07] Esteban A.,Frutos-Vivar F., Ferguson N., Penuelas O., Lorente J.Al, Gordo F., Honrubia T., Algora A., Bustos A., Garcia G., Rodriguez I., Ruiz. R. Sepsis incidence and outcome: contrasting the intensive care unit with the hospital ward. *Crit Care Med*, 35(5):1284–1289, 2007.

[GAG$^+$00] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[Gig00] Giglio B., Riccomagno E., Wynn H. Grbner basis strategies in regression. *Journal of Applied Statistics*, 27(7):923–938, 2000.

[I.J38] Schoenberg I.J. Metric spaces and positive definite functions. In *Transactions of the American Mathematical Society*, volume 44, pages 522–536, 1938.

[Kaj05] Kajdacsy-Balla A.C., Moreira Andrade F., Moreno R., Artigas A., Cantraine F., Vincent J.L. Use of the sequential organ failure assessment score as a severity score. *Intensive Care Med*, 33(10):2194–2201, 2005.

[Kul51] Kullback S., Leibler R.A. On information and sufficiency. *Annals of Mathematical Statistics*, 22:49–86, 1951.

[Le 84] Le Gall J.R., Loirat P., Alperovitch A., Glaser P., Granthil C., Mathieu D., Mercier P., Thomas R., Villers D. A simplified acute physiology score for icu patients. *Crit. Care*, 12(11):975–977, 1984.

[Le 05] Le Gall J.R., Neuman F.H., Bleriot J.P., Fulgencio J.P., Garrigues B., Gouzes C., Lepage E., Moine P., Villers D. Mortality prediction using SAPS II: an update for French intensive care units. *Crit. Care*, 9(6):R645–R652, 2005.

[Lev03] Levy M.M., Fink M.P., Marshall J.C., Edward Angus A., Cook D., Cohen J., Opal S.M., Vincent J.L., Ramsay G. for the International Sepsis Definitions Conference (2003). 2001 SCCM/ESICM/ACCP/ATS/SIS International sepsis definitions conference. *Int. Care Med.*, 29:530–538, 2003.

[Lev05] Levy M.M., Macias W.L., Vincent J.L., Russell J.A., Silva E., Trzaskoma B., Williams D. Early changes in organ function predict eventual survival in severe sepsis. *Crit. Care Med*, 31:243–249, 2005.

[Mar03] Martin G.S., Mannino D.M., Eaton S., Moss M. The epidemiology of sepsis in the united states from 1979 through 2000. *N Engl J Med.*, 348:1546–1554, 2003.

[MMA$^+$05a] Philipp Metnitz, Rui Moreno, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, Jean-Roger Le Gall, and . Saps 3 from evaluation of the patient to evaluation of the intensive care unit. part 1: Objectives, methods and cohort description. *Intensive Care Medicine*, 31:1336–1344, 2005.

[MMA$^+$05b] Rui Moreno, Philipp Metnitz, Eduardo Almeida, Barbara Jordan, Peter Bauer, Ricardo Campos, Gaetano Iapichino, David Edbrooke, Maurizia Capuzzo, Jean-Roger Le Gall, and . Sap3 from evaluation of the patient to evaluation of the intensive care unit. part 2: Development of a prognostic model for hospital mortality at icu admission. *Intensive Care Medicine*, 31:1345–1355, 2005.

[Pac05] Pachter L., Sturmfels B. *Algebraic Statistics for Computational Biology.* Cambridge University Press, 2005.

[PRW01] G. Pistone, E. Riccomagno, and H.P. Wynn. *Algebraic Statistics: Computational Commutative Algebra in Statistics*, volume 89 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, CRC Press, Boca Raton, 2001.

[Rib11] Ribas V.J., Ruiz-Rodríguez J.C., Wojdel A., Caballero-López J., Ruiz-Sanmartín A., Rello J., and Vellido A. Severe sepsis mortality prediction with relevance vector machines. In *Procs. of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 100–103, 2011.

[Rib12] Ribas V.J., Vellido A., Ruiz-Rodríguez J.C., Rello J. Severe sepsis mortality prediction with logistic regression over latent factors. *ESWA*, 39(2):1937–1943, 2012.

[SLRM02] M. Saeed, C. Lieu, G. Raber, and R.G. Mark. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. *Comput Cardiol*, 29, 2002.

[Vin96] Vincent J.L., Moreno R., Takala J., Willats S., De Mendoa A., Bruining H., Reinhart C.K., Suter P.M., Thijs L.G. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Crit. Care Med*, 22:707–710, 1996.

Vicent J. Ribas Ripoll
Centre de Recerca Matemàtica
Campus de Bellaterra, Edifici C
08193 Bellaterra (Barcelona)

Alfredo Vellido, Enrique Romero
Soft Computing (SOCO) Research Group
Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
Edifici Omega, Campus Nord
08034, Barcelona - Spain

Juan Carlos Ruiz-Rodríguez *MD*
Critical Care Department, SODIR Research Group
Vall d'Hebron University Hospital
Vall d'Hebron Research Institute
Autonomous University of Barcelona

CENTRE DE RECERCA MATEMÀTICA

Facultat de Ciències, UAB
08193 Bellaterra
Barcelona, Spain
www.crm.cat