Artificial Intelligence in Medicine 61 (2014) 45–52

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Sepsis mortality prediction with the Quotient Basis Kernel

Vicent J. Ribas Ripoll[a,*], Alfredo Vellido[b], Enrique Romero[b], Juan Carlos Ruiz-Rodríguez[c]

[a] *Centre de Recerca Matemàtica, Campus de Bellaterra, Edifici C, 08193 Bellaterra (Barcelona), Spain*
[b] *Soft Computing (SOCO) Research Group, Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Edifici Omega, Campus Nord, 08034 Barcelona, Spain*
[c] *Shock Organ Dysfunction and Resuscitation (SODIR) Research Group, Critical Care Service, Vall d'Hebron University Hospital, Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona, Passeig de la Vall d'Hebron, 119-129, 08035 Barcelona, Spain*

ARTICLE INFO

Article history:
Received 7 April 2013
Received in revised form 12 March 2014
Accepted 16 March 2014

Keywords:
Kernels
Support vector machines
Sepsis
Mortality prediction
Critical care

ABSTRACT

Objective: This paper presents an algorithm to assess the risk of death in patients with sepsis. Sepsis is a common clinical syndrome in the intensive care unit (ICU) that can lead to severe sepsis, a severe state of septic shock or multi-organ failure. The proposed algorithm may be implemented as part of a clinical decision support system that can be used in combination with the scores deployed in the ICU to improve the accuracy, sensitivity and specificity of mortality prediction for patients with sepsis.

Methodology: In this paper, we used the Simplified Acute Physiology Score (SAPS) for ICU patients and the Sequential Organ Failure Assessment (SOFA) to build our kernels and algorithms. In the proposed method, we embed the available data in a suitable feature space and use algorithms based on linear algebra, geometry and statistics for inference. We present a simplified version of the Fisher kernel (practical Fisher kernel for multinomial distributions), as well as a novel kernel that we named the Quotient Basis Kernel (QBK). These kernels are used as the basis for mortality prediction using soft-margin support vector machines. The two new kernels presented are compared against other generative kernels based on the Jensen–Shannon metric (centred, exponential and inverse) and other widely used kernels (linear, polynomial and Gaussian). Clinical relevance is also evaluated by comparing these results with logistic regression and the standard clinical prediction method based on the initial SAPS score.

Results: As described in this paper, we tested the new methods via cross-validation with a cohort of 400 test patients. The results obtained using our methods compare favourably with those obtained using alternative kernels (80.18% accuracy for the QBK) and the standard clinical prediction method, which are based on the basal SAPS score or logistic regression (71.32% and 71.55%, respectively). The QBK presented a sensitivity and specificity of 79.34% and 83.24%, which outperformed the other kernels analysed, logistic regression and the standard clinical prediction method based on the basal SAPS score.

Conclusion: Several scoring systems for patients with sepsis have been introduced and developed over the last 30 years. They allow for the assessment of the severity of disease and provide an estimate of in-hospital mortality. Physiology-based scoring systems are applied to critically ill patients and have a number of advantages over diagnosis-based systems. Severity score systems are often used to stratify critically ill patients for possible inclusion in clinical trials. In this paper, we present an effective algorithm that combines both scoring methodologies for the assessment of death in patients with sepsis that can be used to improve the sensitivity and specificity of the currently available methods.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Sepsis is a clinical syndrome defined by the presence of both infection and systemic inflammatory response syndrome (SIRS). Sepsis can lead to severe sepsis, which is defined by organ dysfunction or an even more severe condition, septic shock and multi-organ failure [1].

This pathology has clearly increased over the last 20 years, rising to 750,000 cases per year in the United States of America alone. As the population ages and treatment becomes more aggressive, this figure is likely to grow [2,3]. In western health systems, patients with sepsis account for as high as 25% of bed utilisation in the

* Corresponding author. Tel.: +34 680435939.
E-mail address: vribas@crm.cat (V.J. Ribas Ripoll).

http://dx.doi.org/10.1016/j.artmed.2014.03.004
0933-3657/© 2014 Elsevier B.V. All rights reserved.

intensive care unit (ICU), and the pathology occurs in 1–2% of all hospitalisations. The mortality rates of sepsis are very high, ranging from 12.8% for sepsis to 45.7% for septic shock [4].

These figures alone justify the need for a quantitative approach to predict mortality due to sepsis in the ICU. The extreme demands of this clinical environment further require prediction methods that are both robust and feasible within the constraints of a busy ICU.

In this paper, we describe a novel sepsis mortality prediction method that embeds the available data in a suitable feature space and uses algorithms based on linear algebra, geometry and statistics for inference. More specifically, we present a novel kernel for multinomial distributions, the Quotient Basis Kernel (QBK), which is based on the re-parametrisation 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 to build a set of efficient, closed-form kernels that are best suited for multinomial distributions.

The QBK is the result of calculating the covariance of the design matrix of a Gröbner basis. In this paper, we hypothesise that the QBK is particularly well suited for predicting sepsis-related mortality. Not only does it exploit the inner structure of the data (i.e., it is generative), but it also provides a geometric representation accounting for the inner dependencies between its inputs [5] (in our case, these inputs are the Sequential Organ Failure Assessment (SOFA) and Simplified Acute Physiology Score (SAPS) at ICU admission).

This representation is closely related to graphical models [6] in such a way that these kernels could be considered *open-box* methods.

We compared the performance of the proposed QBK method for the prediction of mortality due to sepsis to methods obtained using a number of alternative kernels in the well-known multiparameter intelligent monitoring in intensive care (MIMIC II) database using soft-margin support vector machines (SVMs) [7]. We also compared the QBK to a standard method used in clinical practice that is based on the basal SAPS score [8] (i.e., through the automatic selection of a threshold) and logistic regression.

The paper is organised as follows: Section 2 presents the database used in this study along with the two main indices used for mortality prediction: the SOFA and SAPS scores. In Section 3, 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 [9], with a special emphasis on reparametrisation of the log-Laplace transform term of a regular exponential family. The section closes with the formal definition of the novel QBK and a short overview of SVMs. In Section 4, we show the experimental prediction results for each different kernel and their comparison with standard mortality prediction based on the basal SAPS score.

2. Background

In normal clinical practice, clinicians often treat severely ill patients in the later stages of sepsis or in its more severe manifestations. In many cases, these patients may be suffering from a combination of chronic and acute disease.

Illness scoring systems are commonplace in the ICU. The rationale for using these systems in the 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 each patient to describe the relative risks they face and identify where the patient can be located along the continuum of illness severity.

It is increasingly evident that the ultimate goal of severity scoring is more than just obtaining an overall figure representing the degree of physiological 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 that a patient presents for care or at the time of entry into a clinical trial. Therefore, scoring systems can serve as a pre-treatment protocol. Moreover, they can also be updated during the course of therapy to describe the course of illness and provide an alternative for the evaluation of treatment response.

2.1. *Sequential Organ Failure Assessment Score*

In 1994, the European Society of Intensive Care Medicine (ESICM) [10] organised 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 individuals. The following represent the two main applications of this system:

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 to characterise patients at admission into the ICU (and even serve as an ICU entry criterion) or evaluate treatment efficacy.

Originally, the SOFA score was not designed to predict outcome (mortality) but to describe a series of complications of the critically ill. Although any assessment of morbidity is related to mortality to some extent, the SOFA score was not designed to describe organ dysfunction/failure according to mortality. However, and as described elsewhere [11], a SOFA score greater than 7 has 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.

SOFA limits the number of organs/systems to six: respiratory (inspiration air pressure), coagulation (platelet count), liver (bilirubin), cardiovascular (hypotension), central nervous system (Glasgow coma score), and 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 six organs/systems.

From a clinical perspective, a SOFA score greater than 1 corresponds to multiple organ dysfunction syndrome (MODS), and 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) [12,13] provide valuable information regarding organ dysfunction evolution and prognosis. To expedite the calculation of the Gröbner bases presented below, the input values for SOFA have been transformed into deciles before calculating all kernels in the reported experiments.

2.2. *Simplified Acute Physiology Score for ICU patients*

The SAPS uses 14 routinely measured biologic and clinical variables [8] to develop a simple scoring system to calculate 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 $14 \times 4 = 56$).

Table 1
Data summary.

	% Unit	Median	IQR
Gender (male)	56.3		
Age		65.2	26.1
ICU length of stay		2.2	3.3
Mechanical ventilation	47.3		
Invasive blood pressure	55.5		
Vasoactive medications	34.3		
ICU mortality	21.1		

It has been reported that SAPS presents a sensitivity and specificity of 0.69 for a cut-off value of 12 [8] and a population with a broader pathology base than sepsis. This score was recently updated [14–16], but the version used in this paper is the only publicly available option for research purposes. The version used here was made available by the Massachusetts Institute of Technology (MIT) team that built the MIMIC II database [7]. Regarding the SOFA scores, the input values for SAPS were transformed into deciles before calculating all kernels.

2.3. The MIMIC II database

MIMIC II data were collected from Boston's Beth Israel Deaconess Medical Center (BIDMC) over a seven-year period beginning in 2001, and the data were collected as part of a Bioengineering Research Partnership (BRP) grant.

The project was formally established in 2003, encompassing an interdisciplinary team from MIT, industry (Philips medical systems) and clinical medicine (Beth Israel Deaconess Medical Center), with the objective of developing and evaluating advanced ICU patient monitoring systems that would substantially improve the efficiency, accuracy and timeliness of clinical decision making in intensive care. The requirement for individual patient consent was waived, as the study did not impact clinical care and all data were de-identified.

The database was accessed in March 2012 and searched for patients with sepsis. This search yielded 2002 entries with no missing values. In our experimental work, we investigate the prognosis of sepsis from the SAPS and SOFA scores for ICU patients at ICU admission. The mortality rate for this study was 21.10%. The average length-of-stay (LoS) in the ICU was 4.7 days, 56.3% of patients were male, 47.3% received ventilatory support and ICU mortality was 21.51%. Tables 1 and 2 summarise the characteristics of the analysed data.

3. Methods

In this section, we start by defining a regular exponential family. This representation is particularly useful because its properties simplify the Fisher kernel for multinomial families. The Fisher kernel is calculated as the distance to the mean of the sufficient statistics of the underlying regular exponential family. The likelihood function of the regular exponential family can be parametrised using algebraic geometry. Another interesting property of regular exponential families is that they allow a convex dual in the Legendre sense, which corresponds to negative entropy. This convex-dual will be used later as the building block by which the Jensen–Shannon metric is applied to the kernels derived from this metric.

Table 2
Sequential organ failure and simplified acute physiology score at admission in the intensive care unit presented as median (interquartile range).

SOFA/SAPS	Median (IQR)
SOFA admission	8 (7)
SAPS I admission	16 (8)

Finally, we present the QBK, which is the main theoretical contribution of this study. This kernel is defined as the covariance of the design matrices obtained from a Gröbner basis, which is a generative basis of a polynomial ideal.

3.1. 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 $T : \mathcal{X} \to \mathbb{R}^k$ be a measurable map [17]. 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 the matrix/vector transpose. Let P_θ be the probability measure on $(\mathcal{X}, \mathcal{A})$ that has the 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 $\upsilon^t(\mathbb{R}^k \backslash A) = 0$.

Definition 1 *([17]).* 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 referred to as 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, the canonical statistic becomes $\sum_{i=1}^{n} T(X_i) = n\bar{T}_x$ and the log likelihood function takes the form

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

Definition 2 *([17] Score function).* The score function is the gradient

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

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, because it does not depend on X:

$$\mathrm{cov}(U(\bar{T}, \eta)) = n\frac{\partial^2}{\partial \eta^2}\phi(\eta) = E_\eta\{(n\bar{T} - \zeta(\eta))(n\bar{T} - \zeta(\eta))^t\} \tag{6}$$

Definition 3 *([18]).* The Fisher kernel for a regular exponential family is defined as:

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

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 the following simplified (practical) Fisher kernel is often implemented.

Definition 4 *([18]).* 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, respectively.

Intuitively, the Fisher kernel is a function that measures the similarity of two objects based on 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 an average of the Fisher kernel distance, across classes, from the new object to each known member of the given class. For multinomial families, the Fisher kernel for exponential families is quite simple because it only requires calculation of the distance to 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
      $k(i, j) \leftarrow (T_{x_i} - \mu_x)(T_{z_j} - \mu_z)^t$ {Product of distances of each point to their mean}
  end for
end for
```

3.2. Kernels based on the Jensen–Shannon metric

For the maximum likelihood estimation on a regular exponential family $P_M = (P_\eta, \eta \in M)$, $M \subseteq N$, we need to maximise $l(\eta|\bar{T})$ over the set M. Let A and g be the semi-algebraic set and the diffeomorphism 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-parametrisation by g [17]. Then, the maximisation problem can be relaxed to

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

where $l(\gamma|\bar{T}) = g^{-1}(\gamma)^t \bar{T} - \phi(g(\gamma)^{-1})$. In our case, we utilise the probability simplex as a semi-algebraic set [17] for discrete random variables, which is a convex polyhedron in any dimension. Therefore, the optimisation 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 shown below.

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

Definition 5 *([9] Bregman divergence).*

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

According to the Legendre dual, we have

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

Additionally, F and G are Legendre functions if their derivatives are inverse functions of each other (i.e., $\nabla F(\nabla\,\phi(g^{-1}(\gamma)) = g^{-1}(\gamma))$. Because $F(\bar{T})$ does not depend on parametrisation, we encounter the following optimisation problem:

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

In this respect, we can apply the idea that given new data x_k, a new distribution parametrised by η_i should be chosen. This distribution should be as difficult to discriminate from the original parametrisation η as possible, so that the new data produces as small an information gain in $KL(\eta_i||\eta)$[1] or $B_F(\bar{T}||\nabla\phi(g^{-1}(\gamma_i))$ as possible. In other words, we hope to achieve minimal cross-entropy. Kullback and Leibler previously exploited this approach [19] and termed it the *principle of minimum discrimination information* (MDI).

Therefore, it is natural to use the Jensen–Shannon divergence[2] as a metric to build kernels that exploit the generative properties of the data. In contrast to [9], the main contribution here is bridging together the use of semi-algebraic sets (which are needed for the parametrisation) and the dual structure induced by the diffeomorphism g, which re-parametrises the optimisation problem.

Now, we only have to apply the Jensen–Shannon metric over the dual space. In particular,

Definition 6 *([9,20,21]).* 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 ([9,20,21] Centred kernel). *Let $x_0 \in X$ define the centred kernel as $\psi : X \times X \to \mathbb{R}$*

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

Proposition 2 ([9,20,21] Exponentiated kernel). *We define the exponential kernel as $\psi : X \times X \to \mathbb{R}$*

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

$\forall t > 0$.

Proposition 3 ([9,20,21] Inverse kernel). *We define the inverse kernel as $\psi : X \times X \to \mathbb{R}$*

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

$\forall t > 0$.

It is obvious that the most important aspect involved in the calculation of the kernels outlined above is that 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 $F(\gamma_x)$
  $N = \sum \gamma_x$
  $F \leftarrow \gamma_x \log\left(\frac{\gamma_x}{N}\right)$
```

[1] KL is a Bregman divergence.

[2] Note that the KL divergence is not a metric.

3.3. Quotient Basis Kernel

In this section, we present the definition of algebraic models, as described in [22], where inputs are denoted by x, responses or outputs are denoted by y, and parametric functions are denoted by η, or functions of η. These are related by polynomial algebraic relations, which are possibly implicit. Another feature of this definition is that polynomial constraints can be included in the model specification. Implicit models and the introduction of constraints can lead to the use of dummy variables.

The statistical parameters of the model, are functions of any form, with the restriction that they belong to a specified field. For example, $\mathbb{Q}(\eta_1, \ldots, \eta_p)$ is the set of all rational functions in $\eta_1, \ldots, \eta_p$ with rational coefficients. Similarly, $\mathbb{Q}(e_1^{\eta}, \ldots, e_p^{\eta})$ is the set of all exponential rational functions. Parameters are treated as unknown quantities, and in most cases, they 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 $x_1, \ldots, x_s$ including 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 $\mathbb{Z}_+^s$ [22] (set of positive integers).

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

Definition 7 *([22] 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 $s \in \mathbb{K}$ and 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; 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 [22] demonstrates that every ideal has a finite basis. This provides a very powerful result because it means that any ideal is finitely generated, even if the generating set is not necessarily unique. Another powerful aspect of this result is that this generation basis is a Gröbner basis, which is defined below. This basis will become essential for the derivation of regression/interpolation polynomials and the algebraic derivation of the Fisher and the proposed QBK kernels.

Definition 8 *([22]).* Let τ be a term ordering on $\mathbb{K}[x]$ and f a polynomial in $\mathbb{K}[x]$. The leading term of f, $\mathrm{LT}_\tau(f)$ is the largest term with respect to τ among the terms in f.

Definition 9 *([22] 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

$$\langle \mathrm{LT}_\tau(g_1), \ldots, \mathrm{LT}_\tau(g_t) \rangle = \langle \mathrm{LT}_\tau(I) \rangle \tag{17}$$

where $\mathrm{LT}_\tau(I) = \{\mathrm{LT}_\tau(f) : f \in I\}$.

Theorem 1 ([5,22]). *Given a term ordering, every ideal I except* $\{0\}$ *has a Gröbner basis and any Gröbner basis is a basis of I.*

Definition 10 *([22] 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 *([22,5] Gröbner basis of unique points).* Let A be a set of n unique points $A = \{\mathrm{a}_1, \ldots, \mathrm{a}_n\}$ and τ a term ordering. A Gröbner basis of A, $G = g_1, \ldots, g_t$, is a Gröbner basis of $I(A)$. Therefore, the points in A can be presented as the set of solutions of

$$\begin{cases} g_1(\mathbf{a}) = 0 \\ g_2(\mathbf{a}) = 0 \\ \ldots \\ g_t(\mathbf{a}) = 0 \end{cases} \tag{18}$$

Let us formally define the quotient basis EST_τ that shall be used in the algorithm below.

Definition 12 *([22] Quotient basis).* Let A, be a set of $n \times s$ unique support points $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ and τ a term ordering. A monomial basis of the set of polynomial functions over A is

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

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 ([22]). *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 $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$. We call design matrix (i.e., EST_τ evaluated in A) the following $n \times c$ matrix

$$Z = [\mathrm{EST}_\tau]|_A \tag{20}$$

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

Proposition 4. *The covariance of Z,*

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

is a kernel.

Definition 14 *(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 was achieved with ApCoCoA [23].
4. Calculate the reduced Gröbner Basis G (this can also be calculated with the function IdealOfPoints [24] 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 $Z = [\mathrm{EST}_\tau]|_A$ with the terms of EST_τ.

This algorithm was originally developed for the derivation of interpolation/regression polynomials in [5].

Table 3
Results summary for all kernels with support vector machines, logistic regression and basal simplified acute physiology score.

Kernel	Correct rate (%)	Sens. (%)	Spec. (%)	AUC (%)
Quotient	80.18	79.34	83.24	82.23
Fisher	73.94	71.78	82.72	79.41
Centred	73.01	70.48	82.26	67.17
Exponential	72.53	69.94	81.97	66.81
Inverse	72.08	69.41	81.88	67.97
Linear	72.04	69.11	82.19	79.52
Poly (order 2)	72.26	69.40	82.17	79.21
Gaussian	72.33	69.57	82.12	80.31
LR	71.55	68.57	80.32	79.77
Basal SAPS	71.32	68.42	80.41	68.23

Algorithm 4. Pseudocode of the QBK.

```
Input: A Input dataset, $x$, $y$ and $EST_\tau$
Output: QBK $k(x, y)$
    $\mu \leftarrow$ mean(A)
    $Z_x \leftarrow [\mathrm{EST}_\tau]|_x$ {Evaluate $EST_\tau$ in $x$}
    $Z_y \leftarrow [\mathrm{EST}_\tau]|_y$ {Evaluate $EST_\tau$ in $y$}
    $k(x, y) \leftarrow (Z_x - \mu)(Z_y - \mu)^t$
```

3.4. 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) was compared in their ability to predict mortality using soft-margin SVM [18]. This technique is well suited for the structure of the problem, as the relationships of SAPS and SOFA with mortality are not linear. Note that some patients with high values of SAPS and SOFA 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, one for each class studied. Soft-Margin SVMs [18] 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:

$$\underset{\alpha}{argmax}\left(\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\alpha^t\mathbf{H}\alpha\right)$$
$$\text{s.t.} \quad 0 < \alpha_i < C \quad \text{and} \quad \sum_{i=1}^{N}\alpha_i y_i = 0. \tag{21}$$

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. Parameter C controls the trade-off between the penalty and the size of the margin. Therefore, parameter C can also be interpreted as a factor controlling the number of support vectors.

4. 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 classifier. The inputs to the classifiers are the basal SOFA and SAPS scores at ICU admission. The classifiers were evaluated over a stratified test dataset that was obtained by taking 20% out-of-sample data before cross-validation, and this dataset was used to evaluate all models (400 patients, with a statistical power 97.40%). Table 3 displays the results using the test dataset. The methods presented here have also been compared against a logistic regression (LR) over the basal SAPS and SOFA scores at ICU admission. In Table 4, we summarise the coefficients, confidence intervals and p-values for this LR. QBK were calculated using Algorithm 4, and the kernels based on the Jensen–Shannon metric were calculated with Definitions 1–3 from Algorithm 3.2. The simplified Fisher kernel was calculated using algorithm in Section 3.1. The standard linear, Gaussian and polynomial kernels were also tested. We used Matlab ® SVM QP solver in the BioInformatics and Optimisation Toolboxes. The cross validation experiment (during training and validation) yielded the following appropriate C parameter values for each kernel:

Table 4
Logistic regression over basal simplified acute physiology score and sequential organ failure assessment scores at intensive care unit admission.

	Coeff.	CI	p-Value
Intercept	8.09	[6.30, 9.88]	0
SOFA	−0.24	[−0.33, −0.16]	0.15×10^{-6}
SAPS	−0.20	[−0.28, −0.11]	2.26×10^{-6}

- for the QBK, $C = 1$.
- for the Fisher kernel, $C = 1$.
- for the kernels based on Jensen–Shannon metric, $C = 10$. 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 3 demonstrates that the QBK consistently yielded the best results out of all of the metrics considered: accuracy, sensitivity, specificity and AUC (which, in a way, summarises sensitivity and specificity). The Fisher kernel yielded better accuracy than the centred, exponential, inverse, Gaussian, linear and polynomial kernels. In our study, we did not observe significant differences in accuracy for the centred, linear, Gaussian and polynomial kernels. The results for these 5 kernels were not very different from that obtained using an LR of the SOFA and SAPS scores collected at ICU admission. The other three kernels based on the Jensen–Shannon metric (centred, inverse and exponential) yielded the worst AUC results. Finally, statistical relevance was tested using the McNemar's test between all pairs of results [25]. Table 5 illustrates the χ^2 values between all pairs of results. These χ^2 values indicate that most predictions are significantly different at $p \ll 0.05$, with the exception of the predictions obtained with the polynomial and Gaussian kernels, which are statistically equivalent.

5. Discussion

Over the past 30 years, scoring systems for intensive care unit patients have been introduced and developed. They allow for the assessment of disease severity and provide an estimate of in-hospital mortality. Physiology-based scoring systems are applied to critically ill patients, and they have a number of advantages over diagnosis-based systems. Additionally, severity scoring systems are often used to stratify critically ill patients for possible inclusion in clinical trials. In this study, we enrolled a significant number of patients, which allowed us to evaluate the SAPS and SOFA at ICU admission as a combined prognostic factor for sepsis. Our study followed the methodology presented in [8], where a threshold was selected to assess whether the patient survives. The work from Le Gall et al. [8] used a threshold of 12, which yields a sensitivity and specificity of 0.69. In our case, we automatically selected the threshold that resulted in the maximum accuracy in assessing the risk of death.

However, it is important to note that our population is not as general as that presented in [8], as it only comprises patients with sepsis. In our case, the threshold was 19.5 for the basal SAPS (i.e.,

Table 5
χ^2 values for McNemar test between pairs of predictions.

	Fish.	Cent.	Exp.	Inv.	Lin.	*Poly*.	Gauss.	LR
Quot.	25.20	96.00	71.40	92.29	128	98.61	98.61	216.00
Fish.		42.30	26.25	29.32	42.30	41.90	98.61	216.00
Cent.			21.01	4.05	200.40	183.08	183.10	481.00
Exp.				10.08	166.91	151.83	151.83	481.00
Inv.					196.12	178.94	178.94	476.00
Lin.						10.02	10.02	87.00
Poly.							n/s	97.00
Gauss.								97.00

SAPS at ICU admittance). This threshold yielded an accuracy of 71.32%, a sensitivity of 68.42% and a specificity of 80.41%. The accuracy, sensitivity and specificity in this study are very similar to those obtained with our Jensen–Shannon kernels. The poor sensitivity may be because SAPSs include non-sepsis specific clinical traits (i.e., the performance of haemo-cultures, antibiotic administration or vasoactive drug administration [26]). Therefore, a combined approach between SOFA and SAPS was ideal for our population and the overall goal of assessing risk of death.

We also compared our results to LR, which is a method widely used in clinical practice that uses basal SAPS and SOFA at ICU admission. LR yielded an accuracy of 71.55%, a sensitivity of 68.57% and a specificity of 80.32%. Regarding LR, overall accuracy was improved compared to [8] and the results presented in this paper for the SAPS. Thus, a score combining SAPS and SOFA is clinically relevant. The QBK proposed in this paper is a generalisation of LR, and it exploits the data structure to provide a composite scoring system that improves general performance, particularly sensitivity and specificity. The method described in this paper to assess the risk of death can be implemented as part of a decision support system in the ICU in combination with the SOFA and SAPS. One of the main benefits of such a system is that it uses the basal values of SOFA and SAPS at ICU admission, which is the most critical moment for a patient with sepsis. Indeed, taking proper action during the first six hours of evolution is critical for survival [27].

One limitation of this study, compared with other studies found in the literature [28], is that the only data available are the static values of the SOFA and SAPS. However, the methods presented in this paper can be applied to the physiological values used to calculate these scores. Another limitation of this study is that the performance of the proposed method was evaluated in a single ICU and a limited population sample. Future work should include a multi-centric prospective study to validate its generalisation.

Regarding computation time, it is very important to note that even though the QBK yields the best results in terms of accuracy and balance between sensitivity and specificity for high-dimensional datasets or large input/design matrices, the calculation of a Gröbner basis is very time-consuming. Therefore, for large datasets, we propose the use of the simplified Fisher kernel. To this end, computation time improves because we only need to calculate the covariance of the sufficient statistics of a multinomial distribution (i.e., a vector of zeroes with a 1 in the position corresponding to the actual value of SOFA and SAPS). This increases the sparsity and accelerates the SVM optimisation, making it even faster than that of a linear kernel and yielding acceptable results (Table 3).

6. Conclusions

Sensitivity and specificity are important measures of performance when predicting mortality in patients with sepsis because more aggressive treatment and therapeutic interventions may result in better outcomes for high-risk patients. Here, the SVMs were trained with eight different kernels applied to the basal SOFA and SAPS at ICU admission, of which five were generative and the other three were considered well-suited kernels for the problem at hand. Overall, the investigated kernels provide accurate and medically actionable results, while keeping an acceptable balance between the different parameters of interest (accuracy rate, sensitivity, specificity and AUC). The new kernel proposed in the study, QBK, is defined through the Gröbner basis of an algebraic ideal and the sufficient statistics of a regular exponential family. The proposed simplified Fisher kernel was derived through 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 novel result from the application of regular exponential families for the definition of kernels.

The QBK and simplified Fisher kernels outperformed not only the alternative kernels but also the clinical standard method based on the SAPS for predicting mortality prediction in patients with sepsis. In particular, in this study we report accuracies of 80.18% and 73.94% for the QBK and simplified Fisher kernel, respectively.

The sepsis mortality model using QBK has greater sensitivity and specificity than SOFA and SAPS. This model may be a useful alternative method of severity adjustment for benchmarking purposes, conducting studies of patients with sepsis or providing a decision support method that could be used in addition to SAPS and SOFA for the assessment of outcome in patients with sepsis. In the future, our goal is to generalise the algebraic methods presented here with the application of graphical models over the input variables used to calculate the SOFA and SAPS according to the findings described in the previous section. These graphical models would be a natural expansion of the algebraic models presented in this paper through the use of the Hammersley–Clifford theorem (factorisation of regular exponential families).

References

[1] Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS international sepsis definitions conference. Intensive Care Med 2003;29:530–8.

[2] Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinski MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 2001;29(7):1303–10.

[3] Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. N Engl J Med 2003;348: 1546–54.

[4] Esteban A, Frutos-Vivar F, Ferguson ND, Peñuelas O, Lorente JA, Gordo F, et al. Sepsis incidence and outcome: contrasting the intensive care unit with the hospital ward. Crit Care Med 2007;35(5):1284–9.

[5] Giglio B, Riccomagno E, Wynn HP. Gröbner basis strategies in regression. J Appl Stat 2000;27(7):923–38.

[6] Pachter L, Sturmfels B, editors. Algebraic statistics for computational biology. Cambridge, UK: Cambridge University Press; 2005.

[7] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol 2002;29:641–4.

[8] Le Gall JR, Loirat P, Alperovitch A, Glaser P, Granthil C, Mathieu D, et al. A simplified acute physiology score for ICU patients. Crit Care Med 1984;12(11):975–7.

[9] Agarwal A, Daumé III H. Generative kernels for exponential families. J Mach Learn Res 2011;15:85–92.

[10] Vincent JL, Moreno R, Takala J, Willats S, De Mendonça A, Bruining H, et al. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Crit Care Med 1996;22:707–10.

[11] Ribas VJ, Vellido A, Ruiz-Rodríguez JC, Rello J. Severe sepsis mortality prediction with logistic regression over latent factors. Expert Syst Appl 2012;39: 1937–43.

[12] Levy MM, Macias WL, Vincent JL, Russell JA, Silva E, Trzaskoma B, et al. Early changes in organ function predict eventual survival in severe sepsis. Crit Care Med 2005;31:2194–202.

[13] Kajdacsy-Balla AC, Andrade FM, Moreno R, Artigas A, Cantraine F, Vincent JL. Use of the sequential organ failure assessment score as a severity score. Intensive Care Med 2005;31(2):243–9.

[14] Le Gall JR, Neumann A, Hemery F, Bleriot JP, Fulgencio JP, Garrigues B, et al. Mortality prediction using SAPS II: an update for French intensive care units. Crit Care 2005;9(6):R645–52.

[15] Metnitz PGH, Moreno RP, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3 – from evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description. Intensive Care Med 2005;31:1336–44.

[16] Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3 – 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 Med 2005;31:1345–55.

[17] Drton M, Sullivant S. Algebraic statistical models. Stat Sin 2007;17:1273–97.

[18] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge, UK: Cambridge University Press; 2000.

[19] Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat 1951;22:79–86.

[20] Berg C, Christensen JPR, Ressel P. Harmonic analysis on semigroups. New York: Springer-Verlag; 1984.

[21] Schoenberg IJ. Metric spaces and positive definite functions. Trans Am Math Soc 1938;44:522–36.

[22] Pistone G, Riccomagno E, Wynn HP. Algebraic statistics: computational commutative algebra in statistics. Monographs on statistics and applied probability, vol. 89. Boca Raton, FL: Chapman and Hall/CRC Press; 2001.

[23] CoCoATeam. CoCoA: a system for doing computations in commutative algebra; 2012 http://cocoa.dima.unige.it

[24] Abbott J, Bigatti A, Kreuzer M, Robbiano L. Computing ideals of points. J Symb Comput 2000;30:341–56.

[25] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 1947;12(2):153–7.

[26] Ribas VJ, Ruiz-Rodríguez JC, Wojdel A, Caballero-López J, Ruiz-Sanmartín A, Rello J, et al. Severe sepsis mortality prediction with relevance vector machines. In: Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE. Boston: IEEE; 2011. p. 100–3.

[27] Dellinger RP, Carlet JM, Masur H, Gerlach H, Calandra T, Cohen J, et al. Surviving sepsis campaign guidelines for management of severe sepsis and septic shock. Intensive Care Med 2004;30:536–55.

[28] Fialho AS, Cismondi F, Vieira SM, Sousa JMC, Reti SR, Howell MD, et al. Predicting outcomes of septic shock patients using feature selection based on soft computing techniques. In: Hüllermeier E, Kruse R, Hoffmann F, editors. Information processing and management of uncertainty in knowledge-based systems. Communications in computer and information science, vol. 81. Berlin, Heidelberg: Springer; 2010. p. 65–74.