MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Information Criteria for Kernel Machines

Kei KOBAYASHI and
Fumiyasu KOMAKI

METR 2005–23 Aug 2005

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Information Criteria for Kernel Machines

Kei Kobayashi* and Fumiyasu Komaki†

Aug, 2005

Abstract

We present kernel regularization information criterion (KRIC), which is a new criterion for tuning regularization parameters in kernel logistic regression (KLR) and support vector machines (SVMs). The main idea of the KRIC is based on the regularization information criterion (RIC). Although the RIC is a useful criterion for tuning regularization parameters in statistical regularization models, it cannot be directly applied to parameter tuning for the kernel machines because kernel functions define only inner products in feature spaces. We derive an eigenvalue equation to calculate the KRIC and solve the problem.

The computational cost for parameter tuning by the KRIC is reduced drastically by using the Nyström approximation. The test error rate of SVMs or KLR with the regularization parameter tuned by the KRIC is comparable with the one by the cross validation or evaluation of the evidence. The computational cost of the KRIC is significantly lower than the one of the other criteria.

1 Introduction

In recent years, there have been plenty of researches on support vector machines (SVMs) and kernel logistic regression (KLR). An overview can be found in Vapnik (1998), Cristianini & Shawe-Taylor (2000), and Schölkopf & Smola (2002) for the SVMs and in Jaakkola & Haussler (1999), Keerthi et al. (2002), and Zhu & Hastie (2002) for the KLR.

In particular, parameter and kernel selection becomes a very important theme because they have serious effect on the performance of classifications by the KLR and the SVMs. Both the KLR and the SVMs have two kinds of parameters to tune, the misclassification penalty (often denoted by C), and parameters specifying the kernel function. In this chapter, we mainly deal with the optimization for misclassification penalty C. The parameter C in the SVMs and the KLR is recognized as the regularization parameter in the corresponding statistical regularization models. The value of parameter C determines the tradeoff between training error and model complexity.

*kkoba@stat.t.u-tokyo.ac.jp, The Institute of Statistical Mathematics.

†Graduate School of Information Science and Technology, University of Tokyo.

There have been a lot of researches for tuning the parameter C. One of the most widely used approaches is based on data resampling methods such as cross validation and bootstrapping. They require high computational cost. Another approach is to optimize generalization error bounds (for example, evaluating structural risk minimization (Vapnik (1998)) and the span method (Vapnik & Chapelle (2000))). However, most of these works consider the worst case with respect to the probability measure of population. Therefore the upper bounds on the test error rate are sometimes very loose. The performance of Wahba's GACV (Wahba (1998)) and $\xi\alpha$-estimation by Joachims (2000) is between the two approaches. The experimental comparison between these methods has been studied in Duan et al. (2003).

Kwok (1999) constructed a Bayesian model approximating SVMs and tuned the regularization and kernel parameters by maximizing the type II likelihood (Good (1965)). The type II likelihood is sometimes called as the evidence. He also proposed an approximation method for calculating the type II likelihood (Kwok (2000)). This study based on the Bayesian inference framework (MacKay (1992a) and MacKay (1992b)) was innovative.

In the model in Kwok (1999), the conditional probabilities for each class y_i given data x_i do not add to one. Therefore, his model is not an exact statistical model. A naive but effective solution for this problem is to approximate a hinge loss function of SVMs by a logistic function. We call such models as the logistic Bayesian models for SVMs. Sollich (2002) proposes another Bayesian model for SVMs. He introduces the idea of the normalization of likelihood and constructs an exact Bayesian model for SVMs.

In most of these Bayesian framework approaches, the evidence value is used for tuning the regularization parameter. The calculation of the evidence is a difficult problem and there are several approaches to solve it. Opper & Winther (2000) derived some approximations and bounds of that by the cavity method borrowed from statistical mechanics. Seeger (2000) used a Gaussian variational approach to estimate the evidence. Kwok (2000) used the Laplace approximation.

Another statistical model for the SVMs is the regularization model. Many people have noted the relationship between the SVMs and the regularization function estimation in the reproducing kernel Hilbert spaces (RKHS). An overview can be found in Hastie et al. (2001), Wahba (1998), and Schölkopf & Smola (2002).

In this chapter, first, two different regularization models are considered as the statistical models for SVMs. These models correspond to the Bayesian models, the logistic Bayesian model and Sollich's model. In the regularization models, the regularization parameter corresponds to the hyperparameter in the Bayesian models.

Next, we introduce a criterion, kernel regularization information criterion (KRIC), for tuning the regularization parameter. The KRIC corresponds to the regularization information criterion (RIC) in the feature space.

It is known that the RIC is an effective criterion for tuning parameters in regularization models. The main idea of the RIC is to minimize the expectation of the Kullback-Leibler divergence from the true distribution function to the estimated distribution function by the regression model. See Shibata (1989) for details of the RIC.

The RIC is often easier to compute than the type II likelihood. It is because the RIC

uses a plug-in estimator while the calculation of the type II likelihood needs the integration with respect to the model parameters. The computational cost of model selections by the RIC is often lower than other methods because it needs no data resampling or recursive optimization.

However, for SVMs, the RIC cannot be directly calculated because kernel functions define only the inner products of vectors in feature spaces. We reduces the calculation of the RIC for SVMs to an eigenvalue problem and introduces the KRIC.

The KRIC can be used for tuning the regularization parameters in the kernel logistic regression (KLR). The KLR corresponds to the penalized logistic classification in the RKHS (See the details in Jaakkola & Haussler (1999), Keerthi et al. (2002), and Zhu & Hastie (2002)). KLR is easier to analyse than SVMs because KLR has the logistic loss function and SVMs have a hinge loss function (see Wahba et al. (1995) and Hastie & Tibshirani (1990)). The computational cost for KLR is much higher than SVMs with the sequential minimal optimization (SMO) introduced by Platt (1998) or other decomposition algorithms. This is one of the reasons why the KLR is less popular than the SVMs as an application tool. In recent years, however, the SMO algorithm for the KLR is introduced (Keerthi et al. (2002) and Zhu & Hastie (2002)).

We propose to tune the regularization parameter for the KLR by using the KRIC. We can use the result on SVMs directly because the Bayesian model for the KLR is similar to the logistic Bayesian model for SVMs.

2 The support vector machines and the kernel logistic regression

In this section, we summarize the general algorithm for SVMs and KLR in binary classification problems. Let the input space be $\mathcal{X} \subset \mathbb{R}^m$. Let D be the training set $\{(\boldsymbol{x}_i, y_i)\}$ $(i = 1, \ldots, l)$, where $\boldsymbol{x}_i \in \mathcal{X}$ is the input and $y_i \in \{\pm 1\}$ is the output label. In the binary classification (or pattern recognition) problem, we predict $\tilde{y}_i$ for each future input data $\tilde{\boldsymbol{x}}_i$. The statistical model for this problem is introduced after section 3. We first explain the algorithm for SVMs and KLR.

2.1 Support vector machine

SVMs map the inputs $\boldsymbol{x} \in \mathbb{R}^m$ to vectors $\boldsymbol{z} = \boldsymbol{\phi}(\boldsymbol{x})$ in a l^2 space, called feature space $\mathcal{F}$. The inner product of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ in the l^2 is represented by $\boldsymbol{z}_1 \cdot \boldsymbol{z}_2$. SVMs construct the optimal hyperplane $\boldsymbol{w} \cdot \boldsymbol{z} - b = 0$ in the feature space and classify data by a classification function $f(\boldsymbol{w}; \boldsymbol{z}, b) := \mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{z} - b)$. Here, $\boldsymbol{w}$ and b are obtained by solving the following quadratic problem:

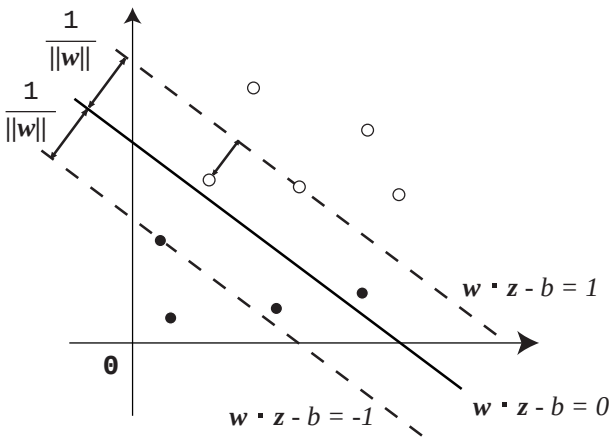

Figure 1: An example of hyperplane by an SVM.

The main quadratic problem of SVMs in the feature space

$$\text{min.} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}\xi_i, \tag{1}$$

$$\text{s.t.} \quad y_i(\boldsymbol{w}\cdot\boldsymbol{z}_i - b) \geq 1-\xi_i, \text{ and } \xi_i \geq 0 \quad \text{for } i = 1,\ldots,l. \tag{2}$$

Here, C is a positive parameter, determining the regularization property of SVMs. Inequalities (2) are represented by

$$\xi_i = [1 - y_i(\boldsymbol{w}\cdot\boldsymbol{z}_i - b)]_+, \quad \text{where } [\cdot]_+ := \max(0,\cdot). \tag{3}$$

We give intuitive explanation of quadratic problem (1) and (2). We consider the problem separating data by hyperplane with a margin $T = \{\boldsymbol{z} \in l^2; |\boldsymbol{w}\cdot\boldsymbol{z} - b| \leq 1\}$. (See figure 1.) The width of hyper plane T is $\|\boldsymbol{w}\|^{-1}$. Then minimization of $\|\boldsymbol{w}\|^2/2$ in (1) corresponds to maximization of the width of hyper plane T. On the other hand, for a data of class $y_i = 1$, ξ_i in (3) is the distance from the "correct" half-space $\boldsymbol{w}_i\cdot\boldsymbol{z}_i - b \geq 1$ to the data $\boldsymbol{z}_i$. Since ξ_i is a misclassification level of each data by hyper plane T, the second term of (1) is proportional to the sum of the level of misclassification.

Therefore, the SVMs search the optimal hyperplane with a margin T such that the width of T is large and the misclassification of data is small at the same time.

The solution of the main quadratic problem (1) and (2) becomes

$$\hat{\boldsymbol{w}} = \sum_{i=1}^{l}\alpha_i y_i \boldsymbol{z}_i,$$

$$\hat{b} = -\frac{\max_{y_i=-1}\hat{\boldsymbol{w}}\cdot\boldsymbol{z}_i + \max_{y_i=1}\hat{\boldsymbol{w}}\cdot\boldsymbol{z}_i}{2},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^\top$ are the solutions of the following dual quadratic problem:

The dual problem of SVMs in the feature space

$$\begin{aligned} \max. \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{z}_i \cdot \boldsymbol{z}_j, \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0 \ \text{ for } i = 1, \ldots, l, \\ & \sum_{i=1}^{l} \alpha_i y_i = 0. \end{aligned}$$

See, for example, Cristianini & Shawe-Taylor (2000) and Vapnik (1995) for derivation of the dual problem of SVMs.

2.2 Kernel functions

Instead of $\boldsymbol{z}_i \cdot \boldsymbol{z}_j$ in the dual problem, we use a function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) := \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{\phi}(\boldsymbol{x}_j)$ for each $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$. The function $K(\cdot, \cdot)$ is called a kernel function. We need some conditions on K for being an inner product in a feature space.

Theorem 2.1 (Mercer's theorem (explained in Cristianini & Shawe-Taylor (2000))) *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^n$. Suppose K is a continuous symmetric function such that the integral operator $T_K : L_2(\mathcal{X}) \to L_2(\mathcal{X})$,*

$$(T_K f)(\cdot) := \int_{\mathcal{X}} K(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

is positive, that is

$$\int_{\mathcal{X} \times \mathcal{X}} K(\boldsymbol{x}_i, \boldsymbol{x}_j) f(\boldsymbol{x}_i) f(\boldsymbol{x}_j) d\boldsymbol{x}_i d\boldsymbol{x}_j \geq 0,$$

for all $f \in L^2(\mathcal{X})$. Then we can expand $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in a uniformly convergent series (on $\mathcal{X} \times \mathcal{X}$) in terms of T_K's eigen-functions $\phi_k \in L_2(\mathcal{X})$, normalized in such a way that $\|\phi_k\|_{L_2} = 1$, and positive associated eigenvalues $\lambda_k \geq 0$,

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{k=1}^{\infty} \lambda_k \phi_k(\boldsymbol{x}_i) \phi_k(\boldsymbol{x}_j).$$

We use the term kernel to refer to functions satisfying this property, but in the literature these are often called *Mercer kernels*.

We introduce some examples of kernel functions.

1. Example: Gaussian kernel (Radial Basis Function kernel)

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$$

where $\sigma > 0$.

2. Example: the polynomial kernel

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + c\boldsymbol{x}_i \cdot \boldsymbol{x}_j)^d$$

where $d \in \mathbb{N}$ and $c > 0$.

3. Example: the neural network kernel (sigmoid kernel)

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(b(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) - c)$$

where $b > 0$ and $c \in \mathbb{R}$.

The parameters in kernels, as σ, c, d and b above, are called *kernel parameters*.

Gaussian kernel and the polynomial kernel are Mercer kernels. The neural network kernel is *not* a Mercer kernel for any b and c. (See Smola et al. (2000).)

Using the kernel functions, the dual problem of SVMs becomes as follows:

The dual problem of the SVMs with kernel function in the feature space

$$\begin{aligned} \max. \quad & \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0 \ \text{ for } i = 1, \ldots, l, \\ & \sum_{i=1}^{l} \alpha_i y_i = 0. \end{aligned}$$

Let

$$a(\boldsymbol{x}, \boldsymbol{w}, b) := \boldsymbol{w} \cdot \boldsymbol{z} - b = \sum_i \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) - b, \tag{4}$$

and

$$a_i := a(\boldsymbol{x}_i, \boldsymbol{w}, b).$$

Then the optimal hyperplane is $a(\boldsymbol{x}, \hat{\boldsymbol{w}}, \hat{b}) = 0$ and the classification function is $I(\boldsymbol{x}) = \text{sign}(a(\boldsymbol{x}, \hat{\boldsymbol{w}}, \hat{b}))$.

2.3 The kernel logistic regression (KLR)

The KLR solves the following optimization problem.

The optimization problem of the KLR

$$\begin{aligned} \min. \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{l} \xi_i, \\ \text{s.t.} \quad & \xi_i = g(y_i(\boldsymbol{w} \cdot \boldsymbol{z}_i - b)) \ \text{ for } \ i = 1, \ldots, l, \end{aligned}$$

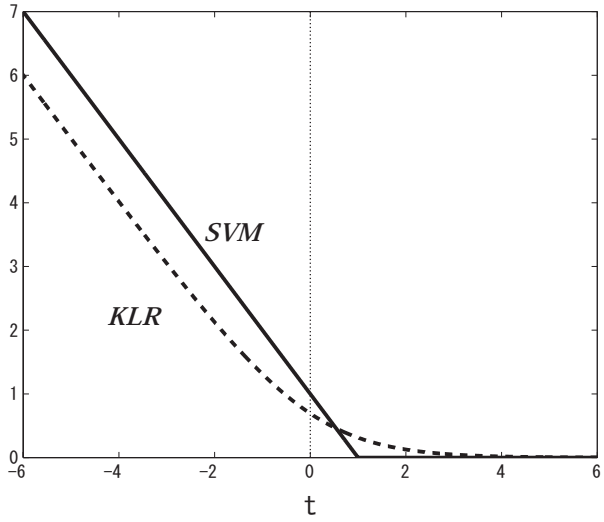

Figure 2: Comparison of loss functions of the SVMs and the KLR.

where $g(t) := \log(1 + e^{-t})$.

The KLR uses the loss function $g(t)$ instead of the SVMs' loss function $[1-t]_+$. Figure 2 shows the difference of these two loss functions.

Let $G(\delta) := \delta \log \delta + (1 - \delta) \log(1 - \delta)$. The Wolfe dual problem for KLR is

The dual problem of the KLR

$$\begin{aligned} \max. \quad & -C \sum_{i=1}^{l} G(\frac{\alpha_i}{C}) - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \\ \text{s.t.} \quad & \sum_{i=1}^{l} \alpha_i y_i = 0. \end{aligned}$$

See Keerthi et al. (2002) for derivation of the dual problem.

3 The logistic Bayesian framework for SVMs and KLR

We assume that each sample $(\boldsymbol{x}_i, y_i)$ is independently and identically distributed. We also assume that $p(\boldsymbol{x}_i)$ is independent of $\boldsymbol{w}$ and $\boldsymbol{\pi}$.

We consider a logistic predictive density

$$p(y_i \mid \boldsymbol{x}_i, \boldsymbol{w}, b, \boldsymbol{\pi}, K) = \frac{1}{1 + \exp(-\eta a_i y_i)}, \tag{5}$$

where η is a fixed positive constant. In the left-hand side, K represents the kernel function and $\boldsymbol{\pi}$ represents the kernel parameters. We omit K in the following.

Next, we assume a normal prior density:

$$p(\boldsymbol{w}, b \mid \lambda, \boldsymbol{\pi}) \propto \exp(-\frac{\lambda}{2}\|\boldsymbol{w}\|^2), \quad (\lambda > 0)$$

where λ is a positive hyper-parameter. From Bayes' rule, the posterior distribution is

$$\begin{aligned} p(\boldsymbol{w}, b \mid D, \lambda, \boldsymbol{\pi}) &\propto p(\boldsymbol{w}, b \mid \lambda, \boldsymbol{\pi}) p(D \mid \boldsymbol{w}, b, \boldsymbol{\pi}) \\ &= p(\boldsymbol{w}, b \mid \lambda, \boldsymbol{\pi}) \prod_i p(y_i \mid \boldsymbol{x}_i, \boldsymbol{w}, b, \boldsymbol{\pi}) p(\boldsymbol{x}_i). \end{aligned}$$

Therefore, the negative logarithmic posterior distribution is

$$\begin{aligned} -\log p(\boldsymbol{w} \mid D, \lambda, \boldsymbol{\pi}) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_i \log\{1 + \exp(-\eta a_i y_i)\} \\ -\sum_i \log p(\boldsymbol{x}_i) + \text{constant (not depend on } \boldsymbol{w}). \end{aligned}$$

4 Sollich's Bayesian framework for SVM

We summarize the Bayesian model by Sollich. We abbreviate $a(\cdot, w, b)$ in (4) to a. As in the Gaussian process (Williams & Barber (1998)), we consider a prior distribution of the function a. Let $Q(\boldsymbol{x})$ be a true density function of each sample $\boldsymbol{x}_i$. Then the likelihood functions of the Bayesian model are given by

$$\begin{aligned} p(\boldsymbol{x}|a) &= Q(\boldsymbol{x})\nu(a(\boldsymbol{x}))/N(a), \\ p(y|\boldsymbol{x}, a) &= \exp(-C[1 + ya(\boldsymbol{x})]_+)\tau(C)/\nu(a(\boldsymbol{x})) \end{aligned}$$

where

$$\begin{aligned} N(a) &:= \int_{\boldsymbol{x}} Q(\boldsymbol{x})\nu(a(\boldsymbol{x}))d\boldsymbol{x}, \\ \nu(a(\boldsymbol{x})) &:= \tau(C)[\exp\{-C[1 - a(\boldsymbol{x})]_+\} + \exp\{-C[1 + a(\boldsymbol{x})]_+\}] \end{aligned} \tag{6}$$

and $\tau(C) = \{1 + \exp(-2C)\}^{-1}$. The integrability of the right-hand side of (6) is assumed. The coefficient $\tau(C)$ is set to ensure that $\nu(a(\boldsymbol{x})) \leq 1$. The prior density function is defined as follows:

$$\pi(a) \propto \exp(-\frac{1}{2}\int a(\boldsymbol{x})K^{-1}(\boldsymbol{x}, \boldsymbol{x}')a(\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}')N^l(a)$$

Assume the Bayesian model above, the logarithm of the posterior probability is

$$\begin{aligned}\log p(a|D) &= \sum_i \log p(y_i|\boldsymbol{x}_i, a) + \log p(\boldsymbol{x}_i|a) + \log \pi(a) + const. \\ &= -\frac{1}{2}\int a(\boldsymbol{x})K^{-1}(\boldsymbol{x},\boldsymbol{x}')a(\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}' - C\sum_{i=1}^{l}[1-y_i\theta(\boldsymbol{x}_i)]_+ + const. \\ &= -\frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}b^2\int K^{-1}(\boldsymbol{x},\boldsymbol{x}')d\boldsymbol{x}d\boldsymbol{x}' - C\sum_{i=1}^{l}[1-y_i a(\boldsymbol{x}_i)]_+ + const.\end{aligned}$$

We adopt the Gaussian process prior on $a(\boldsymbol{x}) + b = \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x})$ instead of $a(\boldsymbol{x})$, i. e.,

$$\pi(a) \propto \exp\{-\frac{1}{2}\int (a(\boldsymbol{x})+b)K^{-1}(\boldsymbol{x},\boldsymbol{x}')(a(\boldsymbol{x}')+b)d\boldsymbol{x}d\boldsymbol{x}'\}N^l(a). \tag{7}$$

Then the posterior probability becomes

$$\log p(a|D) = -\frac{1}{2}\|w\|^2 - C\sum_i [1-y_i a(\boldsymbol{x}_i)]_+ + const.$$

Therefore the maximization of the posterior probability corresponds to the optimization in the SVMs' quadratic problem.

In Sollich's paper, it is suggested to maximize

$$-\frac{1}{2}\|w\|^2 - \frac{1}{2}b^2B^{-2} - C\sum_i [1-y_i a(x_i)]_+ \tag{8}$$

for an adequately selected $B > 0$. This corresponds to assuming the prior distribution on the bias parameter b. While this modification of the original SVMs sometimes approves the performance, we use the original SVMs and the corresponding Bayesian model with the prior (7). We give further discussions on the bias term b in section 9.

5 Regularization information criterion (RIC)

The regularization information criterion (RIC) for a regularization model

$$L(D;\theta;\lambda) = \log p(D|\theta) - \lambda k(\theta)$$

is defined by

$$\mathrm{RIC} := 2[-\log p(D \mid \theta^*) + \mathrm{trace}[\boldsymbol{I}\boldsymbol{J}^{-1}]], \tag{9}$$

where

$$\begin{aligned}\boldsymbol{J} &:= \mathrm{E}_{p_0}\left[-\frac{\partial^2 L(D;\theta;\lambda)}{\partial\theta\partial\theta^\top}\mid_{\theta=\theta^*}\right], \\ \boldsymbol{I} &:= \mathrm{Var}_{p_0}\left[\frac{\partial L(D;\theta;\lambda)}{\partial\theta}\mid_{\theta=\theta^*}\right],\end{aligned}$$

p_0 is the unknown true probability measure on D and $\theta^* = \theta^*(D, \lambda)$ maximizes $L(D; \theta; \lambda)$. For calculation of J and I, instead of expectations with respect to the unknown true density p_0, the sample means are usually used.

The derivation of RIC is similar to that of Takeuchi's information criterion (TIC) (Takeuchi (1978) and Burnham & Anderson (2002)), which is a modification of AIC. It is proved in Shibata (1989) that

$$\begin{aligned} &\int p_0(D') KL(p_0(D)||p(D|\theta^*(D', \lambda)))dD' \\ &\quad = \int p_0(D) \log p_0(D) dD + \frac{1}{2} RIC + o(1) \end{aligned} \tag{10}$$

where $KL(p||q)$ represents the Kullback-Leibler divergence from p to q. The first term of the right-hand side does not depend on λ. Thus, the minimization of the RIC with respect to the parameter λ corresponds to the minimization of the expectation of the Kullback-Leibler divergence from the true probability measure to the estimated probability measure in the regularization model (9) with parameter λ.

6 The Kernel regularization information criterion (KRIC)

In this section, we recognize KLR and SVMs as regularization models and propose a novel criterion for parameter tuning, kernel regularization information criterion (KRIC).

We assume that the reproducing kernel Hilbert space is finite dimensional. Let d be the dimension. It is for using the matrix representation instead of the operational representation on the Hilbert space. This assumption does not hold for general kernel functions (e.g. the Gaussian kernel). However, if the kernel function satisfies Mercer's condition, for any degree of accuracy, there is a matrix representation for each Hilbert space that approximates the operational representation to the degree of accuracy. Therefore, we use the matrix representation for simplicity of the notation.

First, we consider the logistic Bayesian model for SVMs. In the logistic Bayesian framework, we maximize

$$L(D; \boldsymbol{w}, b; \lambda, \boldsymbol{\pi}) = \log p(D \mid \boldsymbol{w}, b, \boldsymbol{\pi}) - \frac{\lambda}{2} \|\boldsymbol{w}\|^2. \tag{11}$$

This model can be recognized as a regularization model (Smola et al. (1998)). Therefore, we can use RIC in order to optimize the regularization parameter λ.

We evaluate the value of RIC. Let $\tilde{\boldsymbol{w}}^\top = [\boldsymbol{w}^\top \ \gamma^{-1} b]$ and $\tilde{\boldsymbol{z}}_i^\top = [\boldsymbol{z}_i^\top \ -\gamma]$ where γ is an arbitral positive constant. Then $\boldsymbol{J}$ and $\boldsymbol{I}$ in the RIC is evaluated as follows:

$$\begin{aligned} \boldsymbol{J} &= \mathrm{E}_{p_0}\Big[\frac{\partial^2}{\partial \tilde{\boldsymbol{w}} \partial \tilde{\boldsymbol{w}}^\top}\{-\sum_i \log p(\boldsymbol{x}_i) p(y_i \mid \boldsymbol{x}_i, \tilde{\boldsymbol{w}}, \boldsymbol{\pi}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\} \mid \tilde{\boldsymbol{w}} = \tilde{\boldsymbol{w}}^*\Big] \\ &= \sum_i t_i \tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_i^\top (1 + O(l^{-1/2})) + \lambda \tilde{\boldsymbol{I}}_d, \end{aligned} \tag{12}$$

where

$$\tilde{\boldsymbol{I}}_d := \begin{bmatrix} \boldsymbol{I}_d & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{bmatrix} \tag{13}$$

and

$$t_i := \frac{\partial^2}{\partial a_i^2} \log p(y_i = 1 \mid \boldsymbol{x}_i, \tilde{\boldsymbol{w}}^*, \boldsymbol{\pi}). \tag{14}$$

In the same way,

$$\begin{aligned} \boldsymbol{I} &= \mathrm{Var}_{p_0}\Big[\frac{\partial}{\partial \tilde{\boldsymbol{w}}} \{ -\sum_i \log p(y_i \mid x_i, \boldsymbol{w}, \boldsymbol{\pi}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \} \mid_{\tilde{\boldsymbol{w}} = \tilde{\boldsymbol{w}}^*} \Big] \\ &= (\sum_i m_i^2 \tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_i^\top - \frac{1}{l} \sum_i m_i \tilde{\boldsymbol{z}}_i \sum_j m_j \tilde{\boldsymbol{z}}_j^\top)(1 + O(l^{-1/2})), \end{aligned} \tag{15}$$

where

$$m_i := \frac{\partial}{\partial a_i} \log p(y_i = 1 \mid \boldsymbol{x}_i, \tilde{\boldsymbol{w}}^*, \boldsymbol{\pi}). \tag{16}$$

In (12) and (15), the law of large numbers is used. Since $\boldsymbol{J}$ and $\boldsymbol{I}$ depend on the unknown real probability measure p_0, we replace $\boldsymbol{J}$ and $\boldsymbol{I}$ by

$$\hat{\boldsymbol{J}} = \sum_i t_i \tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_i^\top + \lambda \tilde{\boldsymbol{I}}_d, \tag{17}$$

$$\hat{\boldsymbol{I}} = \sum_i m_i^2 \tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_i^\top - \frac{1}{l} \sum_i m_i \tilde{\boldsymbol{z}}_i \sum_j m_j \tilde{\boldsymbol{z}}_j^\top.$$

For the logistic Bayesian model of SVMs, $p(y_i | \boldsymbol{x}_i, \tilde{\boldsymbol{w}}, \boldsymbol{\pi})$ is defined as (5). Thus, t_i and m_i become as

$$\begin{aligned} t_i &:= \eta^2 \frac{\exp(-\eta a_i y_i)}{\{1 + \exp(-\eta a_i y_i)\}^2}, \\ m_i &:= -\eta \frac{y_i \exp(-\eta a_i y_i)}{1 + \exp(-\eta a_i y_i)}, \quad \text{for} \quad i = 1, \ldots, l. \end{aligned} \tag{18}$$

The direct calculation of matrices $\hat{\boldsymbol{J}}$ and $\hat{\boldsymbol{I}}$ needs the evaluation of each $\tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_j^\top$.

However, it is difficult to obtain the value of $\tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_j^\top$ because $\tilde{\boldsymbol{z}}_i$ and $\tilde{\boldsymbol{z}}_j$ cannot be calculated explicitly. In order to solve this difficulty, we use the eigenvalues of the matrix $\hat{\boldsymbol{J}}^{-1} \hat{\boldsymbol{I}}$, which can be calculated explicitly[1].

Let $\{\rho_k\}$ and $\{\tilde{\boldsymbol{v}}_k\}$ be the eigenvalues and the corresponding normalized eigenvectors of $\hat{\boldsymbol{J}}^{-1} \hat{\boldsymbol{I}}$, respectively. Then we have

$$\begin{aligned} \hat{\boldsymbol{J}}^{-1} \hat{\boldsymbol{I}} \tilde{\boldsymbol{v}}_k = \rho_k \tilde{\boldsymbol{v}}_k, \quad \text{and} \\ \hat{\boldsymbol{I}} \tilde{\boldsymbol{v}}_k = \rho_k \hat{\boldsymbol{J}} \tilde{\boldsymbol{v}}_k. \end{aligned} \tag{19}$$

[1]Since every t_i in (18) is positive, $\hat{\boldsymbol{J}}$ is positive definite and, therefore, invertible.

Substituting (12) and (15) into (19), we obtain

$$\sum_i \{m_i^2(\tilde{\boldsymbol{z}}_i^\top \tilde{\boldsymbol{v}}_k) - \frac{1}{l} m_i \sum_j m_j(\tilde{\boldsymbol{z}}_j^\top \tilde{\boldsymbol{v}}_k)\}\tilde{\boldsymbol{z}}_i = \sum_i t_i(\tilde{\boldsymbol{z}}_i^\top \tilde{\boldsymbol{v}}_k)\tilde{\boldsymbol{z}}_i + \lambda \tilde{\boldsymbol{I}}_d \tilde{\boldsymbol{v}}_k.$$

Because $\tilde{\boldsymbol{v}}_k$ is normalized, when we take $\gamma \to 0$, the equation (20) becomes

$$\sum_i \{m_i^2(\boldsymbol{z}_i^\top \boldsymbol{v}_k) - \frac{1}{l} m_i \sum_j m_j(\boldsymbol{z}_j^\top \boldsymbol{v}_k)\}\boldsymbol{z}_i = \sum_i t_i(\boldsymbol{z}_i^\top \boldsymbol{v}_k)\boldsymbol{z}_i + \lambda \boldsymbol{I}_d \boldsymbol{v}_k. \tag{20}$$

Therefore, $\boldsymbol{v}_k$ is represented as a linear combination of $\boldsymbol{z}_i$. We set $\boldsymbol{v}_k = \sum_i \mu_{ki}\boldsymbol{z}_i$.

Take the inner products of each side of (20) and a particular $\boldsymbol{z}_q$. The left hand side of the equation is equal to

$$\sum_i \mu_{k_i}\Big(\sum_j m_j^2 K_{ij}K_{jq} - \frac{1}{l}\sum_j m_j K_{ji}\sum_p m_p K_{pq}\Big)$$

where $K_{ij} := \boldsymbol{z}_i^\top \boldsymbol{z}_j$ and $\boldsymbol{K} = (K_{ij})$. The right hand side is equal to

$$\rho_k \sum_i \mu_{ki}\sum_j t_j K_{ji}K_{iq} + \lambda K_{kq}$$

Thus we have

$$\boldsymbol{K}(\mathrm{diag}(\boldsymbol{m}) - \frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\boldsymbol{K}\boldsymbol{\mu}_k = \rho_k(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)\boldsymbol{K}\boldsymbol{\mu}_k.$$

Let $\boldsymbol{\mu}'_k = \boldsymbol{K}\boldsymbol{\mu}_k$. Since $(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)$ is invertible[2], we get the following eigenvalue equation:

$$(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)^{-1}(\boldsymbol{K}\mathrm{diag}(\boldsymbol{m})^2 - \frac{1}{l}\boldsymbol{K}\boldsymbol{m}\boldsymbol{m}^\top)\boldsymbol{\mu}'_{\mathrm{k}} = \rho_{\mathrm{k}}\boldsymbol{\mu}'_{\mathrm{k}}. \tag{21}$$

By solving the eigenvalue equation (21), we have

$$\begin{aligned}\mathrm{trace}(\hat{\boldsymbol{J}}^{-1}\hat{\boldsymbol{I}}) &= \sum_i \rho_i \\ &= \mathrm{trace}[(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)^{-1}(\boldsymbol{K}\mathrm{diag}(\boldsymbol{m})^2 - \frac{1}{l}\boldsymbol{K}\boldsymbol{m}\boldsymbol{m}^\top)].\end{aligned} \tag{22}$$

By substituting (22) into (11) and approximating $\tilde{\boldsymbol{w}}^{*T}$ by $[\hat{\boldsymbol{w}}^\top \ \gamma^{-1}\hat{b}]$ selected by the SVM, we obtain the kernel regularization information criterion (KRIC)

The KRIC for the logistic Bayesian model for SVMs

$$\begin{aligned}\mathrm{KRIC} = 2\Big[&\sum_i \log\{1 + \exp(-\eta a_i y_i)\} \\ &+ \mathrm{trace}\{(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)^{-1}(\boldsymbol{K}\mathrm{diag}(\boldsymbol{m})^2 - \frac{1}{l}\boldsymbol{K}\boldsymbol{m}\boldsymbol{m}^\top)\}\Big]\end{aligned} \tag{23}$$

[2] $|\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_1| = |\mathrm{diag}(\boldsymbol{t})||\boldsymbol{K} + \lambda \mathrm{diag}(\boldsymbol{t})^{-1}| > 0$. Thus, $(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_1)$ is invertible.

where $\boldsymbol{t}_i$ and $\boldsymbol{m}_i$ are (14) and (16). The KRIC for the KLR is the same as the one for the logistic Bayesian model (23) with $\eta = 1$ in (14) and (16).

Next, we construct the KRIC for Sollich's Bayesian model for SVMs. The Bayesian model corresponds to the following regularization model,

$$L_S(D; \boldsymbol{w}, \lambda, \boldsymbol{\pi}) = \log p_S(D|\boldsymbol{w}, \boldsymbol{\pi}) + k_S(\boldsymbol{w}; \lambda)$$

where

$$\log p_S(D|\boldsymbol{w}, \boldsymbol{\pi}) = \log p(D|\boldsymbol{w}, \boldsymbol{\pi}) - l \log N(a)$$

and

$$k_S(\boldsymbol{w}, \lambda) = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + l \log N(a).$$

Since $L_S(D; \boldsymbol{w}; \lambda, \boldsymbol{\pi}) = L(D; \boldsymbol{w}; \lambda, \boldsymbol{\pi})$, the value of $\hat{\boldsymbol{J}}$ and $\hat{\boldsymbol{I}}$ are same as those of the logistic Bayesian model. Therefore, the difference between KRIC for Sollich's model and that for the logistic model is only the likelihood term.

Since $p_0(x)$ is the unknown true probability measure, we use $\hat{N}(a) = \sum_{i=1}^{l} \nu(a(\boldsymbol{x}_i))$ instead of $N(a)$. Consequently, the KRIC becomes as follows:

the KRIC for Sollich's Bayesian model for SVMs

$$\begin{aligned}\mathrm{KRIC} = 2\Big[&\sum_i \log\{1 + \exp(-\eta a_i y_i)\} - l \log \sum_{i=1}^{l} \nu(a(\boldsymbol{x}_i)) \\ &+ \mathrm{trace}\{(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)^{-1}(\boldsymbol{K}\mathrm{diag}(\boldsymbol{m})^2 - \frac{1}{l}\boldsymbol{K}\boldsymbol{m}\boldsymbol{m}^\top)\}\Big]\end{aligned}$$

where $\boldsymbol{t}_i$ and $\boldsymbol{m}_i$ are (14) and (16).

7 The Nyström approximation method for calculation of the KRIC

In this section, we present an approximation for the KRIC by the Nyström method (Williams & Seeger (2001)). The computational cost for the KRIC is $O(l^3)$ because it solves the eigenvalue equation for $l \times l$ matrices. When we use the SMO algorithm, the computational cost for the SVMs' optimization is smaller than $O(l^3)$. Thus, the computational cost for the M-fold cross validation is smaller than $O(Ml^3)$ for a fixed M. This means that if the size of the training data becomes large, KRIC requires higher computational cost than the M-fold cross validation does. We apply the Nyström method to calculations of the KRIC and solve this problem.

In Nyström method, the Gram matrix $\boldsymbol{K} = (K_{ij})$ is approximated by a reduced-rank matrix $\tilde{\boldsymbol{K}}$. Let $(i_1, \ldots, i_m)$ be randomly chosen m indexes such as $1 \leq i_1 < i_2 < \cdots < i_m \leq l$. Define matrices $\boldsymbol{K}_{m,m}$ and $\boldsymbol{K}_{l,m}$ as $\boldsymbol{K}_{m,m}(j,k) = \boldsymbol{K}(i_j, i_k)$ for $1 \leq j, k \leq m$ and $\boldsymbol{K}_{l,m}(i,k) = \boldsymbol{K}(i, i_k)$ for $1 \leq i \leq l$ and $1 \leq k \leq m$. Then $\boldsymbol{K}_{m,m}$ is the restriction of the

Gram matrix $\boldsymbol{K}$ to the randomly chosen m rows and columns and $\boldsymbol{K}_{l,m}$ be the $l \times m$ restricted matrix of $\boldsymbol{K}$ with the same rows as $\boldsymbol{K}_{m,m}$.

Let $\lambda_i^{(m)}$ and $\boldsymbol{u}_i^{(m)}$ be the i-th largest eigenvalue and the corresponding eigenvector. We approximate $\boldsymbol{K}$ by

$$\tilde{\boldsymbol{K}} = \sum_{i=1}^{p} \tilde{\lambda}_i^{(l)} \tilde{\boldsymbol{u}}_i^{(l)} (\tilde{\boldsymbol{u}}_i^{(l)})^\top$$

where

$$\tilde{\lambda}_i^{(l)} := \frac{l}{m} \lambda_i^{(m)},$$

and

$$\tilde{\boldsymbol{u}}_i^{(l)} := \sqrt{\frac{m}{l}} \frac{1}{\lambda_i^{(m)}} \boldsymbol{K}_{l,m} \boldsymbol{u}_i^{(m)}.$$

Let $\tilde{\boldsymbol{U}}$ be an $l \times p$ matrix whose column vectors are $\tilde{\boldsymbol{u}}_i^{(l)}$ $(i = 1, \ldots, p)$ and $\tilde{\boldsymbol{\Lambda}}$ be a $p \times p$ diagonal matrix whose (i, i) component is $\tilde{\boldsymbol{\Lambda}}_i^{(l)}$. Thus,

$$\tilde{\boldsymbol{K}} = \tilde{\boldsymbol{U}} \tilde{\boldsymbol{\Lambda}} \tilde{\boldsymbol{U}}^\top.$$

Using $\tilde{\boldsymbol{K}}$ as an approximation of $\boldsymbol{K}$, the value of $\mathrm{trace}(\hat{\boldsymbol{J}}^{-1} \hat{\boldsymbol{I}})$ becomes as follows:

$$\begin{aligned}
&\mathrm{trace}(\hat{\boldsymbol{J}}^{-1} \hat{\boldsymbol{I}}) \\
&\quad = \mathrm{trace}[(\boldsymbol{K}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)^{-1} (\boldsymbol{K}\mathrm{diag}(\boldsymbol{m}) - \tfrac{1}{l} \boldsymbol{K}\boldsymbol{m}\boldsymbol{m}^\top)] \\
&\quad \simeq \mathrm{trace}[(\tilde{\boldsymbol{K}}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)^{-1} (\tilde{\boldsymbol{K}}\mathrm{diag}(\boldsymbol{m})^2 - \tfrac{1}{l} \tilde{\boldsymbol{K}}\boldsymbol{m}\boldsymbol{m}^\top)] \qquad (24)
\end{aligned}$$

By the matrix inversion lemma, the most right-hand side of (24) is described using $\tilde{\boldsymbol{U}}$ and $\tilde{\boldsymbol{\Lambda}}$ as follows (see the appendix A):

$$\begin{aligned}
&\mathrm{trace}[(\tilde{\boldsymbol{K}}\mathrm{diag}(\boldsymbol{t}) + \lambda \boldsymbol{I}_l)^{-1} (\tilde{\boldsymbol{K}}\mathrm{diag}(\boldsymbol{m})^2 - \frac{1}{l} \tilde{\boldsymbol{K}}\boldsymbol{m}\boldsymbol{m}^\top)] \\
&\quad = \mathrm{trace}[(\tilde{\boldsymbol{U}}^\top \mathrm{diag}(\boldsymbol{t}) \tilde{\boldsymbol{U}} + \lambda \tilde{\boldsymbol{\Lambda}}^{-1})^{-1} \tilde{\boldsymbol{U}}^\top (\mathrm{diag}(\boldsymbol{m})^2 - \frac{1}{l} \boldsymbol{m}\boldsymbol{m}^\top) \tilde{\boldsymbol{U}}]. \qquad (25)
\end{aligned}$$

This is the approximation of the second term of the KRIC by the Nyström method.

Consequently, the KRIC by the Nyström method becomes as follows:

The KRIC for the SVMs by the Nyström method

$$\begin{aligned}
\mathrm{KRIC} = 2\Big[&\log p(D|\hat{\boldsymbol{w}}, \hat{b}, \boldsymbol{\pi}) \\
&+ \mathrm{trace}[(\tilde{\boldsymbol{U}}^\top \mathrm{diag}(\boldsymbol{t}) \tilde{\boldsymbol{U}} + \lambda \tilde{\boldsymbol{\Lambda}}^{-1})^{-1} \tilde{\boldsymbol{U}}^\top (\mathrm{diag}(\boldsymbol{m})^2 - \frac{1}{l} \boldsymbol{m}\boldsymbol{m}^\top) \tilde{\boldsymbol{U}}].
\end{aligned}$$

The computational cost for calculating the KRIC by the Nyström method is $O(m^2 l)$. If we set $m \ll l$, the computational cost for the KRIC becomes much smaller than that for the M-fold cross validation especially when the sample size l is large.

Table 1: Properties of each data set.

data property			
name	dim	#train.	#test
bld	6	230	115
cra	6	133	67
hea	13	180	90
ion	33	234	117
rsy	2	250	1000
snr	60	138	70

8 Experiments

We compared the performance of the KRIC and three other criteria, 10-fold cross validation, the evidence calculated by Laplace's method (Kwok (2000)) and the $\xi\alpha$-estimation (Joachims (2000)).

When we calculate the evidence, we assume the logistic Bayesian model for SVMs. The evidence becomes

$$\begin{aligned}&\log \int p(D|\boldsymbol{w})p(\boldsymbol{w}|\lambda)d\theta \\ &\quad \simeq \log p(D|\hat{\boldsymbol{w}}) - \frac{1}{2}\log|\frac{1}{\lambda}\mathrm{diag}(\boldsymbol{t})^{1/2}\boldsymbol{K}\mathrm{diag}(\boldsymbol{t})^{1/2} + \boldsymbol{I}_1| + (\text{independent of } \lambda)\end{aligned}$$

where $\hat{\boldsymbol{w}}$ is the optimal value calculated by SVMs or KLR, and $\boldsymbol{t}$ and $\boldsymbol{K}$ are the same value used in the KRIC.

The $\xi\alpha$-estimation is defined by

$$Err^l_{\xi\alpha} = \frac{1}{l}|\{i; (\rho\alpha_i R^2_\Delta) \geq 1\}|$$

where $\rho = 2, R^2_\Delta = 1$ as set in Joachims (2000). We omit the detail of Wahba's GACV. See Wahba (1998), for example.

In the experiment, Gaussian kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2)$ is used. The regularization parameter C is tuned by each criterion. We use the data sets in Gestel et al. (2002). The BUPA Liver Disorders (bld), the Statlog Heart Disease (hea), the John Hopkins University Ionosphere (ion), and the Sonar (snr) were downloaded from UCI benchmark repository (Blake & Merz (1998)). The synthetic data set (rsy) and Leptograpsus crabs (cra) are described in Ripley (1996). The property of each data set is Table 1. "dim", "#train", and "#test" denote the input dimension, the size of training data and the size of test data, respectively.

Each data set is split up into training data sets (2/3) and test data sets (1/3), except for the rsy data set, which was originally divided 250 training data set and 1000 test

data set. Each data set is divided into training datasets and test datasets randomly. We repeated the random separation 100 times and make 100 different data sets. Each data set $\{\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})^\top\}_{i=1}^L$ is normalized as $\sum_{i=1}^l x_{ij}^2 = 1$ for $j = 1, \ldots, n$, where L is the sum of the number of the training and test datasets.

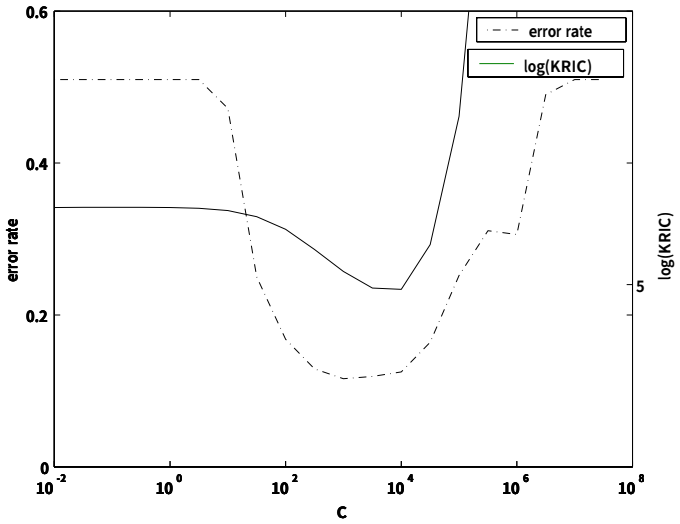

Figure 3: An example of regularization parameter tuning by the KRIC. We use `rsy` data set. The solid line and dashed line represent the KRIC value and test error rate, respectively. The scale parameter in the Gaussian kernel, σ, is fixed at 10^{-1}.

We set the parameter space as $\mathcal{C} = \{10^{k/2-2}\}_{k=0}^{19}$. We select the best value of C in the set $\mathcal{C}$ by maximization or minimization of each criterion. The scale parameter in the Gaussian kernel σ is set as 10.0.

Tables 2 and 3 are the results of comparison of the performance of the criteria for SVMs and KLR, respectively. We compare KRIC (the logistic model), GACV, 10-fold cross validation, KRIC (Sollich's model), the evidence calculated by Laplace's method and $\xi\alpha$-estimation. The two types of KRIC and the evidence are calculated by the Nyström approximation. In the Nyström approximation, the size of randomly selected submatrix is set as $m = 50$. The number of selected principal components is set as $p = 30$. ERROR denotes the average test error rate together with the standard deviation. TIME denotes the average CPU time in seconds consumed by calculation of each criterion for tuning the regularization parameter together with the standard deviation.

Each SVM and 10-fold cross validation was calculated by OSU-SVM algorithm, which is a Matlab's mex program created and distributed by J. Ma, Y. Zhao, and S. Ahalt. We used 1.6-GHz Pentium machine. We made a Matlab code (which is not compiled as mex program) for evaluating KLR by the SMO algorithm based on the works, Keerthi et al. (2002) and Zhu & Hastie (2002).

We computed p-value for the paired t-test on error difference between the KRIC with Nyström method and the other criteria. We also computed the p-value on CPU time

difference. A p-value threshold of 0.05 was used to decide the significant difference. In table 2 and 3, each mark * (or @) on the left of the p-value of a criterion represents that the performance of the criterion is significantly better (or worse) than that of the KRIC.

From the result, we say that there is no criterion whose average test error is significantly lower than that of the KRIC. The error rate of the KRIC for the logistic Bayesian model is lower than that of the GACV and $\xi\alpha$-estimation. Average computational cost of the KRIC is lower than that of the 10-fold cross validation and calculation of the evidence by Laplace's method. On the other hand, the calculation of $\xi\alpha$-estimation and GACV is faster than that of KRIC. The average test error given by the KRIC for Sollich's Bayesian model is significantly worse than that for the logistic Bayesian model for some data set.

We note that the test error rate given by the KRIC for KLR is better than that for SVMs in comparison with other criteria. In particular, the average error rate given by the KRIC is better than the one given by the 10-fold cross validation and the evidence for most of the data sets though the difference is not significant in the t-test.

Figure 3 is an example of regularization parameter tuning by the KRIC. We used `rsy` data set. The kernel parameter σ is fixed at 10^{-1}.

9 Conclusion and discussion

The kernel information criterion (KRIC) was introduced for tuning the regularization parameter in the SVMs and the KLR. The criterion can be calculated with low computational cost if we used some approximation methods.

First we constructed the statistical models of SVMs and KLR. Two different Bayesian models for SVMs are considered. In the first model, the logistic Bayesian model, is the hinge loss function of SVMs is approximated by a logistic function. Thus, the Bayesian model corresponds to that of KLR. The second model, proposed by Sollich, is an exact Bayesian model for SVMs.

We considered two regularization models corresponding to each of the Bayesian models for SVMs. The regularization information criterion (RIC) for these models cannot be directly evaluated because the parameter space is the reproducing kernel Hilbert space. In order to derive the RIC, we introduce an eigenvalue problem and prove that the RIC is calculated by the eigenvalues. We call the RIC calculated in this way as the kernel RIC (KRIC).

Next an approximation of the KRIC by using the Nyström method was presented. In the Nyström method, the Gram matrix is approximated by the submatrix and the principal components. Therefore, the computational cost for calculation of the KRIC by Nyström method becomes very low. Then the model selection by the KRIC with the approximation becomes much faster than the one by the cross-validation when the sample size is large.

From the experimental result, the average test error rate with the regularization parameter tuned by KRIC for the logistic Bayesian model is lower than the one by the GACV and the $\xi\alpha$-estimation. The average error rate is comparable with the one by the

10-fold cross validation and the evidence evaluated by Laplace's method. Computational cost for calculating the KRIC is significantly lower than the one for these two criteria.

In the present work, the KRIC is used only for tuning the regularization parameter. However, the kernel parameters and kernel function can be tuned by the KRIC.

We explain the reason. The KRIC evaluates the left-hand side of (10) where $p_0(D)$ is the true density function and $\theta^*(D', \lambda)$ maximizes $L(D'; \theta; \lambda)$. Since the value of $L(D'; \theta; \lambda)$ depends on the kernel parameter $\boldsymbol{\pi}$ and the kernel function K, $\theta^*(D', \lambda)$ also depends on $\boldsymbol{\pi}$ and K. It is the reason why the KRIC can be used for selecting $\boldsymbol{\pi}$ and K, too.

We carried out some experiments for evaluating the tuning of the kernel parameter by the KRIC. We use the same data sets as the one used in section 8 for selecting both of the scale parameter σ in the Gaussian kernel and the regularization parameter C. The scale parameter σ is tuned well by the KRIC when we restrict σ on $\sigma \gtrsim 10^{-4}$. When we consider $\sigma \lesssim 10^{-4}$, the KRIC takes a global minimum in such a region of σ. However, the global minimum of the test error is not attained in the region.

One naive solution of this problem is to exclude the case σ is smaller than a particular value (for example 10^{-4} in the above experiments.) At least in the experiments, the optimal parameters (in the sense of the lowest test error) were not in the excluded region because σ is so small that the overfitting occurs. Thus the setting of the region of σ is reasonable. Moreover, the boundary value as 10^{-4} should be taken depending on the sample size. Further research on the selection of kernel parameters and kernel functions by the KRIC is a future work.

In Sollich (2002), he suggested introducing a prior distribution on the bias parameter b in the Bayesian model for the SVMs. If we assume a prior distribution on the bias parameter b, the original quadratic problem of SVMs are changed. In the present work, we have been assumed no prior distribution on the bias b.

However, our results can be easily applied to the modified SVMs which have a square penalty on the bias. Consider the modified objective function (8). Then the regularization likelihood (11) changes to

$$L(D; \boldsymbol{w}, b; \lambda, \boldsymbol{\pi}) = \log p(D \mid \boldsymbol{w}, b, \boldsymbol{\pi}) - \frac{\lambda}{2}\|\boldsymbol{w}\|^2 - \frac{\lambda B^{-2}}{2} b^2.$$

Let $\tilde{\boldsymbol{w}}^\top = [\boldsymbol{w}^\top B^{-1} b]$ and $\tilde{\boldsymbol{z}}^\top = [\boldsymbol{z}^\top - B]$. The KRIC for this model is given by substituting

$$\tilde{\boldsymbol{I}}_d := \boldsymbol{I}_{d+1}$$

to (12) and (17) instead of (13).

The above regularization model has two regularization parameters. If we tune both of the regularization parameters, the least value of the KRIC becomes smaller than the one given by tuning only one regularization parameter. This means that we cannot compare models which have different numbers of regularization parameters by the KRIC. In other words, the KRIC cannot avoid the overfitting with respect to the regularization parameters.

As the KRIC in the regularization models, we cannot compare Bayesian models which have different number of hyper parameters by the type-II likelihood. One solution of this

problem for the type-II likelihood is using the Akaike's Bayesian information criterion (ABIC) (Akaike (1980)):

$$ABIC = -2\log \int p(D|w)p(w|\lambda)dw + 2\mathrm{dim}(\lambda).$$

ABIC is recognized as AIC for the type-II likelihood.

As the KRIC, ABIC can be used not only for selecting the regularization parameter but also for the kernel parameters i.e.

$$ABIC = -2\log \int p(D|w,\boldsymbol{\pi})p(w|\lambda,\boldsymbol{\pi})dw + 2\mathrm{dim}(\lambda,\boldsymbol{\pi}).$$

If we use some kernel generating methods presented in Cristianini & Shawe-Taylor (2000), we can obtain very large class of kernel functions. When kernel functions are generated by using these methods, the dimensions of the kernel parameters can become large.

Although there are a lot of researches on evaluating the first term of ABIC (i.e. the type-II likelihood), there is no method whose accuracy is reliable comparing with the second term. The model selection by the KRIC, ABIC and other criteria is a future work when the number of the regularization and kernel parameters in each model is different.

The effective number of parameters (Bishop (1995)) for the regularization model with the regularization parameter λ is

$$\sum_j \frac{\lambda_j}{\lambda_j + \lambda}$$

where λ_j are eigenvalues of the Hessian matrix. In the Bayesian model for the SVMs,

$$\lambda_j = \sum_i t_i(\tilde{\boldsymbol{z}}_i^\top \tilde{\boldsymbol{z}}_j)$$

and t_i is given by (18). This might seem to be similar to the KRIC's penalty term. However, these two criteria are totally different ones. The effective number of parameters is derived from evaluation of the evidence value for Gaussian inferences. The KRIC is derived from minimization of a Kullback-Leibler divergence. We have interest in some potential connections between these two criteria if they exist.

If the dimension of parameters is finite, RIC is proved to be an asymptotic unbiased estimator of the penalized log likelihood (Konishi & Kitagawa (1996)). However, the Bayes models corresponding to SVMs have parameters whose dimension is as many as the number of the samples; sometimes infinity. Thus, further studies are required to validate KRIC rigorously.

Recently, a lot of attention is paid to least-square SVMs, flexible discriminant analysis (Hastie et al. (2001)) and dual penalized logistic regression machines (Tanabe (2001)) as substitutions of ordinary SVMs (Gestel et al. (2002)). It has been presented that all of them sometimes perform better than ordinal SVMs. Our research presented in this chapter could be generalized for these models.

10 Appendix: The derivation of (25)

We prove the (25):

$$\begin{aligned}&\mathrm{trace}[(\tilde{\boldsymbol{K}}\mathrm{diag}(\boldsymbol{t})+\lambda\boldsymbol{I}_l)^{-1}(\tilde{\boldsymbol{K}}\mathrm{diag}(\boldsymbol{m})^2-\frac{1}{l}\tilde{\boldsymbol{K}}\boldsymbol{m}\boldsymbol{m}^\top)]\\&\quad=\mathrm{trace}[(\tilde{\boldsymbol{U}}^\top\mathrm{diag}(\boldsymbol{t})\tilde{\boldsymbol{U}}+\lambda\tilde{\boldsymbol{\Lambda}}^{-1})^{-1}\tilde{\boldsymbol{U}}^\top(\mathrm{diag}(\boldsymbol{m})^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}].\end{aligned}$$

Let $\boldsymbol{M}:=\mathrm{diag}(\boldsymbol{m})$ and $\boldsymbol{T}:=\mathrm{diag}(\boldsymbol{t})$ for shortage of the description.
Using the matrix inversion lemma,

$$\begin{aligned}&\mathrm{trace}[(\boldsymbol{K}\boldsymbol{T}+\lambda\boldsymbol{I}_l)^{-1}\boldsymbol{K}(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)]\\&\quad=\mathrm{trace}[\frac{1}{\lambda}\boldsymbol{T}^{-1}\{\boldsymbol{T}-\boldsymbol{T}\tilde{\boldsymbol{U}}(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\}K(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)]\\&\quad=\frac{1}{\lambda}\mathrm{trace}[\{\boldsymbol{I}_l-\tilde{\boldsymbol{U}}(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\}K(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)]\\&\quad=\frac{1}{\lambda}\mathrm{trace}[\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)]\\&\qquad-\frac{1}{\lambda}\mathrm{trace}[(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\end{aligned}\tag{26}$$

The second term of the most right side is evaluated as follows:

$$\begin{aligned}&\mathrm{trace}[(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\\&\quad=\mathrm{trace}[(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\\&\qquad-\mathrm{trace}[\lambda(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\\&\quad=\mathrm{trace}[\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]-\mathrm{trace}[\lambda(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\end{aligned}$$

Substituting this to (26),

$$\begin{aligned}&\mathrm{trace}[(\boldsymbol{K}\boldsymbol{T}+\lambda\boldsymbol{I}_l)^{-1}\boldsymbol{K}(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)]\\&\quad=\frac{1}{\lambda}\mathrm{trace}[\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}^{-1}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)]-\frac{1}{\lambda}\mathrm{trace}[\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\\&\quad+\mathrm{trace}[(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\\&\quad=\mathrm{trace}[(\lambda\tilde{\boldsymbol{\Lambda}}^{-1}+\tilde{\boldsymbol{U}}^\top\boldsymbol{T}\tilde{\boldsymbol{U}})^{-1}\tilde{\boldsymbol{U}}^\top(\boldsymbol{M}^2-\frac{1}{l}\boldsymbol{m}\boldsymbol{m}^\top)\tilde{\boldsymbol{U}}]\end{aligned}$$

Thus, (25) is valid.

References

Akaike, H. (1980). Likelihood and Bayes procedure with discussion. In *Bayesian Statistics*, D. L. J. Bernardo, M. DeGreet & A. Smith, eds. Valencia, Spain: University Press.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. New York: Oxford University Press Inc.

Blake, L. & Merz, J. (1998). Uci repository of machine learning databases. Tech. rep., University of California, Department of Information and Computer Science, Irvine, CA.

Burnham, K. P. & Anderson, D. P. (2002). *Model Selection and Multimodel Inference — Practical Information Theoretic Approach*. New York: Springer-Verlag, 2nd ed.

Cristianini, N. & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

Duan, K., Keerthi, S. S. & Poo, A. N. (2003). Evaluation of simple performance measures for tuning svm hyperparameters. *Neurocomputing* **51**, 41–59.

Gestel, T., Suykens, J., Lanckried, G., Lambrechts, A., Moor, B. & Vandewalle, J. (2002). Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. *Neural Computation* **14**, 1115–1147.

Good, I. J. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning – Data mining, inference, and prediction*. springer series in statistics. New York: Springer.

Jaakkola, T. & Haussler, D. (1999). Probabilistic kernel regression models. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*. San Francisco: Morgan Kaufmann.

Joachims, T. (2000). Estimating the generalization performance of a svm efficiently. In *Proceedings of the International Conference on Machine Learning*.

Keerthi, S. S., Duan, K., Shevade, S. K. & Poo, A. N. (2002). A fast dual algorithm for kernel logistic regression. In *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002)*, C. Sammut & A. G. Hoffmann, eds. University of New South Wales, Sydney, Australia: Morgan Kaufmann.

Konishi, S. & Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika* **83**, 875–890.

Kwok, J. (1999). Moderating the outputs of support vector machine classifiers. *IEEE Transactions on Neural Networks* **10**, 1018–1031.

Kwok, J. (2000). The evidence framework applied to support vector machines. *IEEE Transactions on Neural Networks* **11**, 1162–1173.

MacKay, D. (1992a). Bayesian interpolation. *Neural Computation* **4**, 415–447.

MacKay, D. (1992b). The evidence framework applied to classification networks. *Neural Computation* **4**, 698–714.

Opper, M. & Winther, O. (2000). Gaussian processes for classification: Mean field algorithms. *Neural Computation* **12**, 2655–2684.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. Tech. Rep. MSR-TR-98-14, Microsoft Research, Redmond.

Ripley, G. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

Schölkopf, P. & Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press.

Seeger, M. (2000). Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers. In *Advances in Neural Information Processing Systems 12*, T. L. S. Solla & K. R. Müller, eds., vol. 12. Cambridge, MA: MIT Press.

Shibata, R. (1989). Statistical aspects of model selection. In *From data to model*, J. C. Williams, ed. London: Springer-Verlag, pp. 215–240.

Smola, A., Ovari, Z. & Williamson, R. (2000). Regularization with dot-product kernels. *Advances in Neural Information Processing Systems* **13**.

Smola, A., Schölkopf, B. & Müller, K. (1998). The connection between regularization operators and support vector kernels. *Neural Networks* **11**, 637–649.

Sollich, P. (2002). Bayesian methods for support vector machines: evidence and predictive class probabilities. *Machine Learning* **46**, 21–52.

Takeuchi, K. (1978). Distribution of information statistics and validity criteria of models. *Mathematical Science* **153**, 12–18.

Tanabe, K. (2001). Penalized logistic regression machines: New methods for statistical prediction. *ISM cooperative research report 143 — Estimation and Smoothing Methods in Nonparametric Statistical Models* , 163–194.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Vapnik, V. & Chapelle, O. (2000). Bounds on error expectation for support vector machine. *Neural Computation* **12**, 2013–2036.

Wahba, G. (1998). Support vector machines, reproducing kernel hilbert spaces, and randomized gacv,. In *Advances in Kernel Methods — Support Vector Learning*, B. Shölkoph, C. J. C. Burges & A. J. Smola, eds. Cambridge, MA: MIT Press, pp. 69–88.

Wahba, G., Gu, C., Wang, Y. & Chappell, R. (1995). Soft classification, a.k.a. risk estimation, via penalized log likelihood and smoothing spline analysis of variance. In *Mathematics of Generalization*. Addison-Wesley Publisher, pp. 329–360.

Williams, C. & Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **20**, 1342–1351.

Williams, C. & Seeger, M. (2001). Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*. Cambridge, MA: MIT Press.

Zhu, J. & Hastie, T. (2002). Kernel logistic regression and the import vector machine. In *Advanced in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press.

Table 2: Comparison of criteria for tuning the regularization parameter of SVMs (Gaussian kernel, $\sigma = 10.0$). We compare KRIC (for the logistic model), GACV, 10-fold cross validation, KRIC (for Sollich's model), the evidence calculated by Laplace's method, and the $\xi\alpha$-estimation. The Nyström approximation is used in the calculation of the two types of KRIC and the evidence.

data set		`bld`		`cra`	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.2821 ± 0.0344	-	0.0000 ± 0.0000	-
	TIME(s)	1.6829 ± 1.1600	-	0.1070 ± 0.1428	-
KRIC	ERROR	0.2888 ± 0.0355	-	0.0009 ± 0.0036	-
(logistic)	TIME(s)	0.2456 ± 0.0128	-	0.1088 ± 0.0080	-
GACV	ERROR	0.4131 ± 0.0381	@0.0456	0.5391 ± 0.0285	@ 0.0000
	TIME(s)	0.1575 ± 0.0040	*0.0000	0.0621 ± 0.0026	* 0.0000
10-fold	ERROR	0.2908 ± 0.0377	0.4892	0.0004 ± 0.0013	0.4559
CV	TIME(s)	13.2987 ± 4.2776	@0.0012	0.8745 ± 0.5552	0.0870
KRIC	ERROR	0.3891 ± 0.0458	0.1085	0.4582 ± 0.0818	@ 0.0000
(Sollich)	TIME(s)	0.2465 ± 0.0164	0.4883	0.1037 ± 0.0076	0.3705
evidence	ERROR	0.3204 ± 0.0383	0.3340	0.0009 ± 0.0036	0.5000
(Lap.)	TIME(s)	0.5258 ± 0.0201	@0.0000	0.1933 ± 0.0063	@ 0.0000
$\xi\alpha$-est.	ERROR	0.4131 ± 0.0381	@0.0456	0.0009 ± 0.0036	0.5000
	TIME(s)	0.1869 ± 0.0041	*0.0003	0.0826 ± 0.0017	* 0.0034

data set		`hea`		`ion`	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.2109 ± 0.0307	-	0.0855 ± 0.0221	-
	TIME(s)	2.2332 ± 1.6830	-	0.4544 ± 0.2921	-
KRIC	ERROR	0.2340 ± 0.0371	-	0.1342 ± 0.0193	-
(logistic)	TIME(s)	0.1959 ± 0.0144	-	0.1450 ± 0.0184	-
GACV	ERROR	0.4622 ± 0.0645	@0.0124	0.1598 ± 0.0221	0.2680
	TIME(s)	0.1474 ± 0.0058	*0.0083	0.2369 ± 0.0208	@ 0.0097
10-fold	ERROR	0.2305 ± 0.0349	0.4806	0.1024 ± 0.0324	0.2696
CV	TIME(s)	17.9704 ± 8.2644	@0.0159	2.9289 ± 1.3944	@ 0.0244
KRIC	ERROR	0.3983 ± 0.1020	0.1188	0.1342 ± 0.0193	0.5000
(Sollich)	TIME(s)	0.1975 ± 0.0159	0.4786	0.1424 ± 0.0174	0.4714
evidence	ERROR	0.2321 ± 0.0341	0.4894	0.1111 ± 0.0248	0.3007
(Lap.)	TIME(s)	0.4804 ± 0.0293	@0.0000	0.6176 ± 0.0746	@ 0.0000
$\xi\alpha$-est.	ERROR	0.4233 ± 0.0980	0.0807	0.1205 ± 0.0195	0.3624
	TIME(s)	0.1845 ± 0.0054	0.2828	0.0599 ± 0.0056	* 0.0000

data set		rsy		snr	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.1072 ± 0.0069	-	0.2057 ± 0.0345	-
	TIME(s)	7.3261 ± 8.1287	-	0.0793 ± 0.0136	-
KRIC	ERROR	0.1112 ± 0.0060	-	0.2429 ± 0.0517	-
(logistic)	TIME(s)	0.1956 ± 0.0173	-	0.1090 ± 0.0277	-
GACV	ERROR	0.4835 ± 0.0755	@ 0.0000	0.3071 ± 0.0828	0.3164
	TIME(s)	0.1062 ± 0.0044	* 0.0000	0.1027 ± 0.0197	0.4472
10-fold	ERROR	0.1100 ± 0.0073	0.4639	0.2371 ± 0.0358	0.4740
CV	TIME(s)	38.4810 ± 14.2007	@ 0.0035	0.5163 ± 0.0670	@0.0000
KRIC	ERROR	0.2858 ± 0.0706	@ 0.0113	0.2443 ± 0.0454	0.4941
(Sollich)	TIME(s)	0.2113 ± 0.0095	0.2788	0.1069 ± 0.0182	0.4822
evidence	ERROR	0.1138 ± 0.0065	0.4171	0.2229 ± 0.0382	0.4120
(Lap.)	TIME(s)	0.3859 ± 0.0184	@ 0.0000	0.2975 ± 0.0436	@0.0041
$\xi\alpha$-est.	ERROR	0.1080 ± 0.0076	0.4067	0.2430 ± 0.0516	0.4994
	TIME(s)	0.1461 ± 0.0024	* 0.0058	0.2403 ± 0.0099	*0.0002

Table 3: Comparison of criteria for tuning the regularization parameter of KLR (Gaussian kernel, $\sigma = 10.0$). We compare KRIC (the logistic model), GACV, 10-fold cross validation, KRIC (Sollich's model), the evidence calculated by Laplace's method, and $\xi\alpha$-estimation. The two types of KRIC and the evidence are calculated by the Nyström approximation. The p-value is for the paired t-test on error rate and CPU time with respect to KRIC (the logistic model).

data set		bld		cra	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.3374 ± 0.0426	-	0.0060 ± 0.0077	-
	TIME(s)	12.2935 ± 0.4063	-	8.1325 ± 0.4327	-
KRIC	ERROR	0.3409 ± 0.0430	-	0.0060 ± 0.0077	-
(logistic)	TIME(s)	0.1519 ± 0.0102	-	0.0881 ± 0.0036	-
GACV	ERROR	0.4200 ± 0.0318	0.1448	0.5313 ± 0.0316	@0.0000
	TIME(s)	0.8097 ± 0.5817	0.1332	0.0893 ± 0.0047	0.4396
10-fold	ERROR	0.3452 ± 0.0382	0.4786	0.0164 ± 0.0192	0.3489
CV	TIME(s)	108.8243 ± 7.3829	@0.0000	71.6293 ± 5.6273	@0.0000
KRIC	ERROR	0.4217 ± 0.0326	0.1422	0.3940 ± 0.0581	@0.0000
(Sollich)	TIME(s)	0.1517 ± 0.0093	0.4971	0.0941 ± 0.0168	0.3844
evidence	ERROR	0.3417 ± 0.0420	0.4959	0.0060 ± 0.0077	0.5000
(Lap.)	TIME(s)	0.3304 ± 0.0238	@0.0000	0.1311 ± 0.0111	@0.0017
$\xi\alpha$-est.	ERROR	0.4217 ± 0.0326	0.1422	0.5313 ± 0.0316	@0.0000
	TIME(s)	0.1016 ± 0.0072	*0.0019	0.0327 ± 0.0043	*0.0000

data set		hea		ion	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.2011 ± 0.0430	-	0.1077 ± 0.0352	-
	TIME(s)	10.9960 ± 0.5419	-	14.9930 ± 0.8140	-
KRIC	ERROR	0.2167 ± 0.0511	-	0.1103 ± 0.0355	-
(logistic)	TIME(s)	0.1515 ± 0.0139	-	0.2555 ± 0.0370	-
GACV	ERROR	0.2878 ± 0.1476	0.3603	0.1863 ± 0.0516	0.1910
	TIME(s)	0.2352 ± 0.0278	@0.0224	1.0713 ± 0.1496	@0.0000
10-fold	ERROR	0.2181 ± 0.0342	0.4931	0.1098 ± 0.0351	0.4976
CV	TIME(s)	98.4162 ± 8.6210	@ 0	129.8559 ± 17.2304	@0.0000
KRIC	ERROR	0.2456 ± 0.0543	0.3920	0.1265 ± 0.0471	0.4221
(Sollich)	TIME(s)	0.1526 ± 0.0167	0.4858	0.2557 ± 0.0365	0.4991
evidence	ERROR	0.2167 ± 0.0511	0.5000	0.1128 ± 0.0351	0.4855
(Lap.)	TIME(s)	0.2826 ± 0.0308	@0.0017	1.3124 ± 0.2161	@0.0000
$\xi\alpha$-est.	ERROR	0.4344 ± 0.1036	0.0797	0.3530 ± 0.0472	@0.0017
	TIME(s)	0.0642 ± 0.0057	*0.0000	0.5334 ± 0.0868	*0.0124

data set		rsy		snr	
		ERROR or TIME	p-value	ERROR or TIME	p-value
optimal	ERROR	0.1206 ± 0.0089	-	0.2143 ± 0.0243	-
	TIME(s)	20.6412 ± 0.5797	-	9.3232 ± 0.7657	-
KRIC	ERROR	0.1235 ± 0.0098	-	0.2343 ± 0.0279	-
(logistic)	TIME(s)	0.3786 ± 0.0199	-	0.1755 ± 0.0201	-
GACV	ERROR	0.2593 ± 0.1740	0.2301	0.4671 ± 0.0738	@0.0110
	TIME(s)	0.8985 ± 0.0943	@0.0000	0.2693 ± 0.0488	0.0868
10-fold	ERROR	0.1233 ± 0.0112	0.4968	0.2529 ± 0.0261	0.3656
CV	TIME(s)	88.7065 ± 5.6848	@0.0000	84.3182 ± 10.6565	@0.0000
KRIC	ERROR	0.2535 ± 0.0430	@0.0070	0.2729 ± 0.0353	0.2709
(Sollich)	TIME(s)	0.3917 ± 0.0176	0.3628	0.1665 ± 0.0218	0.4145
evidence	ERROR	0.1283 ± 0.0129	0.4163	0.2371 ± 0.0271	0.4793
(Lap.)	TIME(s)	0.6113 ± 0.0193	@0.0000	0.5545 ± 0.0746	@0.0000
$\xi\alpha$-est.	ERROR	0.4697 ± 0.1159	@0.0030	0.4643 ± 0.0817	@0.0180
	TIME(s)	0.1487 ± 0.0030	*0.0000	0.1640 ± 0.0160	0.3745