Model Selection for Linear Classifiers using Bayesian Error Estimation

Heikki Huttunen[a,*], Jussi Tohka[b]

[a] *Department of Signal Processing, Tampere University of Technology, Finland*
[b] *Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Spain*

Abstract

Regularized linear models are important classification methods for high dimensional problems, where regularized linear classifiers are often preferred due to their ability to avoid overfitting. The degree of freedom of the model is determined by a regularization parameter, which is typically selected using counting based approaches, such as K-fold cross-validation. For large data, this can be very time consuming, and, for small sample sizes, the accuracy of the model selection is limited by the large variance of CV error estimates. In this paper, we study the applicability of a recently proposed Bayesian error estimator for the selection of the best model along the regularization path. We also propose an extension of the estimator that allows model selection in multiclass cases and study its efficiency with L_1 regularized logistic regression and L_2 regularized linear support vector machine. The model selection by the new Bayesian error estimator is experimentally shown to improve the classification accuracy, especially in small sample-size situations, and is able to avoid the excess variability inherent to traditional cross-validation approaches. Moreover, the method has significantly smaller computational complexity than cross-validation.

Keywords: Logistic regression, Support vector machine, Regularization, Bayesian error estimator, Linear classifier

*Corresponding author: Heikki Huttunen, heikki.huttunen@tut.fi

1. Introduction

The task in supervised classification is to learn to make predictions about the class of an unknown object given a training set of P-dimensional feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$ with known class memberships. An important special case of supervised classification problems arises when the number of features P is larger or nearly as large than the number of training samples N. These classification problems are increasingly important, for example, in genomics and neuroimaging [1, 2]. Due to a small number of training samples (compared to the data dimensionality), linear classifiers are preferred in such cases. Also, some form of regularization is necessary to cope with small N.

In this paper, we concentrate on two widely used regularized linear classifiers: L_1 or LASSO regularized logistic regression [3, 4, 5] and support vector machine (SVM) [6, 7]. These classifiers are trained by minimizing a cost function that is a weighted sum of data term and a regularization term. The usual strategy for selecting the weights (or the value for the regularization parameter) is to train classifiers for various values of regularization parameter producing a set of models and then select the best model according to some model selection criteria. The most widely used approach is to select the best model based on a (non-parametric) estimate of classification error, such as cross-validation (CV), bootstrap, or resubstitution error estimators.

The randomness of the cross-validation has certain drawbacks: The model selection depends on the particular split of the data, the approach is time consuming, and the resulting error estimate may have a large variance [8]. In particular, the latter problem has been documented already almost four decades ago [9], but it is still often dismissed [8]. Thus, we are interested in finding a *deterministic*, *accurate* and *fast* approach for choosing the regularization parameter of a regularized linear classification model.

Other approaches for model selection include information theoretic tools, such as the Akaike Information Criterion (AIC) [10], the Bayesian Information Criterion (BIC) [11]; including its use for logistic regression models [12] and

SVMs [13], and extended BIC (EBIC) [12], which corrects drawbacks of the BIC when $P > N$. However, all of the above are based on the likelihood of the model, not on the prediction error. In predictive modeling, the actual model is often of secondary importance, and the critical issue is the prediction ability and the minimal error. For the SVM model selection, various techniques based on different error bounds and concepts from algorithmic information theory have been suggested [14, 15]. However, these are either complicated and expensive to compute or do not yield satisfactory results [16]. Probably for this reason, the K-fold CV is still the most popular model selection criterion also in small sample settings.

Recently, a few alternatives to the CV type methods for the estimation of classification error have been proposed. One of them is the *Bolstered error estimation* [17], which attempts to smooth the empirical distribution of the available data by placing bolstering kernels at each data point location. A more recent approach, the *Bayesian minimum mean-square estimator for classification error* describes the error in a Bayesian framework [18, 19]. Moreover, a closed form expression can be derived for the posterior expectation of the classification error in the binary classification case under mild assumptions about the covariance structure. The method is attractive, because the errors are estimated directly from the training data, and no iterative resampling or splitting operations are required. This results also in a significant speedup, since the classifier training is done only once. For example, the 5-fold CV (CV-5) includes five training iterations on partial data and one on all training data, while the Bayesian error estimator requires only the last training step with all data.

Experimental data suggests that the Bayesian Error Estimator (BEE) can be more accurate in absolute terms than the CV-based classification error estimates, in particular with small sample sizes [19]. In our earlier work, we have shown that the BEE is accurate for model selection as well [20]. More specifically, we compared the BEE with CV and BIC criteria when used for selecting the regularization parameter λ for the binary logistic regression model with LASSO penalty. This paper extends the earlier study by 1) proposing a

Bayesian model selection rule for multinomial classification problems, 2) considering BEE model selection under more general priors than in [20] and 3) studying the rule for selection of the regularization parameter for both SVM and logistic regression classifiers. Moreover, extensive experiments show that the BEE criterion is significantly faster than the CV, and also more accurate unless the model assumptions are severely violated. The implementation of the proposed error estimator in Matlab and Python is available for download[1].

The rest of this paper is organized as follows: In Section 2 we will briefly review the regularized logistic regression and SVM classifiers; Section 3 defines the Bayesian error estimator for binary and multiclass cases; and Section 4 compares the accuracy of the BEE to CV and BIC based model selection in various experimental cases. Finally, Section 5 discusses the applicability and the limits of the proposed method.

2. Linear Classifiers

In the following, we denote the observation matrix as $\mathbf{X} \in \mathbb{R}^{N \times P}$, whose rows $\mathbf{x}_i$ are the samples with corresponding class labels $\mathbf{y} = (y_1, \ldots, y_N)^T$ with $y_i = \{-1, 1\}$ in the 2-class case and $y_i \in \{1, 2, \ldots, C\}$ in the multiclass case. The predicted class label $\hat{y}$ for the feature vector $\mathbf{x}$ is given by $\hat{y} = \text{sign}(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}) \doteq g(\mathbf{x})$ in the binary case and $\hat{y} = \arg\max_c(\beta_{c,0} + \boldsymbol{\beta}_c^T \mathbf{x})$ in the multiclass case, where the classifier parameters $\beta_0, \beta_{c,0} \in \mathbb{R}$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^T, \boldsymbol{\beta}_c = (\beta_{c,1}, \beta_{c,2}, \ldots, \beta_{c,P})^T \in \mathbb{R}^P$ are learned from training data.

2.1. Regularized Logistic Regression

Logistic regression (LR) is a statistical classification method modeling the class conditional probability densities by the logistic function. Binary logistic regression models the class probabilities of the sample $\mathbf{x} = (x_1, x_2, \ldots, x_P)^T \in$

[1] https://sites.google.com/site/bayesianerrorestimate/

$\mathbb{R}^P$ belonging to class $c \in \{-1, 1\}$ as

$$\Pr(c \mid \mathbf{x}) = \frac{1}{1 + \exp\left[c(\beta_0 + \mathbf{x}^T\boldsymbol{\beta})\right]}.$$

A slightly different model is adopted for multinomial logistic regression for C classes [5], which models the probability $\Pr(c \mid \mathbf{x})$ of the sample $\mathbf{x}$ belonging to class $c \in \{1, 2, \ldots, C\}$ as

$$\Pr(c \mid \mathbf{x}) = \frac{\exp(\beta_{c,0} + \boldsymbol{\beta}_c^T\mathbf{x})}{\sum_{k=1}^{C} \exp(\beta_{k,0} + \boldsymbol{\beta}_k^T\mathbf{x})}.$$

Adopting the notation $\boldsymbol{\beta}_1 \doteq \boldsymbol{\beta}$ for the 2-class case, the model parameters are learned from the training data by maximizing the ℓ_1-penalized log-likelihood

$$\sum_{i=1}^{N} \log \Pr(y_i \mid \mathbf{x}_i) - \lambda \sum_{c=1}^{L} ||\boldsymbol{\beta}_c||_1$$

where $L = C$ for the multiclass case and $L = 1$ for the 2-class case.

Although the penalized log-likelihood function is not differentiable everywhere, several approximate algorithms exist for the minimization task [4, 5, 21] and the implementation of this paper uses the GLMNET algorithm [5].

2.2. The Support Vector Machine

Support vector machines (SVM) are widely used due to their maximum margin property. The binary SVM with the linear kernel solves the following problem:

$$\min_{\boldsymbol{\beta},\beta_0,\xi} \left(\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + C^* \sum_{i=1}^{l} \xi_i \right) \text{ such that } \begin{cases} y_i(\boldsymbol{\beta}^T\mathbf{x}_i + \beta_0) \geq 1 - \xi_i, \\ \xi_i \geq 0, \text{ for } i = 1, \ldots, N. \end{cases}$$

where $C^* \in \mathbb{R}$ is the upper bound. Alternatively, the above constrained minimization problem can written in a form emphasizing the regularization [7]

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{N} [1 - y_i(\beta_0 + \boldsymbol{\beta}^T\mathbf{x}_i)]_+ + \lambda||\boldsymbol{\beta}||_2^2,$$

where $[x]_+ = \max(0, x)$ and $\lambda = 1/(2C^*)$. In our work, we use the LIBSVM implementation of the SVM and extend it into the multiclass case using a one-against-one strategy [22].

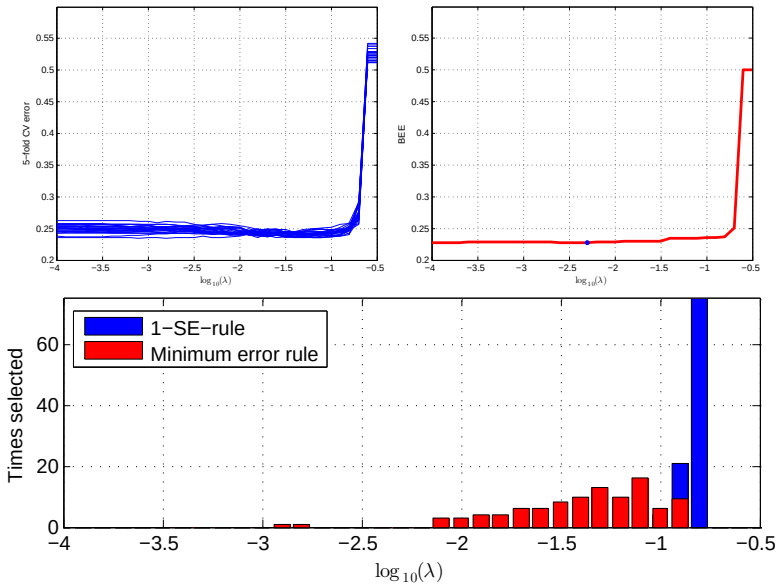


Figure 1: Top left: Examples of regularization path error curves of 5-fold cross-validation for a fixed 20-dimensional data with two classes. Top right: The corresponding BEE curve (blue dot at the minimum). Bottom: The histogram of the minimum location for 500 CV error curves.

2.3. Model selection

With both logistic regression and SVM classifiers, the parameter $\lambda > 0$ controls the strength of the regularization. Different choices of $\lambda \in \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ produce different classifiers and we denote the parameters as $\beta_0(\lambda)$, $\boldsymbol{\beta}(\lambda)$, $\beta_{c,0}(\lambda)$ and $\boldsymbol{\beta}_c(\lambda)$ to make this fact apparent when needed.

The best value of λ is traditionally selected by cross-validation: Either as the minimum of the cross-validation error curve [22] or as the largest λ whose error is within one standard deviation from the minimum [5]. The latter rule favors slightly sparser solutions and tends to decrease the generalization error. However, in our earlier work [20] choosing the minimum CV error solution resulted in more accurate prediction, so from now on we will focus on the minimum of the CV error as the selection rule.

Figure 1 illustrates the model selection using different error estimators.

Examples of error curves for different values of the regularization parameter λ with the logistic regression model are shown in Figure 1. In this example, a 20-dimensional toy dataset with altogether 500 normally distributed samples drawn from two classes was generated. The errors for models with $\log_{10}(\lambda) \in \{-0.5, -0.6, \ldots, -3.9, -4.0\}$ were estimated using 5-fold CV (top left) and the BEE (top right). There is significant variation between resulting error curves as shown in Figure 1 (top left) and there is even more significant variation between the location of the minima of the curves, as seen in Figure 1 (bottom). In particular, note that there are two isolated cases where the minima are far from the majority of cases, with $\lambda = 10^{-2.8}$ and $\lambda = 10^{-2.9}$.

In comparison, the Bayesian error estimator produces a single error curve, as shown in Figure 1 (top right). Although averaging a pool of CV curves would converge to a similar result, the BEE curve can be computed several orders of magnitude faster. In the experimental section, we will show that the location of the minimum is in many cases more accurate and in all cases significantly faster to compute than the CV.

3. Bayesian Minimum Mean-Square Error Estimator for Classification Error

A Bayesian approach to error estimation was recently introduced in the context of discrete classifiers [18] and linear classifiers [19]. We will first briefly review the latter case for binary classifiers. After this we propose an extension of BEE for model selection in multinomial classification problems.

3.1. Binary Classification

Consider a linear classifier with coefficients $\boldsymbol{\beta}(\lambda)$ and intercept $\beta_0(\lambda)$. For notational simplicity, we omit the reference to $\boldsymbol{\beta}$ and β_0 unless explicitly needed, and will be indexing the classifiers by the regularization parameter λ directly.

Assume that γ is the known prior probability of class -1 and let ε_c denote the error contributed by all samples from class $c \in \{-1, 1\}$. Then, the true classification error can be decomposed as $\varepsilon = \gamma\varepsilon_{-1} + (1-\gamma)\varepsilon_1$. The Bayesian error

estimator (BEE) is defined as the minimum mean squared estimator (MMSE) minimizing the expectation between the error estimate and the true error. Assuming that the true classification error ε is a function of the sample $(\mathbf{X}, \mathbf{y})$ and the parameters describing the feature-label joint distribution $\boldsymbol{\theta}_c$ for $c \in \{-1, 1\}$ are independent random variables prior to observing the data, this can be written as [19]:

$$\text{BEE} = E[\varepsilon \mid \mathbf{X}, \mathbf{y}] = \gamma E[\varepsilon_{-1} \mid \mathbf{X}, \mathbf{y}] + (1 - \gamma)\, E[\varepsilon_1 \mid \mathbf{X}, \mathbf{y}], \tag{1}$$

where

$$E[\varepsilon_c \mid \mathbf{X}, \mathbf{y}] = \int \varepsilon_c(\boldsymbol{\theta}_c) p_c(\boldsymbol{\theta}_c \mid \mathbf{X}, \mathbf{y})\, d\boldsymbol{\theta}_c \propto \int \varepsilon_c(\boldsymbol{\theta}_c) p_c(\boldsymbol{\theta}_c) \prod_{i:y_i=c} p_c(\mathbf{x}_i \mid \boldsymbol{\theta}_c)\, d\boldsymbol{\theta}_c, \tag{2}$$

for $c \in \{-1, 1\}$. Above, $\varepsilon_c(\boldsymbol{\theta}_c)$ is the true classification error given parameters $\boldsymbol{\theta}_c$ for the feature-label joint distribution and a fixed classifier, and $p_c(\boldsymbol{\theta}_c)$ is the prior for the parameters of the feature-label joint distribution for class c.

In order to evaluate the integral (2), we assume Gaussian class-conditional densities $p_c(\mathbf{x}_i \mid \boldsymbol{\theta}_c)$ for the data. The parameters of the Gaussian model describe the feature-label distribution and are denoted by $\boldsymbol{\theta}_c = (\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ for the class $c \in \{-1, 1\}$. The choice of the prior and its parameters is important to any Bayesian approach. Following Dalton and Dougherty [19] we assume inverse-Wishart priors for $\boldsymbol{\theta}_c$,

$$\begin{aligned} p_c(\boldsymbol{\theta}_c) \propto \det(\boldsymbol{\Sigma}_c)^{-(\kappa+P+1)/2} \exp\left(-\frac{1}{2}\text{trace}\left(\mathbf{S}\boldsymbol{\Sigma}_c^{-1}\right)\right) \\ \times \det(\boldsymbol{\Sigma}_c)^{-(1/2)} \exp\left(-(\nu/2)(\boldsymbol{\mu}_c - \mathbf{m})^T \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{\mu}_c - \mathbf{m})\right), \end{aligned}$$

with hyperparameters $\nu \in \mathbb{R}, \kappa \in \mathbb{R}, \mathbf{S} \in \mathbb{R}^{P \times P}, \mathbf{m} \in \mathbb{R}^P$. For an in-depth discussion on the role of these hyperparameters, see [19].

We limit our consideration to two specific choices related to the shape of the covariance matrix: 1) scaled identity covariances (i.e., $\boldsymbol{\Sigma}_c = \sigma_c \mathbf{I}$; Section 3.1.1) and 2) general covariances ($\boldsymbol{\Sigma}_c$ are general symmetric positive definite matrices; Section 3.1.2). In the following, we favor simplicity in the choice of hyperpa-

rameters, which are explained in more detail for the two different covariance models below.

3.1.1. Improper Prior— Scaled Identity Covariance

If the covariance matrices $\mathbf{\Sigma}_c$ are assumed to be scaled identity matrices $\sigma_c \mathbf{I}$, the non-informative Jeffrey's prior results by setting $\nu = 0$ and $\kappa = 0$. Note, that the parameters $\mathbf{m}$ and $\mathbf{S}$ need not be defined, because the prior is non-informative. The rationale for non-informative prior is that we regularize the covariance matrices by enforcing them to be of simple shape and do not need strong priors in addition to that. We will abbreviate the resulting error estimator as BEEi.

Denoting the number of samples in class $c \in \{-1, 1\}$ by N_c, the BEEi estimate for class c given in Eq. (2) has a closed form solution given by [19]

$$E[\varepsilon_c \mid \mathbf{X}, \mathbf{y}] = \frac{1}{2} + \frac{\text{sign}(A_c(\lambda))}{2} I\left(\frac{A_c(\lambda)^2}{A_c(\lambda)^2 + (N_c - 1)\text{trace}(\hat{\mathbf{\Sigma}}_c(\lambda))}; \frac{1}{2}, \alpha\right),$$

where $I(x; a, b)$ is the regularized incomplete Beta function and

$$\alpha = \frac{P(\lambda)(N_c + P(\lambda) + 1)}{2} - 1 \quad \text{and} \quad A_c(\lambda) = \frac{-y g_\lambda(\hat{\boldsymbol{\mu}}_c(\lambda))}{||\boldsymbol{\beta}(\lambda)||} \sqrt{\frac{N_c}{N_c + 1}},$$

with $P(\lambda)$ denotes the number of features with non-zero β_i and $\hat{\boldsymbol{\mu}}_c(\lambda)$ and $\hat{\mathbf{\Sigma}}_c(\lambda)$ the sample mean and covariance of class c; both depending on λ since the dimensions with zero coefficients β_i are discarded in their computation.

3.1.2. Proper Prior—General Covariance

To assume general covariance matrices, we need to select a proper prior to handle situations where $N_c < P$. We select $\kappa = P + 2$, $\nu = 0.5$, $\mathbf{S} = \mathbf{I}$, and $\mathbf{m} = 0$. The intuitive meaning of the hyperparameters is that $\mathbf{m}$ and $\mathbf{S}$ act as the most likely targets for the mean and covariance of the distribution, while ν and κ control how much the prior penalizes variability from $\mathbf{m}$ and $\mathbf{S}$, respectively.

With these choices, we define zero mean and identity covariance as the most likely model for the data. The value of κ has a key role in the model, as it acts as a regularizer for the covariance matrix estimate. Particular choices

include $\kappa = -(P+2)$ and $\kappa = 0$, which yield a flat prior and a Jeffrey's rule prior, respectively [19]. Our experiments indicate that a strong regularization is beneficial for model selection, and we choose κ as the negative of that of the flat prior $\kappa = P + 2$. Namely, a small κ may result in an ambiguous solution when number of samples is small. In particular, the flat prior easily becomes underdetermined with few samples.

These hyperparameters lead to a closed form solution [19]:

$$E[\varepsilon_c \mid \mathbf{X}, \mathbf{y}] = \frac{1}{2} + \frac{\text{sign}(A_c(\lambda))}{2} I\left(\frac{A_c(\lambda)^2}{A_c(\lambda)^2 + \boldsymbol{\beta}(\lambda)^T \mathbf{S}_c \boldsymbol{\beta}(\lambda)}; \frac{1}{2}, \frac{N_c+3}{2}\right),$$

where

$$A_c(\lambda) = -y g_\lambda(\mathbf{m}_c(\lambda)) \sqrt{(0.5+N_c)/(1.5+N_c)}$$

and

$$\mathbf{m}_c(\lambda) = \frac{\hat{\boldsymbol{\mu}}_c(\lambda) N_c}{N_c + 0.5} \quad \text{and} \quad \mathbf{S}_c(\lambda) = (N_c - 1)\hat{\boldsymbol{\Sigma}}_c(\lambda) + \mathbf{I}_{N_c} + \frac{0.5 N_c}{N_c + 0.5} \hat{\boldsymbol{\mu}}_c \hat{\boldsymbol{\mu}}_c^T.$$

From now on, we will abbreviate the resulting error estimator as BEEp. Note that all the parameter values are kept fixed throughout the experiments.

3.2. Multinomial Classification

The above error estimators apply only to binary classifiers. In order to extend the scope to multi-class problems we consider *pairwise error* instead of the classification error. The pairwise error is defined as

$$\varepsilon_{\text{pw}} = \frac{1}{N} \sum_{k=1}^{C-1} \sum_{m=k+1}^{C} |\mathbf{y}_{\{k,m\}}| \, \varepsilon(k,m), \tag{3}$$

where $\varepsilon(k,m)$ denotes (the estimate of) the classification error for the binary problem involving only classes k and m, and $|\mathbf{y}_{\{k,m\}}|$ is the number of samples in classes k and m. In other words, we average the number of erroneous classifications within all pairs of classes k and m. Using this concept of pairwise error, we can extend any binary error estimator for multiclass classification problems in a similar manner as the one-vs-one strategy can be used to generalize a binary classifier into a multiclass classifier. Within the scope of this paper, we

Table 1: Summary of the studied datasets. The Ovarian cancer and the Yeast datasets are not split to training and test sets, and the split is randomized in the experiments leaving a part of the data for testing.

Dataset	*Training Samples*	*Test Samples*	*Features*	*Classes*	*Fixed Split*
MEG Binary	727	653	408	2	Yes
Ovarian	216	N/A	15000	2	No
Adult	1605	30956	123	2	Yes
MEG	727	653	408	5	Yes
Yeast	1484	N/A	8	10	No
Satellite	4435	2000	36	6	Yes

substitute the general binary error estimator $\varepsilon(k, m)$ of Eq. (3) by the Bayesian Error Estimator of Eq. (1).

The pairwise error ε_{pw} is closely related to the classification error: It is always greater than the genuine multiclass error ε and upper bounds can be derived for the difference between the two [23]. As our main goal is not to estimate the performance of the classifier in absolute terms, but to only compare their performances, the positive bias of the pairwise error is not a problem for the model selection. The model selection is based on the minimum of the error estimate, and the actual values of the error estimates are not important.

4. Experimental Results

In this section we consider the accuracy of the proposed error estimation method for model selection. The section consists of eight experiments with different data sets. There are two experiments with synthetic data (one for binary and one for multi-class case), and six experiments for real-world data (three binary and three multinomial cases). The properties of the real-world datasets are summarized in Table 1. Note that the priors and their parameters

for BEEi and BEEp are always as described in Section 3 and we do not alter them between different experiments. This way we avoid unfair hand-tuning of the error criterion to match the data.

The BEE model selection is compared to 5-fold cross-validation and the EBIC criterion. As our hypothesis is that the BEE is most efficient with small samples, we will investigate the accuracy as a function of the number of samples by randomly subsampling the training set. In addition to the CV-5, we also experimented with the 10-fold and leave-one-out CV but no significant difference to CV-5 was observed (Table 2 shows few example results). For further discussion on the effect of the number of folds, see [24].

Two of the eight experiments also investigate the effect of repeating the CV-5 procedure in order to decrease the variance of the error estimates. Namely, one may argue that the inaccuracies of the counting based approaches may be avoided by averaging multiple estimates together. Since this is a computationally costly approach, we limit to two cases where the number of samples is smallest (Sections 4.1.3 and 4.2.2 and consider only the faster GLMNET-algorithm based LR classifier.

4.1. Binary Classification

4.1.1. Synthetic Data

The first experiment studies a synthetic 20-dimensional dataset similar to that of [25]. This data consists of two normally distributed classes with different means and a common, non-diagonal covariance matrix. 14 of the features have discriminative power and the number of non-informative features is varied from 6 (as in [25]) to 250. With the SVM, we consider only the case of 6 non-informative features.

The data is designed so that it is challenging for feature ranking or greedy feature selection methods. We use the conditional error rate as the performance criterion, *i.e.*, we estimate the probability of misclassification for a classifier trained on a given training sample [26, 27]. The experiment was run 100 times for each sample size $N = 28, 50, 100, 200$ generating a new training sample each

Table 2: Average conditional error rates with 2-class synthetic data and the LR classifier. Statistically significant improvements ($p < 0.01$) of BEE over CV-5 are highlighted in boldface. The cases where CV-5 outperforms either BEE rule with statistical significance ($p < 0.01$) are highlighted in italics. The significance is tested with a paired permutation test over the 100 resamplings (one-sided). Note that if CV-5 value is in italics and BEEp value in boldface, it means that CV-5 outperformed BEEi but BEEp outperformed CV-5. Average conditional error rates with CV-10 are shown for reference.

Noise Features	*6*	*12*	*18*	*24*	*30*	*50*	*100*	*150*	*250*
					N = 28				
CV-5	0.146	0.152	0.156	0.164	0.163	0.172	0.183	0.189	0.208
CV-10	0.147	0.151	0.158	0.161	0.159	0.168	0.184	0.191	0.208
BEEp	**0.132**	**0.143**	**0.132**	**0.143**	**0.148**	**0.157**	0.176	0.187	0.207
BEEi	**0.127**	**0.134**	**0.130**	**0.137**	**0.145**	**0.152**	**0.173**	0.184	0.205
EBIC	0.137	0.147	0.146	0.166	0.187	0.254	0.355	0.387	0.437
					N = 50				
CV-5	0.104	0.110	0.108	0.116	0.112	0.123	0.130	0.123	0.136
CV-10	0.102	0.111	0.108	0.114	0.111	0.118	0.131	0.126	0.135
BEEp	**0.092**	**0.099**	0.110	0.113	0.117	**0.113**	0.126	0.124	0.137
BEEi	**0.094**	**0.097**	**0.100**	**0.106**	0.111	**0.110**	0.124	0.123	0.136
EBIC	0.101	0.109	0.117	0.118	0.121	0.123	0.173	0.219	0.258
					N = 100				
CV-5	*0.067*	*0.068*	0.073	0.074	0.075	0.080	0.084	0.091	0.088
CV-10	0.066	0.068	0.072	0.074	0.074	0.079	0.083	0.091	0.088
BEEp	**0.057**	**0.061**	**0.065**	**0.069**	**0.072**	0.082	0.084	0.090	0.089
BEEi	0.072	0.072	0.073	0.075	0.076	0.080	0.085	0.090	0.089
EBIC	0.068	0.071	0.076	0.078	0.080	0.086	0.082	0.086	0.085
					N = 200				
CV-5	*0.042*	*0.046*	*0.050*	*0.047*	*0.051*	*0.054*	*0.060*	*0.060*	0.067
CV-10	0.042	0.046	0.049	0.047	0.051	0.053	0.058	0.059	0.065
BEEp	**0.038**	**0.042**	**0.045**	**0.044**	**0.048**	0.053	0.063	0.065	0.062
BEEi	0.062	0.062	0.062	0.062	0.062	0.061	0.063	0.065	0.066
EBIC	0.050	0.053	0.058	0.057	0.060	0.060	0.066	0.066	0.063

Table 3: Average conditional error rates and standard deviations of error rates with 2-class synthetic data and the SVM classifier.The number of noise features was 6. Statistically significant improvements ($p < 0.01$) of BEE over CV-5 are highlighted in boldface. The cases where CV-5 outperforms either BEE rule with statistical significance ($p < 0.01$) are highlighted in italics. The significance is tested with a paired permutation test over the 100 resamplings (one-sided).

N	*28*	*50*	*100*	*200*
CV-5	*0.092* $\pm$ 0.025	*0.075* $\pm$ 0.016	0.063 $\pm$ 0.009	*0.053* $\pm$ 0.008
BEEi	**0.083** $\pm$ 0.015	**0.068** $\pm$ 0.009	**0.059** $\pm$ 0.005	0.057 $\pm$ 0.004
BEEp	0.106 $\pm$ 0.027	0.081 $\pm$ 0.016	0.064 $\pm$ 0.008	**0.048** $\pm$ 0.007

time. The classes were balanced such that there were $14, 25, 50, 100$ training samples per class. The optimal Bayes rate for the problem is 0.023 and the results are shown in Tables 2 and 3. In addition to CV-5 and the two flavors of BEE, Table 2 tabulates the error rates of classifiers selected using Extended Bayesian Information Criterion (EBIC) with the default parameter value (the only parameter $\gamma = 0.5$) as proposed by [12]. We also compute p-values for the significance of the difference in the average accuracies between the selection criteria. To this aim, we used a permutation test with 100 randomly sampled training sets. The average conditional error rates resulting from CV-10 model selection are also shown for reference. As can be seen in Table 2, the average conditional error rates are very similar between CV-5 and CV-10.

One can conclude that with small sample sizes ($N = 28$, 50), the BEE rules are significantly more accurate than either CV-5, CV-10 or the EBIC criteria as long as the number of noise features is not too high. When the sample size is roughly 4 times the number of relevant features per class ($N = 100$, 200) the BEEp rule still has the best performance, but CV-5, CV-10 and the EBIC are now superior to the BEEi. This is not surprising, since the data generation model is in agreement with the assumptions of the BEEp, but not with the assumptions of the BEEi as the features are not independent.

Another aspect of interest is the variability of conditional error rates. We calculated the standard deviations of the classification errors over the 100 iterations. The results are omitted for brevity, but in summary the BEEi criterion results in smallest deviation in all cases, with the second smallest variance given by the BEEp criterion.

Table 3 summarizes the results with the SVM classifier. In this case, the BEEi rule is superior to CV-5 with the three smallest sample sizes, but with the largest sample size, CV-5 becomes more accurate. On the other hand, the BEE with the proper prior is better than CV-5 with the largest sample size and worse with the two smallest sample sizes. This happens probably due to the difficulty in estimating full covariance matrices with small number of samples. The standard deviations of error values are always smallest with the BEEi rule, similarly to the logistic regression case.

4.1.2. MEG Binary Data

As the first real-world experiment, we consider the data from *MEG mind reading competition* of the ICANN 2011 conference [28] using the features extracted by the winning method of the competition [29]. The challenge was to predict which video was shown to the test subject based on MEG brain measurements. The task is originally a 5-class problem, but the organizers also considered a binary task: Separating videos with a plot from those without one. Here, the data represents a case where the dimensionality is not particularly high and $N \approx P$ with $N = 727$ and $P = 408$. To study the effect of sample size, we subsampled the training data randomly with different decimation factors. Thus, we used 10%, 15%, 20%, ..., 100% of the training data. The subsampling was repeated 100 times for each case and the errors were averaged.

The results are summarized in Figure 2 (top), with the test errors shown for logistic regression (left) and SVM (right) classifiers. The plots show the prediction error when the model has been selected using CV-5 and BEE criteria, with proper and improper priors. For reference, the plot also shows the error for the best model selected using the actual test error (termed "Oracle").

For both classifiers, the BEE with proper prior results in the most accurate classification. However, as the number of samples is increased, the difference to the CV-5 approach decreases. This is what one would expect: As more samples are added, the benefit of the prior becomes less important. It can also be seen that the accuracy of the BEE with improper prior is not very good. In particular, the BEEi is not good in selecting the model along the SVM regularization path. Also the choice of an appropriate LR model from the regularization path is less accurate than with a proper prior; although mostly better than with CV-5.

Our interpretation for the poor performance of the BEEi with the SVM is that the underlying assumption of diagonal covariance matrices is not valid as the features are measurements from MEG sensors, and nearby sensors are known to correlate. Moreover, the correlation degrades the performance of the BEEi clearly less in conjunction with the LR classifier because the L1-regularization is able to select uncorrelated subsets of features.

4.1.3. Ovarian Cancer Data

As the third experiment, we study the performance for the *Ovarian cancer data* [30] widely used to benchmark binary classifiers. The data is generated by protein mass spectrometry, using the WCX2 protein array. The data consists of measurements from 121 ovarian cancer patients and 95 healthy controls. In all, there are 15000 mass spectrometry features of which less than a few dozen are helpful for classification. As such, the data is high-dimensional and $N \ll P$ with $N = 216$ and $P = 15000$. For this experiment, we also compare the proposed model selection criteria with the CV-5 and 100 times repeated CV-5. More specifically, the latter error estimate is obtained as the average of hundred 5-fold cross-validation experiments initialized with different random seeds.

The results are shown in Figure 2 (middle) for the LR (left) and SVM (right) classifiers. In this case, the BEEp criterion is the most accurate in both plots. Also, the difference to CV-5 is again most significant with a small number of samples. In the Logistic Regression case, the accuracy of the 100-times repeated CV-5 reaches that of the two BEE criteria as the number of samples increases.

Nevertheless, the performance with small sample size is inferior to the BEE, and the computational cost is about 400 times higher.

One interesting observation is that the performance of the BEEi is equal to that of the BEEp criterion for the LR classifier, but inferior for the SVM classifier. The reason for the poor performance of BEEi is the same as in Section 4.1.2: The 15000 spectrogram features are correlated, and the LR selects a non-correlated subset of them.

4.1.4. Adult Data

The *adult dataset* has become a standard benchmark in particular within the SVM research community. The data originates from a 1994 census database of US adult citizens, and the task is to predict whether the annual income of a household exceeds \$50 000 based on a set of attributes. The data is preprocessed as in [31] such that there are altogether 123 sparse binary features. Moreover, there are 1605 training samples and 30956 test samples, so this data represents a case with $N \gg P$.

The results are shown in Figure 2 (bottom), with logistic regression classifier (left) the SVM (right). For this data set, neither of the BEE implementations is clearly better than the CV. The model selection performance is somewhat better with the SVM, but slightly worse with the LR, although the estimation accuracy improves as more samples are added to the training set.

The poor performance with small number of samples is unexpected, but explained by a significant deviation from the Gaussianity assumption. As mentioned, the features are all binary, and the model assumptions are not correct. The degradation due to the incorrect prior is mitigated as the sample size increases, and the accuracy increases to the level of the CV-5 criterion, eventually surpassing it.

Table 4: Average conditional error rates and their standard error with 4-class synthetic data. N_i is the number of samples per class. Statistically significant differences between the BEE and the CV-5 rule are highlighted in boldface ($p < 0.01$). Moreover, the cases where CV-5 outperforms either BEE rule with statistical significance ($p < 0.01$) are highlighted in italics. The significance is tested with a paired permutation test over 100 resamplings (one-sided).

N_i	*15*	*25*	*50*	*100*
LR CV-5	0.215 ± 0.029	0.164 ± 0.024	*0.111* ± 0.012	*0.083* ± 0.006
LR BEEi	0.210 ± 0.023	0.167 ± 0.014	0.142 ± 0.009	0.134 ± 0.006
LR BEEp	0.212 ± 0.029	**0.156** ± 0.019	**0.106** ± 0.010	**0.080** ± 0.005
SVM CV-5	*0.196* ± 0.022	*0.159* ± 0.017	0.133 ± 0.008	*0.110* ± 0.006
SVM BEEi	**0.180** ± 0.015	**0.151** ± 0.008	0.133 ± 0.004	0.125 ± 0.005
SVM BEEp	0.215± 0.022	0.171 ± 0.016	0.136 ± 0.007	**0.106** ± 0.006

4.2. Multinomial Classification

4.2.1. Synthetic Data

The synthetic dataset is analogous to the one used in the binary classification experiments. It models a four-class problem, where the classes are normally distributed with the same covariance matrix but with different means. 15 of the 21 features are informative and 6 are non-informative. The covariance matrix is similar to the one described for the binary classification experiment. The means are intentionally located so that the problem would be as hard as possible for the pairwise error based error estimation, namely so that the difference between the classification error and the pairwise error would be maximal.

For a 4-class problem this can be achieved by dividing the informative features to the groups of 3 similar features (here 5 groups), and placing the means of the groups of features at the four vertices of the unit 3-simplex (tetrahedron). The code to generate the dataset is available at the supplementary site for this paper[2], where also a more detailed description of the dataset can be

[2] `https://sites.google.com/site/bayesianerrorestimate/`

found. Again, the conditional error rate across 100 resamplings is used as the performance criterion. The conditional error rate was computed with the Monte Carlo integration with 5 million samples simulated per class.

The results are shown in Table 4. The results show no qualitative difference to the 2-class case. BEE with the improper prior is the most accurate method with the SVM and small sample sizes. With the larger sample sizes, the lowest error rates are obtained by the BEE rule with the proper prior, both with logistic regression (when $N_i \geq 25$) and SVM (when $N_i = 100$). As can be noted from Table 4, the standard deviations of the conditional errors of the BEE based classifiers are smaller than their CV-5 based counterparts.

4.2.2. MEG Data

As our second experiment, we consider the *MEG mind reading data* of the ICANN 2011 conference [28], as in Section 3.1, except now treated as a five-class problem (5 movie types) instead of the binary one (plot-vs-no plot). In this case, we also compare the BEE criteria with 100-times repeated 5-fold CV.

The results are summarized in Figure 3 (top). As one can expect, the results are similar to the binary case: The BEE error criterion is able to produce either similar or more accurate results for all training set sizes. The exception is again the model selection of an SVM classifier (right), when improper prior is assumed. Once again, this is due to correlated features of the MEG data, which makes the diagonal covariance assumption invalid.

The BEEp is superior to the BEEi with the LR classifier, as well. However, the difference appears only when the number of samples increases and the invalidity of the diagonal covariance assumption becomes significant and the gain given by the strong prior becomes marginal. We can also see that the repeated cross-validation is not helpful in this case: The accuracy is close to that given by a single CV-5 iteration.

4.2.3. Yeast Data

Our third experiment considers the widely used *Yeast* dataset [32]. The data consists of the cellular location sites in yeast cells together with 8 explanatory variables. Altogether, there are 1484 samples and 10 categories in this set.

Figure 3 (middle) summarizes the results for the two classifiers: LR (left) and SVM (right). For this data, both BEE rules are superior to the cross-validation based model selection, with the BEEp slightly more accurate than the BEEi with the SVM. One can see, however, that the difference of BEE and CV decreases as more training data is used. Moreover, the difference between the optimal model (green curve) and the selected model increases as the amount of training data is increased. This is due to the decrease of test set size: As there are only a few samples in the test data, it becomes less and less representative of the true underlying distribution, and the discrepancy between the test error and the training/true error starts to increase.

4.2.4. Satellite Data

As our final experiment, we consider the *Satellite* data set[3]. In this dataset the goal is to predict the soil type using multispectral pixel values in a satellite image. The data is divided into predetermined training and test sets, consisting of 4435 and 2000 samples, respectively. The data has of six classes and has 36 explanatory variables originating from 3×3 pixel neighborhoods from four spectral bands.

Figure 3 (bottom) illustrates the results for this experiment. For this data, the features are highly correlated, and the assumption of independent features is not valid resulting in a poor model selection accuracy of the BEEi estimator. This is the case also for the logistic regression, which was earlier more tolerant than the SVM to correlated variables. Unlike the earlier experiments, here all variables are correlated, and the LR classifier can not select uncorrelated features.

[3] `http://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)`

Nevertheless, the BEEp estimator has a performance that is comparable to the CV estimator, with the BEEp superior for small number of training samples and the CV-5 superior when the amount of training data increases. This is quite natural, since the prior assumptions of the Bayesian approach become obsolete and even harmful as the number of data becomes large enough. It is noteworthy, that the satellite dataset has the largest number of samples among all our experiments, and the benefit of the Bayesian approach is the smallest here.

4.3. Computational Complexity

A key reason to use the Bayesian error estimator instead of counting based approaches is its computational efficiency. The advantages are rather obvious: The BEE calculates the error estimate directly from the training set after the classifier has been trained. Thus, it is obtained as a side product of the full model training step.

The experimental results are summarized in Figure 4, where the LR and SVM classifiers are trained with the MEG dataset (see Section 4.2.2) with 5 classes. The red curve shows the time complexity (in seconds) for the BEE method and the blue curve that for 5-fold cross-validation for the LR (solid) and SVM (dashed) classifiers. In general, the CV-5 algorithm requires 3 – 3.5 times more computation than the BEE-based model selection. The actual time used by the CV-5 approach is slightly less than the number of folds because the folds are trained with less data, and because the BEE estimate also requires some computation, which becomes more significant with small data sizes. Moreover, as the data consists of 5 classes, the BEE estimate is in fact evaluated for all pairs of classes, altogether $\binom{5}{2} = 10$ times.

5. Conclusion

In this paper, we proposed using a Bayesian error estimator (BEE) for the selection of the best classification model along the regularization path for SVM

and regularized logistic regression classifiers. Also, we proposed a multinomial extension of the Bayesian error estimator, and studied its efficiency in model selection for linear classifiers. The model selection by the new Bayesian error estimator was experimentally shown to improve the classification accuracy in small sample-size situations.

The BEE produces a single deterministic error curve over the regularization path, as opposed to cross-validation based approaches, whose error depends on the particular random split of the data into folds. The BEE is also significantly faster to calculate, because the estimate is obtained directly from the training data; without iterated training over folds.

The experiments show that the BEE criteria, on average, select better classification models along regularization path than CV-5, in particular with small number of samples. The exception to this are the cases, where the assumptions of the BEE do not hold. In particular, BEE with the improper prior assumes identity covariances, and the BEEi is not very tolerant to deviations from this assumption. However, the BEEp is not limited by the independence assumption and is the recommended model selection criterion of the two BEE rules. The BEE rules assume the Gaussianity of the data. However, they are not especially sensitive to the Gaussianity assumption for the model selection purposes and BEEp is less accurate than CV-5 only when the distribution of the features departs markedly from the Gaussian assumption as in Adult dataset.

Model selection using the BEE criterion is an important addition to a practitioner's toolbox in the area of machine learning. This is particularly important when experimenting with classification frameworks, where a large number of iterations is required and training speed becomes critical.

Acknowledgments

This project has received funding from the Universidad Carlos III de Madrid, the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement nr 600371, el Ministerio

de Economa y Competitividad (COFUND2013-40258) and Banco Santander.

References

[1] Y. Saeys, I. Inza, P. Larraaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507 – 17.

[2] F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: a tutorial overview, NeuroImage 45 (Suppl 1) (2009) S199 – S209.

[3] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. R. Stat. Soc., Series B 58 (1994) 267–288.

[4] B. Krishnapuram, L. Carin, M. A. Figueiredo, A. J. Hartemink, Sparse multinomial logistic regression: Fast algorithms and generalization bounds, IEEE Trans Patt Anal Mach Intell 27 (6) (2005) 957–968.

[5] J. H. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (1) (2010) 1–22.

[6] V. Vapnik, The nature of statistical learning theory, Springer, 2000.

[7] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, Springer, 2009.

[8] E. R. Dougherty, C. Sima, B. Hanczar, U. M. Braga-Neto, Performance of error estimators for classification, Current Bioinformatics 5 (1) (2010) 53.

[9] N. Glick, Additive estimators for probabilities of correct classification, Pattern recognition 10 (3) (1978) 211–222.

[10] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: Selected Papers of Hirotugu Akaike, Springer, Budapest, 1998, pp. 199–213.

[11] G. Schwarz, Estimating the dimension of a model, The annals of statistics 6 (2) (1978) 461–464.

[12] J. Chen, Z. Chen, Extended BIC for small-n-large-P sparse GLM, Statistica Sinica 22 (2) (2012) 555.

[13] S. Demyanov, J. Bailey, K. Ramamohanarao, C. Leckie, AIC and BIC based approaches for SVM parameter value estimation with RBF kernels., Journal of Machine Learning Research 25 (2012) 97–112.

[14] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Machine learning 46 (1-3) (2002) 131–159.

[15] U. v. Luxburg, O. Bousquet, B. Schölkopf, A compression approach to support vector model selection, Journal of Machine Learning Research 5 (2004) 293–323.

[16] K. Duan, S. Keerthi, A. N. Poo, Evaluation of simple performance measures for tuning SVM hyperparameters, Neurocomputing 51 (2003) 41 – 59.

[17] U. Braga-Neto, E. Dougherty, Bolstered error estimation, Pattern Recognition 37 (6) (2004) 1267–1281.

[18] L. A. Dalton, E. R. Dougherty, Bayesian minimum mean-square error estimation for classification error—part I: Definition and the Bayesian MMSE error estimator for discrete classification, IEEE Trans. Signal Process 59 (1) (2011) 115–129.

[19] L. A. Dalton, E. R. Dougherty, Bayesian minimum mean-square error estimation for classification error—part II: The Bayesian MMSE error estimator for linear classification of Gaussian distributions, IEEE Trans. Signal Process 59 (1) (2011) 130–144.

[20] H. Huttunen, T. Manninen, J. Tohka, Bayesian error estimation and model selection in sparse logistic regression, in: 2013 IEEE International Work-

shop on Machine Learning for Signal Processing (MLSP), IEEE, 2013, pp. 1–6.

[21] G.-X. Yuan, C.-H. Ho, C.-J. Lin, An improved GLMNET for l1-regularized logistic regression, The Journal of Machine Learning Research 98888 (2012) 1999–2030.

[22] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3) (2011) 27.

[23] F. Garber, A. Djouadi, Bounds on the bayes classification error based on pairwise risk functions, IEEE Trans Pattern Anal Machine Intell 10 (2) (1988) 281–288.

[24] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: IJCAI, Vol. 14, 1995, pp. 1137–1145.

[25] P. Křížek, J. Kittler, V. Hlaváč, Improving stability of feature selection methods, in: Computer Analysis of Images and Patterns, Springer, 2007, pp. 929–936.

[26] S. J. Raudys, A. K. Jain, Small sample size effects in statistical pattern recognition: Recommendations for practitioners, IEEE Trans Patt Anal Mach Intell 13 (3) (1991) 252–264.

[27] G. J. McLachlan, The bias of the apparent error rate in discriminant analysis, Biometrika 63 (2) (1976) 239–244.

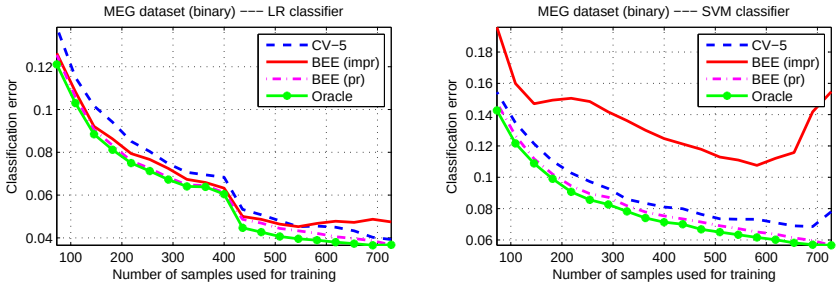
[28] A. Klami, P. Ramkumar, S. Virtanen, L. Parkkonen, R. Hari, S. Kaski, ICANN/PASCAL2 challenge: MEG mind-reading—overview and results (2011).
URL `http://www.cis.hut.fi/icann2011/meg/megicann_proceedings.pdf`

[29] H. Huttunen, T. Manninen, J.-P. Kauppi, J. Tohka, Mind reading with regularized multinomial logistic regression, Machine Vision and Applications 24 (6) (2013) 1311–1325.

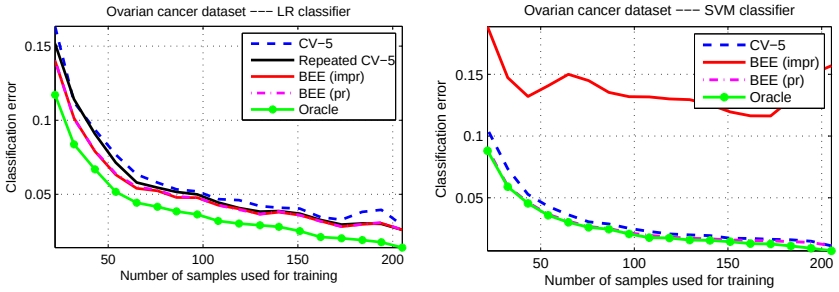
[30] T. P. Conrads, V. A. Fusaro, S. Ross, D. Johann, V. Rajapakse, B. A. Hitt, S. M. Steinberg, E. C. Kohn, D. A. Fishman, G. Whitely, High-resolution serum proteomic features for ovarian cancer detection., Endocrine-Related Cancer 11 (2) (2004) 163–178.

[31] J. C. Platt, Fast training of support vector machines using sequential minimal optimization, in: Advances in kernel methods: support vector learning, MIT press, 1999.

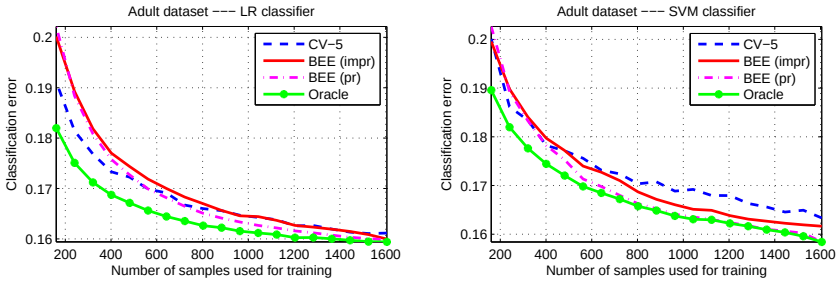
[32] P. Horton, K. Nakai, A probabilistic classification system for predicting the cellular localization sites of proteins, in: Proc Int Conf Intell Syst Mol Biol., Vol. 4, 1996, pp. 109–115.



(a) MEG binary data

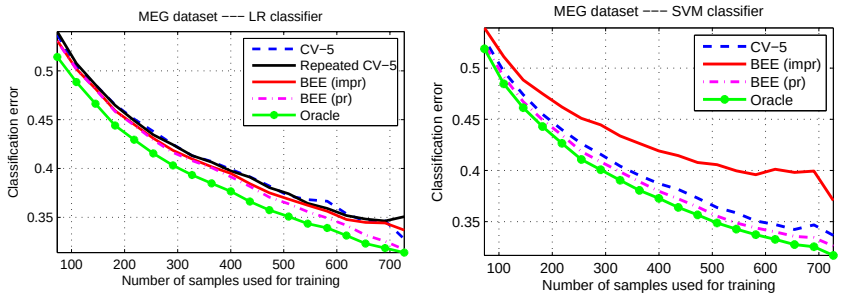


(b) Ovarian cancer data

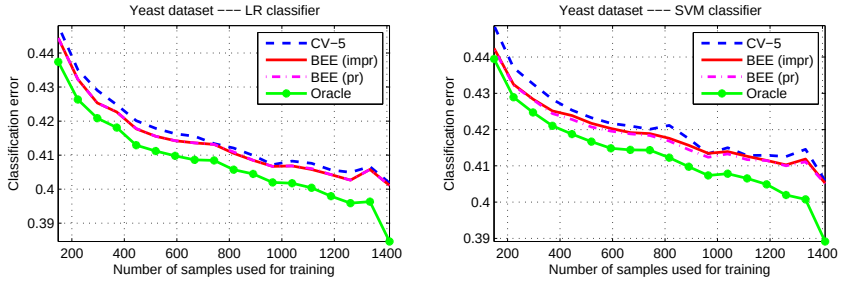


(c) Adult data

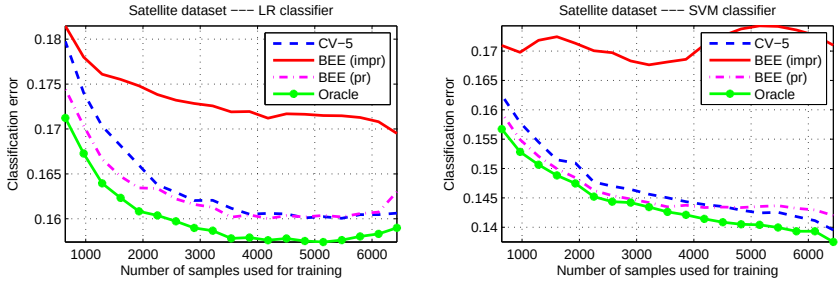
Figure 2: Classification errors for three binary classification problems as a function of the amount of data used for training. In each case, the top plot shows the results for logistic regression and the bottom figure for the SVM classifiers with model selected using CV-5 (dashed blue) and BEE error estimates with proper ($\sigma = 1$; dashed red) and improper (solid red) priors. For reference, also the error of the model minimizing the test error (termed *Oracle*) is shown in green. The middle-left plot includes also the results for the 100-times repeated CV-5 criterion (solid black line).



(a) MEG data



(b) Yeast data



(c) Satellite data

Figure 3: Classification errors for three multinomial classification problems as a function of the amount of data used for training. The top figure shows the results for logistic regression and the bottom figure for the SVM classifiers with model selected using CV-5 and BEE error estimates with proper ($\sigma = 1$) and improper priors. For reference, also the error of the model minimizing the test error (termed *Oracle*) is shown in green. The top-left plot includes also the results for the 100-times repeated CV-5 criterion (solid black line).

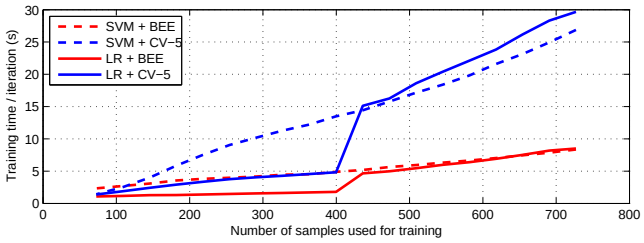


Figure 4: The time complexity of one run of the CV-5 λ selection algorithm (blue) and the BEE λ selection algorithm (red) for the LR classifier (solid line) and the SVM classifier (dashed curve). The curves are averaged over 100 iterations using the MEG data. The step is due to the switch between two coordinate descent algorithms within the glmnet package depending on the P/N ratio.