Variational Relevance Vector Machine for Tabular Data

Dmitry Kropotov DMITRY.KROPOTOV@GMAIL.COM
Dorodnicyn Computing Centre of the Russian Academy of Sciences
119333, Russia, Moscow, Vavilov str., 40

Dmitry Vetrov VETROVD@YANDEX.RU
Lomonosov Moscow State University
119992, Russia, Moscow, Leninskie Gory, 1, 2nd ed. bld., CMC department

Lior Wolf WOLF@CS.TAU.AC.IL
The Blavatnik School of Computer Science, The Tel-Aviv University
Schreiber Building, room 103, Tel Aviv University, P.O.B. 39040, Ramat Aviv, Tel Aviv 69978

Tal Hassner HASSNER@OPENU.AC.IL
Computer Science Division, The Open University of Israel
108 Ravutski Str. P.O.B. 808, Raanana 43107, Israel

Editor: Masashi Sugiyama and Qiang Yang

Abstract

We adopt the Relevance Vector Machine (RVM) framework to handle cases of table-structured data such as image blocks and image descriptors. This is achieved by coupling the regularization coefficients of rows and columns of features. We present two variants of this new gridRVM framework, based on the way in which the regularization coefficients of the rows and columns are combined. Appropriate variational optimization algorithms are derived for inference within this framework. The consequent reduction in the number of parameters from the product of the table's dimensions to the sum of its dimensions allows for better performance in the face of small training sets, resulting in improved resistance to overfitting, as well as providing better interpretation of results. These properties are demonstrated on synthetic data-sets as well as on a modern and challenging visual identification benchmark.

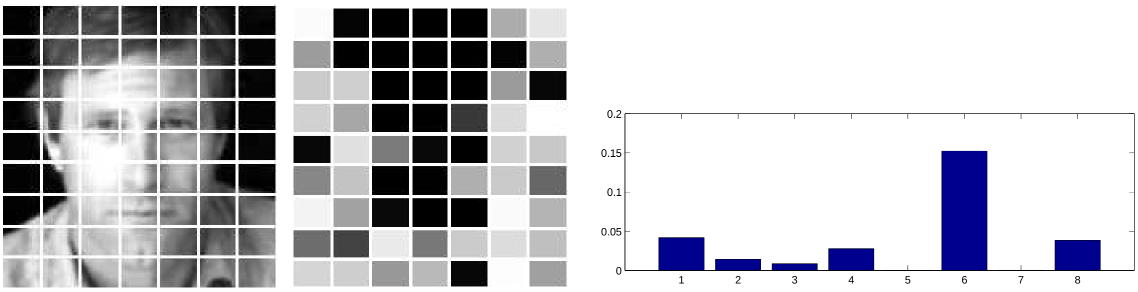
Keywords: Bayesian learning, variational inference, feature selection, Relevance Vector Machine, Automatic Relevance Determination, image classification

1. Introduction

Generalized linear models have been a popular approach to regression and classification problems for decades. Special attention is often paid to obtain sparse decision rules, where most of the assigned weights equal zero. Within a Bayesian framework the detection of relevant features can be done automatically by assigning an individual regularization coefficient to each weight. This process is called automatic relevance determination (ARD). The Relevance Vector Machine (RVM) is an important example of successful application of ARD to linear/logistic regression (see Tipping (2001)).

In this paper we generalize the RVM framework to the case of tabular data, i.e. cases where an object is described by a matrix of features. Tabular data arises in many domains

©2010 Dmitry Kropotov, Dmitry Vetrov, Lior Wolf, Tal Hassner.



(a) (b) (c)

Figure 1: The illustration of "grid" approach on the LFW data. Each image is split into 63 blocks (a) and for each block 8 descriptors are computed. GridRVM assigns individual regularization coefficients for each block and each descriptor. The relevance of blocks (the darker the more informative) is shown in (b) and the relevance of descriptors (inverse regularization coefficient) is shown in (c).

(see section 2). Of course, it is always possible to convert tables to feature vectors and run a standard learning algorithm. We will, however, show that this may sometimes lead to overfitting. Here, we suggest assigning individual regularization coefficients to each column and row of the table. The regularization coefficient of the feature in position ij in the table is then the result of the composition of the coefficients for the i^{th} row and the j^{th} column. We consider two variants of such compositions: product and summation of regularization coefficients, thus deriving p- and s-gridRVM models. Variational inference is used to obtain iterative equations for learning in these models. We demonstrate results on synthetic and real-world problems and show that the gridRVM approach prevents overfitting in case of small datasets. In particular we address the problem of same/not-same face classification in the Labeled Faces in the Wild (LFW) image set (see Huang et al. (2007)). We convert each image to a tabular presentation by computing a set of descriptors on distinct image blocks. GridRVM is then applied to find the most relevant blocks and descriptors (see fig. 1).

The rest of the paper is organized as follows. Motivation and related work are given in sections 2 and 3 respectively. Section 4 presents definitions of the p- and s-gridRVM models and establishes the notation used thereafter. Iterative learning algorithms based on variational inference are described in section 5. We conclude with experiments on illustrative and real-world problems in section 6.

2. Motivation

In classical machine learning theory, a training set consists of a number of objects (precedents) $X = \{\vec{x}_n\}_{n=1}^{N}$, each represented as a vector of features $\vec{x}_n = (x_n(1), \ldots, x_n(d)) \in \mathbb{R}^d$. It is assumed that there is no hierarchy in the space of features. This is not, however, always the optimal representation. In some cases, a tabular presentation is more convenient. Objects are then described by a number of features that form a table rather than a single vector.

A natural example of such case arises in a region/descriptor-based framework for image analysis. Within this framework, an image is split into several regions (blocks) and a set of descriptors is then computed for each region. Then, we may associate each feature with the pair region/descriptor and form a tabular view of a single image. Note that often the number of features extracted from single image exceeds the number of images in the whole training set, resulting in increased risk of overfitting.

Another example is related to the use of radial basis functions (RBF) in regression / classification algorithms. Traditionally, RBF depends only on the distance $\rho(\vec{x}, \vec{y_i})$ between the object $\vec{x}$ and some predefined point $\vec{y_i}$ in the space of features $\mathbb{R}^d$

$$\phi_i(\vec{x}) = f(\rho(\vec{x}, \vec{y_i})), \quad i \in \{1, \ldots, m\}.$$

Each object is described by a vector of m RBF values. Gaussian RBFs $\phi_i(\vec{x}) = \exp(-\delta\|\vec{x} - \vec{y_i}\|^2)$ are a popular "rule-of-thumb" choice in many cases. The obvious drawback of Gaussian RBF is their low discriminative ability in the presence of numerous noisy features. To deal with noisy features one may consider basis functions consisting of a single feature

$$\phi_{ij}(\vec{x}) = f(|x(j) - y_i(j)|).$$

Although representing $\vec{x}$ as a vector $(\phi_{11}(\vec{x}), \ldots, \phi_{m,d}(\vec{x}))$ is still possible, it may be more natural to form a table of m columns and d rows.

The tabular representation of data provides new options in analyzing feature sets. In particular, we may search for relevant columns and rows instead of searching for relevant features by setting one regularization coefficient for each column and row, hence reducing the number of hyperparameters from $m \times d$ to $m + d$. With the reduced number of adjustable parameters we may expect the final decision rule to have better generalization properties.

3. Related work

The idea of treating image features as tables is not new and has been considered by a number of authors. Many papers on tabular data consider the problem of dimensionality reduction (either supervised or unsupervised). In Yang et al. (2004) 2-dimensional PCA is proposed where each data point is treated as a matrix. In Xu et al. (2004) the authors proposed an image reconstruction criterion to obtain the original image matrices using two low dimensional coupled subspaces, which encode the row and column subspaces of the image. They suggested an iterative method, CSA (Coupled Subspaces Analysis) to optimize this criterion. They also prove that PCA and 2D-PCA are special cases of CSA. The generalization of LDA to tabular data has been proposed in Ye et al. (2004) and Li and Yuan (2005). More recently, Yang et al. (2009) has proposed projecting images along both row and column directions, in an effort to maximize the so called Laplacian Bidirectional Maximum Margin Criterion (LBMMC).

A variant of the Zero-norm SVM feature selection algorithm for tabular data was presented in Wolf et al. (2007). It was shown that the family of $m \times d$ tabular linear separators that form matrices of rank k has a VC-dimension of no more than $k(m + d)\log(k(m + d))$, which is much lower than the VC-dimension of $md + 1$ obtained for linear functions over the vector representation.

In the context of sparse methods several non-Bayesian techniques have been proposed, for example, Boser et al. (1992) and Tibshirani (1996). Automatic relevance determination was first proposed in MacKay (1992) which provides a Bayesian framework for determining irrelevant parameters in machine learning models. The assignment of individual regularization coefficient to each weight makes it possible to get extremely sparse decision rules.

Unlike non-Bayesian methods, ARD does not require to set manually the values of the sparsity controlling regularization parameters. The application of ARD to generalized linear models and in particular to linear/logistic regression was proposed in Tipping (2001) as the Relevance vector machine model (RVM).

Since fully Bayesian inference is intractable even for regression problems, different authors have used some approximations of the general Bayesian scheme. These include evidence maximization (see MacKay (1992) and Tipping (2001)), marginalization w.r.t. the hyperparameters (see Williams (1995) and Cawley and Talbot (2006)), and the variational inference (see Bishop and Tipping (2000)).

In case of classification further approximations are necessary to perform inference. Various approximations of the likelihood function with a Gaussian were suggested. In Tipping (2001) the authors used Laplace approximation. Local variational methods were proposed in Jaakkola and Jordan (2000). The closely related expectation propagation technique (see Minka (2001)) for approximate Bayesian inference in generalized linear models was suggested in Qi et al. (2004). Although ARD methods have been applied successfully for the search of relevant features, objects, and basis functions in many domains, over- and underfitting of RVM was reported in some cases (see Qi et al. (2004)).

To our knowledge, ARD methods have so far only been applied to problems where the objects are represented as vectors of features. Here, we extend the ARD framework to the case of tabular data.

4. GridRVM models

Consider a regression and two-class classification problem with tabular data. Let $(X, \vec{t}) = \{\vec{x}_n, t_n\}_{n=1}^N$ be the training set, where $t_n \in \mathbb{R}$ are target values for regression problem and $t_n \in \{-1, 1\}$ are class labels for classification problem. Each object $\vec{x}_n$ is represented as a table of generalized features $(\phi_{ij}(\vec{x}_n))_{i,j=1}^{M_1,M_2}$. Note that we will also use one-index notation $(\phi_k(\vec{x}_n))_{k=1}^M$, $M = M_1 M_2$ when we need to treat the description of the object as a vector. Define the following probabilistic model for regression (p-gridRVR):

$$p(\vec{t}, \vec{w}, \vec{\alpha}, \vec{\beta}, \gamma | X) = p(\vec{t}|X, \vec{w}, \gamma) p(\vec{w}|\vec{\alpha}, \vec{\beta}) p(\vec{\alpha}) p(\vec{\beta}) p(\gamma),$$

$$p(\vec{t}|X, \vec{w}, \gamma) = \prod_{n=1}^N \mathcal{N}(t_n | \vec{w}^T \vec{\phi}(\vec{x}_n), \gamma^{-1}),$$

$$p(\vec{w}|\vec{\alpha}, \vec{\beta}) = \frac{\prod_{i,j=1}^{M_1,M_2} \sqrt{\alpha_i \beta_j}}{\sqrt{2\pi}^{M_1 M_2}} \exp\left(-\frac{1}{2} \sum_{i,j=1}^{M_1,M_2} \alpha_i \beta_j w_{ij}^2\right), \tag{1}$$

$$p(\vec{\alpha}) = \prod_{i=1}^{M_1} \mathcal{G}(\alpha_i | a_0, b_0), \tag{2}$$

$$p(\vec{\beta}) = \prod_{j=1}^{M_2} \mathcal{G}(\beta_j | c_0, d_0), \tag{3}$$

$$p(\gamma) = \mathcal{G}(\gamma | e_0, f_0).$$

Here $\mathcal{G}(\alpha_i | a_0, b_0)$ stands for a gamma distribution over α_i with parameters a_0, b_0 and all $\alpha_i, \beta_j, \gamma \geq 0$. Note that the number of regularization coefficients $\vec{\alpha}$ and $\vec{\beta}$ is $M_1 + M_2$ while the number of weights $\vec{w}$ is M_1M_2. Within this model we assign independent regularization coefficients to each row and column of the tabular presentation. The regularization coefficient for the weight w_{ij} is the result of a combination of α_i and β_j. In p-gridRVR we take the product of the two. Alternatively, we may consider the sum, i.e.

$$p(\vec{w}|\vec{\alpha},\vec{\beta}) = \frac{\prod_{i,j=1}^{M_1,M_2}\sqrt{\alpha_i+\beta_j}}{\sqrt{2\pi}^{M_1M_2}}\exp\left(-\frac{1}{2}\sum_{i,j=1}^{M_1,M_2}(\alpha_i+\beta_j)w_{ij}^2\right), \tag{4}$$

We refer to this model as s-gridRVR. Note that if we consider independent regularization coefficients α_{ij} for each position in tabular representation, we get the standard RVR model.

In p- and s-gridRVR models we consider the joint influence of the row and column of each table entry on the associated feature weight. However the models have one important distinction. In the case of s-gridRVR, large values of α_i mean that all regularization coefficients for the i^{th} row are at least as large as α_i. The same of course holds for large values of β_j. In p-gridRVR the situation is different. Large values of, say, α_i do not necessarily imply large values of the regularization coefficient for a particular weight w_{ij} since the coefficient β_j may have a small value. Thus we may expect a different behavior from these models.

Now we define the probabilistic model for classification (p-gridRVM):

$$p(\vec{t},\vec{w},\vec{\alpha},\vec{\beta}|X) = p(\vec{t}|X,\vec{w})p(\vec{w}|\vec{\alpha},\vec{\beta})p(\vec{\alpha})p(\vec{\beta}),$$

where priors are computed using (1), (2), (3) and data likelihood function is

$$p(\vec{t}|X,\vec{w}) = \prod_{n=1}^{N}\sigma(t_n\vec{w}^T\vec{\phi}(\vec{x}_n)).$$

Here $\sigma(y) = 1/(1+\exp(-y))$ is a logistic function. Similar to regression case we may consider prior (4) instead of (1) thus defining the s-gridRVM model.

5. Variational learning

Variational methods (see Jordan et al. (1998)) are popular technique for inference in Bayesian models. These methods allow to move from hardly computable model evidence to its lower bound, which is much simpler for estimation. In this section we first briefly describe basic ideas of the variational approach and then show its application for learning in the p- and s-gridRVM models.

5.1 Global variational inference

Suppose we are given a probabilistic model with variables $(\vec{t},\vec{\theta})$, where $\vec{t}$ is observable and $\vec{\theta}$ is not. We would like to estimate the model evidence

$$p(\vec{t}) = \int p(\vec{t},\vec{\theta})d\vec{\theta},$$

which we assume cannot be found analytically. Variational inference introduces here some distribution over the unobservable variables $q(\vec{\theta})$. Using this distribution the model evidence can be decomposed as follows

$$\log p(\vec{t}) = \mathcal{L}(q) + \mathrm{KL}(q||p(\vec{\theta}|\vec{t})),$$

where

$$\mathcal{L} = \int q(\vec{\theta}) \log \frac{p(\vec{\theta}, \vec{t})}{q(\vec{\theta})} d\vec{\theta} \tag{5}$$

and $\mathrm{KL}(q||p)$ is the Kullback-Leibler divergence between two distributions. Since $\mathrm{KL}(q||p) \geq 0$, $\mathcal{L}$ is a lower bound on the log-evidence. Besides, $\log p(\vec{t})$ does not depend on $q(\vec{\theta})$ and hence maximization of the lower bound $\mathcal{L}$ w.r.t. $q(\vec{\theta})$ is equivalent to minimization of the KL divergence between $q(\vec{\theta})$ and posterior distribution $p(\vec{\theta}|\vec{t})$.

Now consider the case when the distribution $q(\vec{\theta})$ has a factorized form

$$q(\vec{\theta}) = \prod_i q_i(\vec{\theta}_i).$$

Here $\{\vec{\theta}_i\}$ is a decomposition of a full set of variables so that $\vec{\theta} = \sqcup_i \vec{\theta}_i$. In Jordan et al. (1998) it's shown that maximization of (5) can be done iteratively by the following recalculation formula:

$$q_i(\vec{\theta}_i) = \frac{1}{Z} \exp\left(\int \log p(\vec{t}, \vec{\theta}) \prod_{j \neq i} q_j(\vec{\theta}_j) d\vec{\theta}_j\right), \tag{6}$$

where Z is a normalization constant ensuring that $q_i(\vec{\theta}_i)$ is a distribution. In this recalculation process the lower bound (5) monotonically increases.

5.2 Local variational inference

Global variational methods are supposed to move from the hardly computable model evidence to its lower bound. However, in many practical models (including p- and s-gridRVM) this lower bound is still analytically intractable. The local variational approach (see Jaakkola and Jordan (2000)) introduces a further bound on $p(\vec{\theta}, \vec{t})$:

$$p(\vec{\theta}, \vec{t}) \geq F(\vec{\theta}, \vec{t}, \vec{\xi}) > 0.$$

This bound is tight for some particular value of $\vec{\xi}$ and so it is local. Substituting this bound into (5) gives the following result:

$$\log p(\vec{t}) \geq \mathcal{L} \geq \mathcal{L}_{local} = \int q(\vec{\theta}) \log \frac{F(\vec{\theta}, \vec{t}, \vec{\xi})}{q(\vec{\theta})} d\vec{\theta}.$$

The last expression can be optimized w.r.t. $q(\vec{\theta})$ and $\vec{\xi}$ for a sensible choice of a local variational bound.

5.3 p-gridRVM

In a regression problem we wish to calculate

$$p(t_{new}|\vec{x}_{new}, \vec{t}, X) = \int p(t_{new}|\vec{x}_{new}, \vec{w}, \gamma) p(\vec{w}, \vec{\alpha}, \vec{\beta}, \gamma|\vec{t}, X) d\vec{w} d\vec{\alpha} d\vec{\beta} d\gamma \tag{7}$$

for any new object $\vec{x}_{new}$. For the model p-gridRVR as well as for the model s-gridRVR this integration is intractable and hence some approximation scheme is needed. Here we use the variational approach, which has been successfully applied for the conventional RVM model in Bishop and Tipping (2000), and try to find a variational approximation $q(\vec{w}, \vec{\alpha}, \vec{\beta}, \gamma)$ of the true posterior $p(\vec{w}, \vec{\alpha}, \vec{\beta}, \gamma | \vec{t}, X)$ in the following family of factorized distributions:

$$q(\vec{w}, \vec{\alpha}, \vec{\beta}, \gamma) = q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta}) q_{\vec{\gamma}}(\vec{\gamma}). \tag{8}$$

Then (7) can be reduced to the integration over the factorized distribution q:

$$p(t_{new} | \vec{x}_{new}, \vec{t}, X) \simeq \int p(t_{new} | \vec{x}_{new}, \vec{w}, \gamma) q(\vec{w}, \vec{\alpha}, \vec{\beta}, \gamma) d\vec{w} d\vec{\alpha} d\vec{\beta} d\gamma = \\ \int p(t_{new} | \vec{x}_{new}, \vec{w}, \gamma) q_{\vec{w}}(\vec{w}) q_{\gamma}(\gamma) d\vec{w} d\gamma. \tag{9}$$

Use of the global variational approach for the p-gridRVR model leads to the estimation of the following lower bound of the model log-evidence:

$$\log p(\vec{t} | X) \geq \mathcal{L} = \int \log \frac{p(\vec{t} | X, \vec{w}, \gamma) p(\vec{w} | \vec{\alpha}, \vec{\beta}) p(\vec{\alpha}) p(\vec{\beta}) p(\gamma)}{q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta}) q_{\gamma}(\gamma)} \times \\ q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta}) q_{\gamma}(\gamma) d\vec{w} d\vec{\alpha} d\vec{\beta} d\gamma. \tag{10}$$

Maximization of the criterion function (10) w.r.t. distributions $q_{\vec{w}}(\vec{w})$, $q_{\vec{\alpha}}(\vec{\alpha})$, $q_{\vec{\beta}}(\vec{\beta})$, $q_{\gamma}(\gamma)$ using (6) leads to the following result:

$$q_{\vec{w}}(\vec{w}) = \mathcal{N}(\vec{w} | \vec{\mu}, \Sigma), \tag{11}$$

$$q_{\vec{\alpha}}(\vec{\alpha}) = \prod_{i=1}^{M_1} \mathcal{G}(\alpha_i | a_i, b_i), \tag{12}$$

$$q_{\vec{\beta}}(\vec{\beta}) = \prod_{j=1}^{M_2} \mathcal{G}(\beta_j | c_j, d_j), \tag{13}$$

$$q_{\gamma}(\gamma) = \mathcal{G}(\gamma | e, f), \tag{14}$$

where

$$\Sigma = \left(\mathrm{diag}(\mathbb{E}_{\vec{\alpha}} \alpha_i \mathbb{E}_{\vec{\beta}} \beta_j) + \mathbb{E}_{\gamma} \gamma \Phi^T \Phi \right)^{-1}, \\ \vec{\mu} = \mathbb{E}_{\gamma} \gamma \Sigma \Phi^T \vec{t}, \tag{15}$$

$$a_i = a_0 + \frac{M_2}{2}, \quad b_i = b_0 + \frac{1}{2} \sum_{j=1}^{M_2} \mathbb{E}_{\vec{\beta}} \beta_j \mathbb{E}_{\vec{w}} w_{ij}^2, \tag{16}$$

$$c_j = c_0 + \frac{M_1}{2}, \quad d_j = d_0 + \frac{1}{2} \sum_{i=1}^{M_1} \mathbb{E}_{\vec{\alpha}} \alpha_i \mathbb{E}_{\vec{w}} w_{ij}^2, \tag{17}$$

$$e = e_0 + \frac{N}{2}, \quad f = f_0 + \frac{1}{2} (\vec{t}^T \vec{t} - 2\vec{t}^T \Phi \mathbb{E}_{\vec{w}} \vec{w} + \mathrm{tr}(\Phi \Sigma \Phi^T) + \vec{\mu}^T \Phi^T \Phi \vec{\mu}). \tag{18}$$

The necessary statistics are calculated as follows:

$$\mathbb{E}_{\vec{w}} \vec{w} = \vec{\mu}, \tag{19}$$

$$\mathbb{E}_{\vec{w}} w_{ij}^2 = S_{ij,ij} + \mu_{ij}^2, \tag{20}$$

$$\mathbb{E}_{\vec{\alpha}} \alpha_i = \frac{a_i}{b_i}, \tag{21}$$

$$\mathbb{E}_{\vec{\alpha}} \log \alpha_i = \Psi(a_i) - \log b_i, \tag{22}$$

$$\mathbb{E}_{\vec{\beta}}\beta_j = \frac{c_i}{d_i}, \quad (23) \qquad \mathbb{E}_{\gamma}\gamma = \frac{e}{f}, \quad (25)$$

$$\mathbb{E}_{\vec{\beta}}\log\beta_j = \Psi(c_j) - \log d_j, \quad (24) \qquad \mathbb{E}_{\gamma}\log\gamma = \Psi(e) - \log d, \quad (26)$$

where $\Psi(a) = \frac{d}{da}\log\Gamma(a)$ — digamma function. The lower bound (10) can be calculated analytically.

Coming back to decision making (9) with (11) and (14) it can be shown that the integral (9) equals to

$$\int \mathcal{N}(t_{new}|\vec{\mu}^T\vec{\phi}(\vec{x}_{new}), \gamma^{-1} + \vec{\phi}^T(\vec{x}_{new})\Sigma\vec{\phi}(\vec{x}_{new}))\mathcal{G}(\gamma|e,f)d\gamma.$$

This is the one-dimensional integral and hence it can be easily estimated using the Monte Carlo technique. However, if N is large enough we can use the following estimate (see Bishop and Tipping (2000)):

$$p(t_{new}|\vec{x}_{new}, \vec{t}, X) \simeq \mathcal{N}\left(t_{new}\middle|\vec{\mu}^T\vec{\phi}(\vec{x}_{new}), \frac{1}{\mathbb{E}_{\gamma}\gamma} + \vec{\phi}^T(\vec{x}_{new})\Sigma\vec{\phi}(\vec{x}_{new})\right).$$

In contrast to the p-gridRVR model, the use of the global variational approach for the p-gridRVM model for classification leads to the evidence lower bound that can't be computed analytically. Here we follow the variational method for the conventional RVM model presented in Bishop and Tipping (2000) and use the local variational approach by introducing the Jaakkola-Jordan inequality (see Jaakkola and Jordan (2000)) for the data likelihood function:

$$p(\vec{t}|X, \vec{w}) \geq F(\vec{t}, X, \vec{w}, \vec{\xi}) = \prod_{n=1}^{N}\sigma(\xi_n)\exp\left(\frac{z_n - \xi_n}{2} - \lambda(\xi_n)(z_n^2 - \xi_n^2)\right), \quad (27)$$

where $\sigma(y) = 1/(1+\exp(-y))$ — sigmoid function, $\lambda(\xi) = \tanh(\xi/2)/(4\xi)$, $z_n = t_n\vec{w}^T\vec{\phi}(\vec{x}_n)$. This bound is tight for $\xi_n = z_n$ and is illustrated on Fig. 2, left. Then substituting the inequality (27) into the evidence lower bound we obtain:

$$\log p(\vec{t}|X) \geq \mathcal{L}_{local} = \int \log\frac{F(\vec{t}, X, \vec{w}, \vec{\xi})p(\vec{w}|\vec{\alpha}, \vec{\beta})p(\vec{\alpha})p(\vec{\beta})}{q_{\vec{w}}(\vec{w})q_{\vec{\alpha}}(\vec{\alpha})q_{\vec{\beta}}(\vec{\beta})}\times$$
$$q_{\vec{w}}(\vec{w})q_{\vec{\alpha}}(\vec{\alpha})q_{\vec{\beta}}(\vec{\beta})d\vec{w}d\vec{\alpha}d\vec{\beta}. \quad (28)$$

It can be shown that maximization of the criterion function (28) w.r.t. the distributions $q_{\vec{w}}(\vec{w})$, $q_{\vec{\alpha}}(\vec{\alpha})$, $q_{\vec{\beta}}(\vec{\beta})$ and the variational parameters $\vec{\xi}$ leads to formulae (11), (12), (13), where the corresponding parameters are calculated using (16), (17) and

$$\Sigma = \left(\mathrm{diag}(\mathbb{E}_{\vec{\alpha}}\alpha_i\mathbb{E}_{\vec{\beta}}\beta_j) + 2\Phi^T\Lambda\Phi\right)^{-1}, \quad \Lambda = \mathrm{diag}(\lambda(\xi_n)),$$
$$\vec{\mu} = \frac{1}{2}\Sigma\Phi^T\vec{t}, \quad (29)$$
$$\xi_n^2 = \vec{\phi}^T(\vec{x}_n)\mathbb{E}_{\vec{w}}\vec{w}\vec{w}^T\vec{\phi}(\vec{x}_n). \quad (30)$$

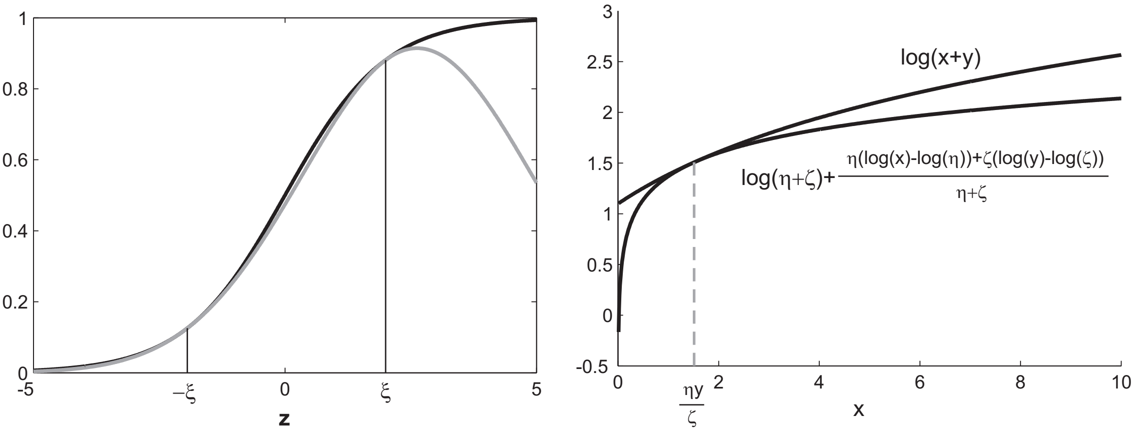


Figure 2: Left: Jaakkola-Jordan bound (27) for case $N = 1$, right: one-dimensional projection of the bound (31) for parameters $(y, \eta, \zeta) = (3, 2, 4)$.

The necessary statistics are computed using (19)-(24). The decision making scheme for this case is the following:

$$p(t_{new}|\vec{x}_{new}, \vec{t}, X) = \int p(t_{new}|\vec{x}_{new}, \vec{w}) q_{\vec{w}}(\vec{w}) q_{\vec{\alpha}}(\vec{\alpha}) q_{\vec{\beta}}(\vec{\beta}) d\vec{w} d\vec{\alpha} d\vec{\beta} =$$

$$\int p(t_{new}|\vec{x}_{new}, \vec{w}) q_{\vec{w}}(\vec{w}) d\vec{w} = \int \sigma(t_{new}\vec{w}^T \vec{\phi}(\vec{x}_{new})) \mathcal{N}(\vec{w}|\vec{\mu}, \Sigma) d\vec{w} =$$

$$\int \sigma(z) \mathcal{N}(z|t_{new}\vec{\mu}^T \vec{\phi}(\vec{x}_{new}), \vec{\phi}^T(\vec{x}_{new}) \Sigma \vec{\phi}(\vec{x}_{new})) dz.$$

The last integral is one-dimensional one and thus can be effectively estimated using the Monte Carlo technique. The useful analytical approximation for this integral is proposed in MacKay (1992):

$$\int \sigma(z) \mathcal{N}(z|m, s^2) dz \simeq \sigma \left(\frac{m}{\sqrt{1 + \frac{\pi s^2}{8}}} \right).$$

5.4 s-gridRVM

Similar to the previous case we propose to apply the variational approach for the s-gridRVM model. In this way we try to find a variational approximation q to the true posterior $p(\vec{w}, \vec{\alpha}, \vec{\beta}, \gamma|\vec{t}, X)$ in the family of factorized distributions (8) by optimizing the lower bound (10). However, in the case of the s-gridRVM model the criterion function (10) becomes intractable and we need a further lower bound in sense of the local variational methods. For this reason let us consider the function $f(x, y) = \log(x + y)$. This function is strictly concave. Now let us substitute the variables $x_1 = \log(x), y_1 = \log(y)$ and consider the function $f_1(x_1, y_1) = f(\exp(x_1), \exp(y_1)) = \log(\exp(x_1) + \exp(y_1))$. The function f_1 is convex and hence satisfies the following inequality:

$$f_1(x_1, y_1) \geq \frac{\partial f_1}{\partial x_1}(\eta)(x_1 - \eta) + \frac{\partial f_1}{\partial y_1}(\zeta)(y_1 - \zeta) + f_1(\eta, \zeta)$$

for arbitrary η and ζ. This inequality is just a relation between the function and its tangent line and becomes equality when $x_1 = \eta, y_1 = \zeta$. Moving back to initial variables x, y, we obtain the following variational bound:

$$\log(x+y) \geq \log(\eta+\zeta) + \frac{\eta(\log(x) - \log(\eta)) + \zeta(\log(y) - \log(\zeta))}{\eta+\zeta}, \tag{31}$$

which is tight when $x/y = \eta/\zeta$. One-dimensional projection of this bound is illustrated in Figure 2, right. Inequality (31) leads to the following bound on $\log p(\vec{w}|\vec{\alpha}, \vec{\beta})$:

$$\log p(\vec{w}|\vec{\alpha},\vec{\beta}) = \frac{1}{2}\sum_{i,j=1}^{M_1,M_2}[\log(\alpha_i+\beta_j) - (\alpha_i+\beta_j)w_{ij}^2] - \frac{M_1M_2}{2}\log 2\pi \geq$$

$$G(\vec{w},\vec{\alpha},\vec{\beta},\vec{\eta},\vec{\zeta}) = \frac{1}{2}\sum_{i,j=1}^{M_1,M_2}\left[\log(\eta_{ij}+\zeta_{ij}) + \frac{\eta_{ij}(\log(\alpha_i) - \log(\eta_{ij}))}{\eta_{ij}+\zeta_{ij}} + \right.$$

$$\left.\frac{\zeta_{ij}(\log(\beta_j) - \log(\zeta_{ij}))}{\eta_{ij}+\zeta_{ij}}\right] - \frac{1}{2}\sum_{i,j=1}^{M_1,M_2}(\alpha_i+\beta_j)w_{ij}^2 - \frac{M_1M_2}{2}\log 2\pi.$$

This bound is tight if $\eta_{ij} = \alpha_i$ and $\zeta_{ij} = \beta_j$. Substituting this inequality into (10) we obtain:

$$\log p(\vec{t}|X) \geq \int \log \frac{p(\vec{t}|X,\vec{w},\gamma)G(\vec{w},\vec{\alpha},\vec{\beta},\vec{\eta},\vec{\zeta})p(\vec{\alpha})p(\vec{\beta})p(\gamma)}{q_{\vec{w}}(\vec{w})q_{\vec{\alpha}}(\vec{\alpha})q_{\vec{\beta}}(\vec{\beta})q_\gamma(\gamma)} \times$$

$$q_{\vec{w}}(\vec{w})q_{\vec{\alpha}}(\vec{\alpha})q_{\vec{\beta}}(\vec{\beta})q_\gamma(\gamma)d\vec{w}d\vec{\alpha}d\vec{\beta}d\gamma. \tag{32}$$

Maximization of the criterion function (32) w.r.t. distributions $q_{\vec{w}}(\vec{w})$, $q_{\vec{\alpha}}(\vec{\alpha})$, $q_{\vec{\beta}}(\vec{\beta})$, $q_\gamma(\gamma)$ and variational parameters $\vec{\eta}, \vec{\zeta}$ leads to the formulae (11)-(14), where the corresponding parameters are calculated using (15), (18) and

$$\Sigma = \left(\operatorname{diag}(\mathbb{E}_{\vec{\alpha}}\alpha_i + \mathbb{E}_{\vec{\beta}}\beta_j) + \mathbb{E}_\gamma\gamma\Phi^T\Phi\right)^{-1},$$

$$a_i = a_0 + \frac{1}{2}\sum_{j=1}^{M_2}\frac{\eta_{ij}}{\eta_{ij}+\zeta_{ij}}, \quad b_i = b_0 + \frac{1}{2}\sum_{j=1}^{M_2}\mathbb{E}_{\vec{w}}w_{ij}^2, \tag{33}$$

$$c_j = c_0 + \frac{1}{2}\sum_{i=1}^{M_1}\frac{\zeta_{ij}}{\eta_{ij}+\zeta_{ij}}, \quad d_j = d_0 + \frac{1}{2}\sum_{i=1}^{M_1}\mathbb{E}_{\vec{w}}w_{ij}^2, \tag{34}$$

$$\eta_{ij} = \exp(\mathbb{E}_{\vec{\alpha}}\log\alpha_i),\ \zeta_{ij} = \exp(\mathbb{E}_{\vec{\beta}}\log\beta_j). \tag{35}$$

Use of the variational approach for the s-gridRVM model for classification requires both local variational bounds (27) and (31). In this case the optimal distribution q is obtained using (11)-(13), where the corresponding parameters are computed using (29), (30), (33)-(35) and

$$\Sigma = \left(\operatorname{diag}(\mathbb{E}_{\vec{\alpha}}\alpha_i + \mathbb{E}_{\vec{\beta}}\beta_j) + 2\Phi^T\Lambda\Phi\right)^{-1}, \quad \Lambda = \operatorname{diag}(\lambda(\xi_n)).$$

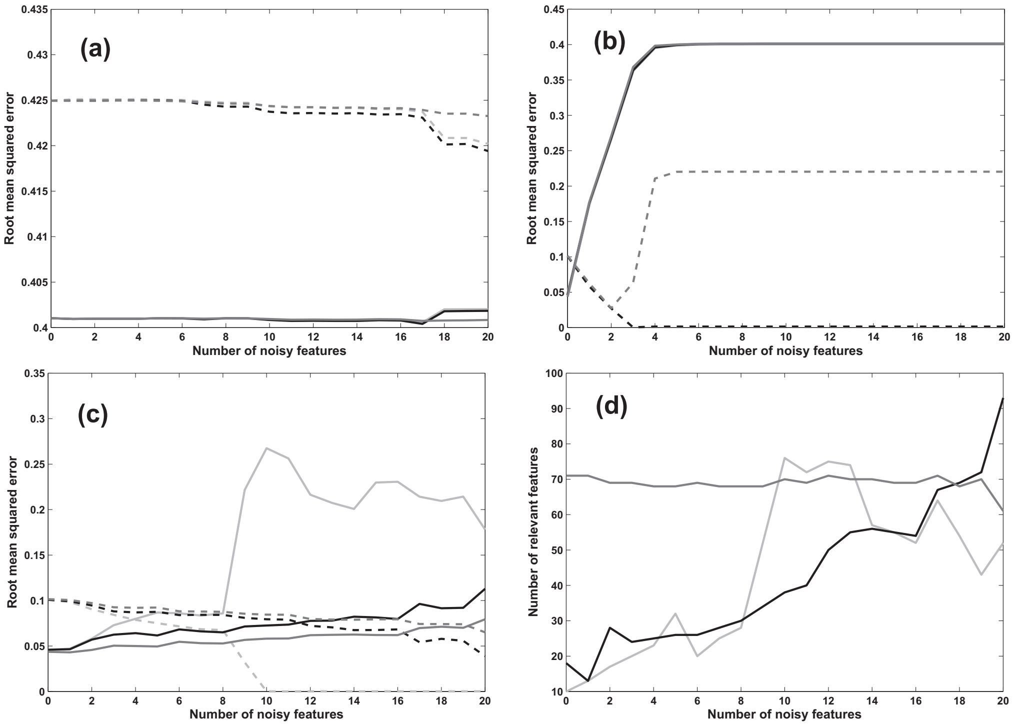


Figure 3: Sinc results. Color legend: black – p-gridRVR, dark grey – s-gridRVR, light grey – RVR. RMSE for train set is shown by dotted line, for test set – by solid line.

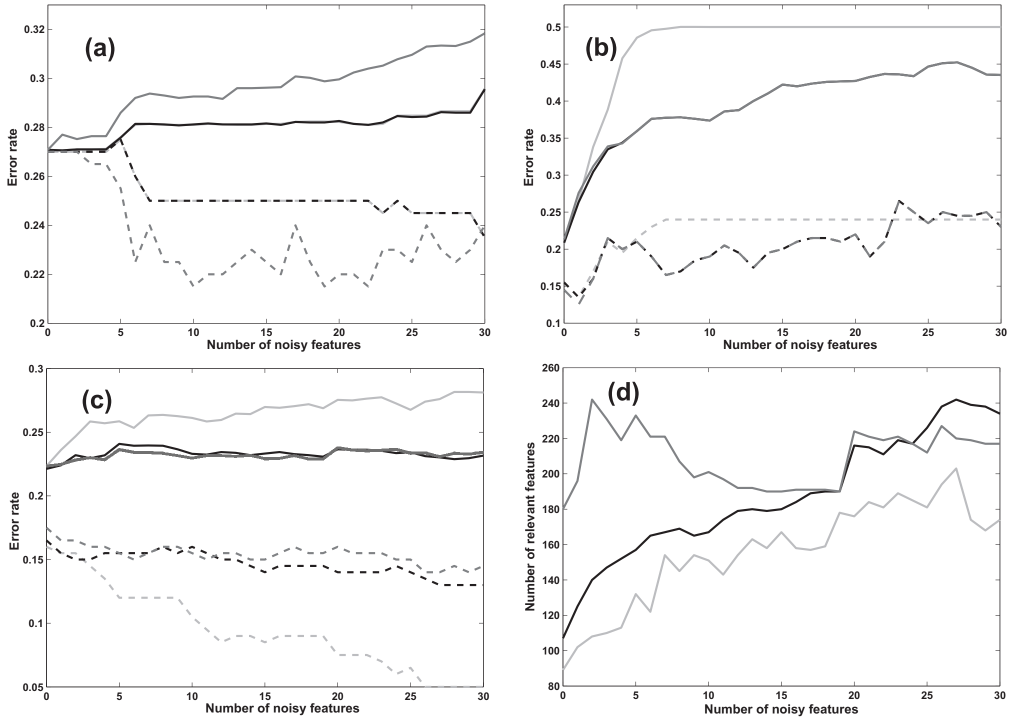


Figure 4: Results for Mixture dataset. Color legend: black – p-gridRVM, dark grey – s-gridRVM, light grey – RVM. Train error is shown by dotted line, test error – by solid line.

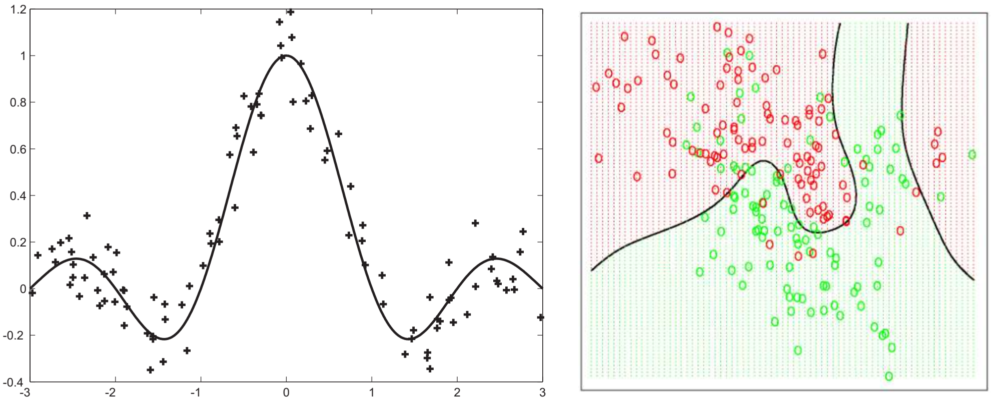


Figure 5: Artificial datasets. Left: noisy sinc, right: classification problem Mixture.

6. Experiments

We start with artificial regression and classification problems and then consider a real-world problem, for which tabular representation of data is natural.

6.1 Toy examples

First consider an artificial regression dataset. The training set consists of 100 points sampled from one-dimensional sinc function on the interval $[-10, 10]$ with additional uniform noise on the interval [-0.2,0.2]. The testing set consists of 600 points without noise. The normalized dataset is shown in Figure 5, left. In this experiment we add up to 20 normally distributed noisy features and investigate the behaviour of the conventional variational RVR (see Bishop and Tipping (2000)), p-gridRVR and s-gridRVR with 3 types of basis functions. In the first case we take initial features, i.e. $\phi_j(\vec{y}) = y(j)$ (total d features). This corresponds to a linear regression function. In the second case we take Gaussian RBFs of the form $\phi_j(\vec{y}) = \exp(-\delta\|\vec{y}-\vec{x}_j\|^2)$, where $\vec{x}_j$ are training objects (total N features). In the third case we take separate RBFs calculated for each dimension, i.e. $\phi_{ij}(\vec{y}) = \exp(-\delta(y(i) - x_j(i))^2)$ (total Nd features). In the first two cases we have a standard vector representation of objects, $M_2 = 1$ for both gridRVRs and hence gridRVRs are very similar to standard RVR here. In the last case we may treat objects' representation both as matrix of size $N \times d$ for gridRVRs and as a vector of length Nd for RVR. The experimental results (root mean squared error, RMSE) are shown in figure 3 (case a for initial features, case b for standard RBFs and case c for RBFs calculated for each dimension). In all cases $\delta = 5.55$. In the first case both train and test RMSE are above 0.4 because linear regression is inadequate for explaining non-linear data. In the second case there is no tabular representation of feature space and hence all three methods show almost similar performance that quickly degrade with the addition of noisy features. In the third case conventional RVR begins to overfit starting from several noisy features while both gridRVR methods show stable performance even with 20 additional noisy features. This is because gridRVR methods require tuning only $100 + d$ regularization parameters compared to $100d$ for conventional RVR. Figure 3,d shows the number of the relevant basis functions (the ones with weights with absolute values more than 0.1) for the third type of basis functions. We can see that s-gridRVR gives less sparse solution compared to p-gridRVR.

Comparing training time for gridRVRs and conventional RVR methods we observed that per iteration training time were almost similar. Hence training times of all methods fully depends on number of iterations for convergence. Although this number may differ for various methods the resulting divergence in training time was at most 50% and in many situations was almost zero.

Now consider an artificial classification dataset[1] taken from Friedman et al. (2001) (see Figure 5, right). This is a 2-class problem with 200 objects in the training set and 5000 objects in the test set. The feature space is two-dimensional and the data are generated from a specified distribution with Bayesian error rate 19%. The optimal discriminative surface is sufficiently non-linear. Here we again add up to 30 normally distributed noisy features and consider behaviour of conventional variational RVM, p-gridRVM and s-gridRVM with 3 types of basis functions: initial (corresponds to a linear hyperplane) (see fig. 4,a), standard Gaussian RBFs (see fig. 4,b) and Gaussian RBFs calculated for each dimension (see fig. 4,c). In all cases $\delta = 5.55$. Here in the first case we have more than 27% error rate for all three methods because linear hyperplane is inadequate for this non-linear data. For the second case all methods show similar performance and quickly overfit with the addition of noisy features. However, the overfit speed for gridRVM methods is less than for RVM. In the last case, where the tabular representation of data is appropriate, gridRVM methods show stable performance resulting in 22–23% of errors even for 30 noisy features while RVM definitely overfits starting from several noisy features. The number of the relevant basis functions for the last type of basis functions is shown on fig. 4,d. We can see that again the s-gridRVM model gives less sparse solution compared to p-gridRVM.

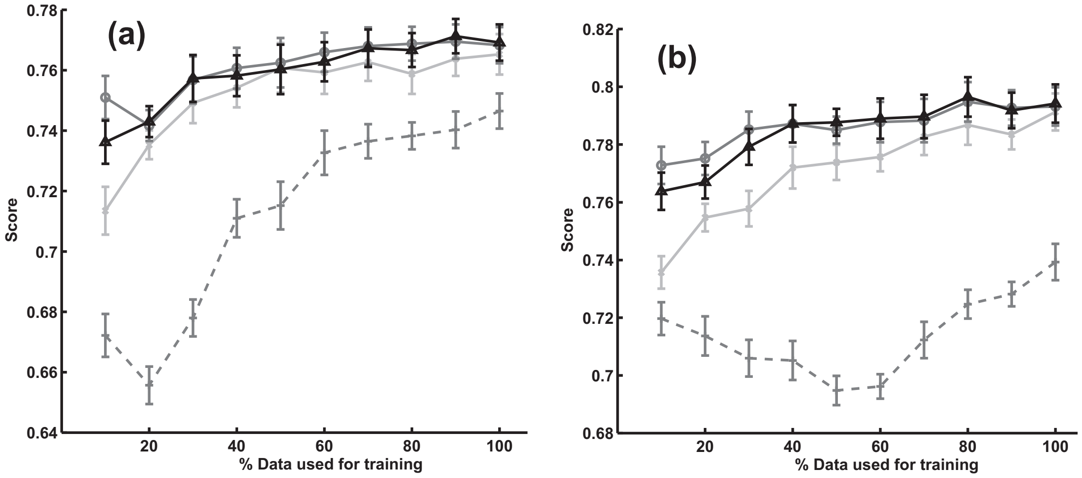


Figure 6: LFW results. Please see the text for more details

6.2 Face image pair-matching

We test our method on the Labeled Faces in the Wild (LFW) pair-matching benchmark (see Huang et al. (2007)). The LFW data set provides around 13,000 facial images of 5,749 individuals, each having from 1 to 150 images. These images were automatically harvested from news websites and thus represent faces under challenging, unconstrained viewing conditions. The goal of the benchmark is to determine, given a pair of images from

[1] http://www-stat.stanford.edu/˜tibs/ElemStatLearn/datasets/mixture.example.data to download

the collection, whether the two images match (portray the same subject) or not. To this end, 6,000 image pairs have been selected, half labeled "same" and half "not-same". For the purpose of testing, these pairs were further divided into ten, cross-validation "splits".[2]

Our tests build on the ones described in Wolf et al. (2008), where state-of-the-art results were obtained on the LFW benchmark. The results reported here employ the best performing descriptors for face-images and classifiers used in Wolf et al. (2008). Although other descriptors have been proposed for general classification of images (e.g., the SIFT (Lowe (1999)) and GIST (Oliva and Torralba (2001)) descriptors), the ones described next have been tailored for images of faces and are therefore used here. Specifically, we use the following four image descriptors: Local Binary Patterns (LBP) Ojala et al. (2001), Center Symmetric LBP (CSLBP) Heikkilä et al. (2006), and the Three and Four Patch LBP descriptors (TPLBP and FPLBP resp.) Wolf et al. (2008). Each face image was subdivided into 63 non-overlapping blocks of 23×18 pixels centered on the face. A separate histogram of codes was computed for each block, with 59 values for the uniform version of the LBP descriptor, 16 values for each of the CSLBP and FPLBP descriptors, and 256 values for the TPLBP descriptor.

Each pair of images to be compared is represented by one table of similarity values. The rows of the tables correspond to types of similarities values, and the columns correspond to the 63 facial regions depicted in Figure 1(a). The types of similarity values are all possible combinations of the four image representation above, and four histogram distances and similarities.

These four different histogram distances/similarities are computed block by block between the corresponding histograms of the two images. They are the L2 norm, the Hellinger Distance obtained by taking the square root of the histogram values, the so called One-Shot Similarity (OSS) measure (see Wolf et al. (2008)) (using the code made available by the authors), and OSS applied to the square root of the histogram values. To compute OSS scores we used 1,200 images in one of the training splits as a "negative" training set.

We report our results in Figure 6 where the pair-matching performance of s-gridRVM and p-gridRVM is compared against two baseline methods. Both figures plot classification scores across the ten-folds of the LFW benchmark, along with standard error values for different amounts of training (measured as the percentage of nine splits used as a training set). Figure 6(a) presents results using an 8×63 features of L2 and Hellinger distances between the four image descriptors; in Figure 6(b) we add also the four OSS scores and four OSS scores applied to the square roots of the histogram values.

As baseline methods we take linear SVM and standard RVM. We note that although non-linear SVM may be used to obtain even better results, we follow here the experiments reported in Wolf et al. (2008) and prefer linear SVM instead. As can be seen, the gridRVM methods show a clear advantage over both baseline methods. This is particularly true when only a small amount of training data is available. Although this advantage diminishes as more training is made available, both grid methods remain superior. Note that the results improve the ones reported in Wolf et al. (2008), where the same features were used for the whole image and the reported accuracy was 0.7847. p-gridRVM and s-gridRVM showed 0.7934 and 0.7942 of correct answers respectively. Note that since the publication of Wolf

[2] Here we use the aligned version of the image set made public at http://www.openu.ac.il/home/hassner/data/lfwa

et al. (2008), higher performance rates were reported on this benchmark. These, however, required additional training information, were obtained through a different protocol, or made possible by further processing of the images.

7. Conclusions

The experiments allow us to draw some conclusions. The first observation is that in some learning scenarios gridRVM is significantly more robust w.r.t. overfitting than standard RVM. This is particularly true for the case of small training samples with large amount of basis functions. It is important to stress that in case of large samples both standard and gridRVMs show almost identical results, so gridRVMs are not affected by underfitting although we reduced the number of adjustable regularization coefficients. The second observation is that both gridRVMs are sparse both in terms of regularization coefficients (many of them having large values) and in terms of the weights (many of whom are close to zero). Therefore, this useful and important property of standard RVM is kept. Comparing p- and s-gridRVM methods we may say that they show comparable performance with slight advantage of s-gridRVM. Also s-gridRVM is more robust to overfitting. However, p-gridRVM should be preferred in cases when sparsity is important.

Note that the grid approach can be straightforwardly generalized for the case of tensors (multidimensional tables). For example, we could treat the blocks in Figure 1 as a two-dimensional array hence obtaining a third dimension (together with the descriptor dimension) in the objects' description.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (project No. 09-01-92474) and Ministry of Science, Culture and Sport, State of Israel (project No. 3-5797).

References

C. Bishop and M. Tipping. Variational relevance vector machine. In *UAI*, 2000.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152, 1992.

G. Cawley and N. Talbot. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*, 22:2348–2355, 2006.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning.* Springer, 2001.

M. Heikkilä, M. Pietikäinen, and C. Schmid. Description of interest regions with center-symmetric local binary patterns. In *Computer Vision, Graphics and Image Processing, 5th Indian Conference*, pages 58–69, 2006.

G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. UMASS, TR 07-49, 2007.

T. Jaakkola and M. Jordan. Bayesian parameter estimation through variational methods,. *Statistics and Computing*, 10:25–37, 2000.

M. I. Jordan, Z. Gharamani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In *M. I. Jordan eds. Learning in Graphical Models*, pages 105–162, 1998.

M. Li and B. Yuan. 2D–LDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognition Letters*, 26(5):527–532, 2005.

D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

D. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

T. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, 2001.

T. Ojala, M. Pietikäinen, and T. Mäenpää. A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. In *ICAPR*, 2001.

A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

Y. Qi, T. Minka, R. Picard, and Z. Gharamani. Predictive automatic relevance determination by expectation propagation. In *ICML*, 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

M. Tipping. Sparse Bayesian learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

P. Williams. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

L. Wolf, H. Jhuang, and T. Hazan. Modeling appearances with low-rank SVM. In *CVPR*, 2007.

L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *Real-Life Images workshop at ECCV*, October 2008.

D. Xu, S. Yan, L Zhang, Z. Liu, and H. Zhang. Coupled subspaces analysis. Technical Report MSR-TR-2004-106, Microsoft Research, 2004.

J. Yang, D. Zhang, A. Frangi, and J. Yang. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131–137, 2004.

W. Yang, J. Wang, M. Ren, J. Yang, L. Zhang, and G. Liu. Feature extraction based on laplacian bidirectional maximum margin criterion. *Pattern Recognition*, 42(11):2327–2334, 2009.

J. Ye, R. Janardan, and Q. Li. Two-dimensional linear discriminant analysis. In *NIPS*, 2004.