Variational Bayes logistic regression as regularized fusion for NIST SRE 2010

Ville Hautamäki[1]*, Kong Aik Lee*[2]*, Anthony Larcher*[2]*, Tomi Kinnunen*[1]*, Bin Ma*[2]*, and Haizhou Li*[2]

[1]School of Computing, University of Eastern Finland (UEF), Finland.
[2]Human Language Technology Department, Institute for Infocomm Research (I^{2}R), Singapore.
{villeh, tkinnu}@cs.joensuu.fi
{kalee, alarcher, mabin, hli}i2r.a-star.edu.sg

Abstract

Fusion of the base classifiers is seen as a way to achieve high performance in state-of-the-art speaker verification systems. Typically, we are looking for base classifiers that would be complementary. We might also be interested in reinforcing good base classifiers by including others that are similar to them. In any case, the final ensemble size is typically small and has to be formed based on some rules of thumb. We are interested to find out a subset of classifiers that has a good generalization performance. We approach the problem from sparse learning point of view. We assume that the true, but unknown, fusion weights are sparse. As a practical solution, we regularize weighted logistic regression loss function by elastic-net and LASSO constraints. However, all regularization methods have an additional parameter that controls the amount of regularization employed. This needs to be separately tuned. In this work, we use *variational Bayes* approach to automatically obtain sparse solutions without additional cross-validation. Variational Bayes method improves the baseline method in 3 out of 4 sub-conditions.

Index Terms: logistic regression, regularization, compressed sensing, linear fusion, speaker verification

1. Introduction

Speaker verification is the task of accepting or rejecting an identity claim based on a person's voice sample [1]. Classification can be done on either *base classifier* level or at the level of *ensemble*, which is a technique known as *classifier fusion*. In this study, we focus on the latter.

In this paper, we consider linear classifier as a fusion device for the base classifer scores. Loss function used to optimize linear classifier parameters, i.e. the weight vector $\boldsymbol{w}$ and the bias b, play an important role as to how well the learned classifier generalizes to unseen data [2]. In this work we focus on the *logistic regression* model, which is a probabilistic discriminative linear model. As a loss function, *log loss* has several desirable properties, such as it does not overfit as easily as optimizing classification error directly. Logistic regression was introduced to speaker verification score fusion in [3] and later popularized by the *fusion and calibration* (FoCal) toolkit[1]. It has subsequently been found to be a useful linear fusion training methodology by in a number of independent studies (e.g. [4, 5]) and is taken here as a reference method.

A complete speaker verification system consisting of an ensemble of base classifiers might utilize, for instance, different speech parameterizations (e.g. spectral, prosodic or high-level features), classifiers (e.g. Gaussian mixture models [6] or support vector machines [7]), channel compensation techniques (e.g. joint factor analysis [8] or nuisance attribute projection [9]) or even selecting different datasets for estimating the hyperparameters. From such a list of features and classifiers a large number of potential ensembles is possible. Usually, it is left for an individual developer to select a suitable ensemble by hand, which may not be optimal in all cases. It is the topic of this paper to pursue an systematic method in fusion ensemble design.

One can just develop a large number of base classifiers and let the weight optimizer to find a good solution. Function to be optimized is a *proxy* of a classification error, upper bound of it that forms a convex function [10]. However, overfitting on the training data is still possible, even though an upper bound is optimized instead of classification error. To avoid overfit, regularization is required. The most common one is the quadratic regularization $\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$, also known as *ridge regression* [11]. Regularization forces parameter shrinkage, where the greater the Lagrange coefficient λ is, the smaller the norm $\|\boldsymbol{w}\|$ will be. Smaller norm implies better generalizability. Reason for this is easy to see, as higher norm means that some classifiers are given a large weight based on the training data. Effectiveness of these classifiers might not be realized on an unseen evaluation data.

In contrast to the ridge regression, an other approach is to regularize with a constraint that enforces *sparse* solutions. Extreme example is to regularize with $\|\boldsymbol{w}\|_0$ constraint [12]. Then λ signifies a maximum number of non-zero weights. This optimization problem is solved by optimizing weights for all the subsets of the constrained ensemble size and picking the one that performs the best. In our previous work [13] we selected the subset based on the training set and not on a cross-validation set, resulting in an underfit.

The time complexity of subset selection is exponential with respect to the number of base classifiers. Sparse solution can still be obtained by using a $\|\boldsymbol{w}\|_1$ constraint instead, also known as *least absolute shrinkage and selection operator* (LASSO) [12]. LASSO shrinks all of the coefficients, where some are forced to exactly zero. By regularizing logistic regression with the LASSO constraint, we can simultaneously optimize the fusion weights and perform classifier subset selection. Convex combination of ridge regression and LASSO leads to another regularization technique known as *elastic-net* (E-net) [14], which is sharp on the zeroing capability and yet smoother than the LASSO type of regularization. In addition, with elastic-net control of the norm of the weight vector can be more fine-grained than using LASSO, by increasing the influ-

The works of T. Kinnunen and V. Hautamäki were supported by the Academy of Finland project numbers 132129 and 253000.

[1]http://sites.google.com/site/nikobrummer/focal

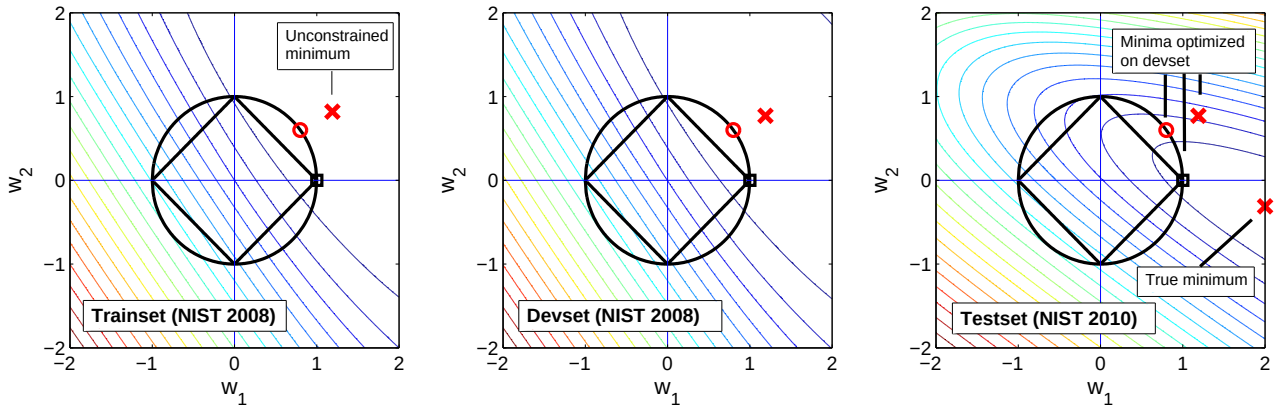


Figure 1: Visual intuition behind sparse classifier fusion. We display the contours of the C_{wlr} cost function for score fusion of two classifiers. In each panel, the global minimum of C_{wlr} is indicated by a red cross. In the constrained optimization case, we search for the minimum instead inside a constraint region specified by $(w_1^p + w_2^p)^{(1/p)} \leq 1$ (here, the cases $p = 1$ and $p = 2$ are visualized). As can be seen, the case $p = 1$ finds a *sparse* solution in the sense that classifier 2 is zeroed out. Luckily this solution hits closest to the true minimum on the unseen test data. Even L_2 regularization ($p = 2$) gives smaller cost than the unconstrained solution, suggesting that regularization and sparsification might be particularly useful for classifier fusion under unpredictable data mismatches.

ence of the ridge constraint. In our previous work [15], we proposed to use LASSO and elastic-net regularization techniques to simultaneously achieve generalizable fusion device and classifier subset selection.

Performance on the evaluation data is crucially dependent on selection (or estimation) of suitable regularization parameters. Cross-validation (CV) is a standard practice used to solve these kind of problems, but it requires a large number of computations. Also the solution can be quite noisy, select one CV set and you get one λ, with slightly a different set the difference in λ can be large [16]. In this work, we attempt to solve this problem by placing a suitable sparsity enforcing prior on a logistic regression and finding the so-called *variational Bayes* solution to it, instead of the MAP solution as we used in [15].

In addition, we were interested to find out whether a compromise could be found between LASSO and elastic-net, where we add an integer constraint to LASSO instead of ridge regression. our integer constraint specifies which base classifier should *not* be zeroed out. All experiments are now performed on the complete NIST 2010 core set.

2. Classifier Fusion

2.1. Problem Setup

We assume that, during the development phase, one has access to a development set $\mathcal{D} = \{(\mathbf{s}_i, y_i), i = 1, 2, \ldots, N_{\text{dev}}\}$ of base classifier score vectors $\mathbf{s}_i \in \mathbb{R}^L$, with $y_i \in \{+1, -1\}$ indicating whether the corresponding speech sample originates from a target speaker ($y_i = +1$) or from a non-target ($y_i = -1$). Using $\mathcal{D}$, the goal is to find the best parameters $\mathbf{w}^*$ of a linear combiner $f_{\mathbf{w}}(\mathbf{s}) = \mathbf{w}^t\mathbf{s} + w_0$ so that a classification error measure is minimized on unseen evaluation data. For our evaluation purposes, we adopt the standard *detection cost function* (DCF) used in the NIST speaker recognition evaluations[2],

$$C_{\text{det}}(\theta) = C_{\text{miss}} P_{\text{miss}}(\theta) P_{\text{tar}} + C_{\text{fa}} P_{\text{fa}}(\theta)(1 - P_{\text{tar}}). \quad (1)$$

[2] http://www.itl.nist.gov/iad/mig/tests/spk/

Here $P_{\text{miss}}(\theta)$ and $P_{\text{fa}}(\theta)$ are the miss and false alarm probabilities as a function of the decision threshold θ where P_{tar} is the prior probability of a target (true) speaker, C_{miss} is the cost of a miss (false rejection) and C_{fa} is the cost of a false alarm (false acceptance). These application-dependent cost parameters can also be summarized as a single cost parameter, *effective prior*:

$$P = \text{logit}^{-1}(\text{logit}(P_{\text{tar}}) + \log(C_{\text{miss}}/C_{\text{fa}})), \quad (2)$$

where $\text{logit}\, P = \log(P/(1-P))$. It is possible to minimize DCF directly (e.g. [17]) or to optimize a surrogate cost such as effective prior weighted logistic regression cost [4]. Here we adopt the latter approach which represents state-of-the-art.

2.2. Logistic regression

Logistic regression is a probabilistic linear model, which begins with the realization that target class posterior can be modeled as $p(y = 1|\boldsymbol{s}) = (1 + \exp\{-(\mathbf{w}^t\mathbf{s} + w_0)\})^{-1} = \sigma(\mathbf{w}^t\mathbf{s} + w_0)$, where $\sigma(.)$ is a logistic sigmoid function [2]. Non-target is then $p(y = -1|\boldsymbol{s}) = 1 - \sigma(\mathbf{w}^t\mathbf{s} + w_0) = \sigma(-\mathbf{w}^t\mathbf{s} - w_0)$, by utilizing the properties of $\sigma(.)$. The quantity $\mathbf{w}^t\mathbf{s} + w_0$ is then interpreted as a log of the ratio of probabilities $\ln[p(y = 1|\boldsymbol{s})/p(y = -1|\boldsymbol{s})]$, so called *log odds* [2]. This is useful, as if the posteriors in the log odds are well estimated then Bayes optimal cost-sensitive decision can be made by placing the threshold to $-\text{logit}\, P$.

Maximum likelihood estimate of the parameters can be found by taking the negative logarithm of the likelihood formulation, yielding the following *cross-entropy* cost [2]:

$$-\sum_{n=1}^{N} \{t_n \ln x_n + (1 - t_n) \ln(1 - x_n)\}, \quad (3)$$

where $t_n \in \{0, 1\}$ is relabeled (for mathematical convenience) class label and $x_n = \sigma(\mathbf{w}^t\mathbf{s}_n + w_0)$. Iterative gradient descent methods can then find parameter estimates.

2.3. Weighted cross-entropy objective

In speaker verification applications, we are usually interested in a specific set of DCF parameters, in effect in the training phase we learing the parameters in a cost-sensitive way. In addition, the ratio of positive and negative examples in the development set might be highly imbalanced. This is the case with the bi-annual NIST evaluation setup.

In the FoCal software package, indirect optimization of the fusion weights and bias given DCF parameters is achieved by modifying the cross-entropy objective C_{wlr}. In C_{wlr}, in addition to a global effective prior based bias, cross-entropy is weighted by the observed ratio of positive and negative examples [4]:

$$\begin{aligned} C_{\mathrm{wlr}}(\boldsymbol{w}, \boldsymbol{s}) &= \frac{P}{N_t} \sum_{i=1}^{N_t} \log\left(1 + e^{-\boldsymbol{w}^t \boldsymbol{s}_i - \mathrm{logit}\, P}\right) \\ &+ \frac{1-P}{N_f} \sum_{j=1}^{N_f} \log\left(1 + e^{\boldsymbol{w}^t \boldsymbol{s}_j + \mathrm{logit}\, P}\right), \end{aligned} \quad (4)$$

where the two sums go through N_t target score vectors $\boldsymbol{s}_i$ and N_f non-target score vectors $\boldsymbol{s}_j$, respectively. We will also do the standard bias encoding by adding one extra element containing 1 to $\boldsymbol{s}$. Global bias can then be extracted from the corresponding position in the weight vector.

3. Regularized Logistic Regression

We extend the weighted logistic regression in Eq. (4) by adding a regularization term [2]. It leads to minimizing,

$$C_{\mathrm{wlr}}(\boldsymbol{w}, \boldsymbol{s}) \quad \text{s.t.} \quad J(\boldsymbol{w}) \leq t, \quad (5)$$

where $J(\boldsymbol{w})$ can be $\frac{1}{2}\|\boldsymbol{w}\|_2^2$, which is called ridge regression, $\|\boldsymbol{w}\|_1 = \sum_{i=1}^{L} |w_i|$, which is called LASSO or $\|\boldsymbol{w}\|_0 = \sum_{i=1}^{L} w_i^0$, which is called subset selection. Quantity w_i^0 is 1 everywhere except when $w_i = 0$, then it will get value 0. In other words, the 0^{th} norm, simply counts the number of non-zero weights. The user specified parameter t indicates the intended amount of parameter shrinkage. The Lagrange coefficients will give us, in the case of LASSO, the following expression,

$$C_{\mathrm{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda \|\boldsymbol{w}\|_1. \quad (6)$$

It is known that the larger λ, the more norm $\|\boldsymbol{w}\|$ will be shrunk [12]. Example of (6) on real data can be seen in Fig. 1, where two base classifiers are fused. From the example it is clear that weights found by the direct optimization of (4) would lead to non-optimal solution for the NIST SRE 2010 data set.

If optimization is based on Eq. (6), then the correspondence between λ and shrinkage threshold t can be found by a binary search on possible λ values. In each iteration we select one λ value and optimize weights using it, output is then the norm of the weights. Final weight vector is the one whose norm is closest to the target t, but does not violate it.

Elastic-net, on the other hand, is based on the idea that we can combine both regularizers into one constraint optimization problem,

$$C_{\mathrm{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda \left(\beta \|\boldsymbol{w}\|_1 + (1-\beta)\|\boldsymbol{w}\|_2^2\right). \quad (7)$$

As can be seen, Eq. (7) is a generalized variant of both LASSO and ridge regression. That is, we can always find such a β where, in terms of performance, elastic-net will at least not lose to LASSO or ridge regression. However, whereas LASSO and ridge regression had to select only one regression parameter, now we need to do crossvalidation over a 2-d space. In this work we first fix the β parameter. Then the shrinkage factor λ can be cross-validated as in LASSO and ridge regression.

Depending on the chosen regularization method, there are different strategies to optimize (5). Since logistic regression using quadratic regularization is differentiable, it can be efficiently optimized using standard packages [2]. Situation is not so simple for LASSO regularization. In [12], a *quadratic programming* (QP) solution was proposed to it by rewriting the constraints in (5) to a more convenient form. However, more recent techniques are faster in practice, for that reason we apply the *projectionL1* algorithm [18] that optimizes the Lagrangian form Eq. (6). We apply the same method to elastic-net, as the sum of two convex functions is still convex; therefore we can minimize $C_{\mathrm{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda(1-\beta)\|\boldsymbol{w}\|_2^2$, given $\lambda\beta\|\boldsymbol{w}\|_1$ as the constraint.

3.1. Restricted LASSO

As λ increases, LASSO tends to zero out a large number of base classifiers. We are interested to find out if regularizing the LASSO regularizer will bring the extra benefit, i.e. in similar vein than in elastic-net, we add extra parameter to LASSO so that result will be less sparse. We call this method *restricted LASSO*. We can exclude any of the base classifiers from being zeroed out by adding an extra constraint $w_j \neq 0$ to (5). Taking Lagrange formulation of the constrained optimization problem, we give λ for $w_j \neq 0$ constraint. We assume that λ associated with $w_j \neq 0$ is the same as one for the LASSO. Thus, summing up both constraints together leads to the fact that λ equals to zero for base classifier j.

3.2. Bayesian interpretation

Regularized logistic regression can also be interpreted as the maximum *a posteriori* (MAP) estimate of the weight vector $\boldsymbol{w}$ [12], where regularization term in (6) acts as a prior.

We are interested in shrinking the parameters w_j towards zero, threrefor a simple prior per weight is univariate Gaussian with zero mean and τ_j variance, $p(w_j|\tau_j) = \mathcal{N}(0, \tau_j)$ [19]. By further assuming that τ_j is distributed according to exponential distribution:

$$p(\tau_j|\gamma_j) = \frac{\gamma_j}{2} \exp\left(-\frac{\gamma_j}{2}\tau_j\right), \quad (8)$$

integrating out τ_j from $p(w_j|\tau_j)p(\tau_j|\gamma_j)$ gives [19]:

$$p(w_j|\lambda_j) = \frac{\lambda_j}{2} \exp\left(-\lambda_j |w_j|\right), \quad (9)$$

where $\lambda_j = \sqrt{\gamma_j}$. Now, setting $\lambda_j = \lambda$ for all the base classifiers we obtain LASSO regularized logistic regression of (6). Similarly, modeling precision instead of variance ($\alpha = 1/\lambda$) the prior corresponding to ridge regression turn out to be,

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}). \quad (10)$$

3.3. Variational Bayes fusion

In contrast to the previous sub sections, where MAP point estimate was the goal, in variational Bayes (VB) approach we try to find the approximate *full* posterior [2]. Then no free parameters are assumed to be fixed or set using cross-validation techniques. The ultimate goal of the variational Bayes is to obtain

an approximate solution to the $p(\mathcal{D})$, where all unobserved parameters and latent variables have been marginalized out. However, direct optimization of $p(\mathcal{D})$ is typically not possible, so in the variational Bayes we iteratively optimize the lowerbound $\ln p(\mathcal{D}) \geq \mathcal{L}(q)$. Where q is the *variational distribution*. The method assumes that the variational posterior distribution then factorizes [2]. Factors can then be independently maximized, which leads to iterative EM-like optimization algorithm. In this work, we use implementation of VB logistic regression by Jan Drugowitsch [3].

In VB, we in addition to the prior in (10), where precision is treated as a free parameter, we place hyper-prior the precision α, $p(\alpha) = \mathrm{Gam}(\alpha|a_0, b_0)$ [2]. Modeling decision can be done because, Gamma is the conjugate prior precision of Normal with the known mean. Scalars a_0 and b_0 are parameters of the hyper-prior, in this work we select them to be non-informative.

In standard VB approach, one hyper-prior was selected for all base classifiers, but in *automatic relevance determination* (ARD) prior we will have per base classifier hyper prior. It aims to utilize data-dependent prior distribution that effectively prunes away redundant or superfluous features [20]. In this work, prior is selected to be the [2]:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1}), \tag{11}$$

where $\boldsymbol{A}$ is a diagonal matrix with diagonal elements $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$. The hyper-prior is then $p(\boldsymbol{\alpha}) = \prod_{i=1}^{L} \mathrm{Gam}(\alpha_i|a_0, b_0)$. Each base classifier has it's own prior, parametrized by precision α_i. We see that if α_i is high then it is likely that the base classifier i does not play any role in the fusion and w_i is most likely zero [2].

4. Corpora, Metrics and Base Classifiers

4.1. Experiments with I4U systems

We utilize the two most recent NIST SRE corpora, namely, NIST 2008 and NIST 2010, in our experiments[4]. The experiments are performed solely on female trials[5]. The audio files from all NIST 2008 speakers were split into two disjoint parts. Trials were then automatically generated from those two sets, while keeping observed p_{target} similar than in the official NIST 2008 SRE trial lists. The first part, *trainset*, is used for training the score warping parameters (S-cal [4] was used as precalibration method), fusion weights and bias. The second part, *cross validation set*, is used for estimating shrinkage parameter (λ) and the tradeoff parameter (β) between LASSO and elastic-net. The optimized parameters are then applied to the NIST 2010 data, which is reserved for the evaluation purposes.

For evaluation of the methods, we consider the detection cost function in (1), where the cost parameters are adopted from the previous NIST SRE evaluation plans, namely, $C_{\mathrm{miss}} = 10$ $C_{\mathrm{fa}} = 1$ and $P_{\mathrm{tar}} = 0.01$. Decision is based on the threshold obtained from the effective prior in Eq. (2). We are interested in comparing the application dependent classification error as measured in actual DCF (ActDCF). ActDCF is the error count after thresholding the scores.

In this study we use the same ensemble setup as in our previous studies [13, 15]. We have twelve subsystems in total, all are based on different cepstral features and four different classifiers, as part of the I4U system. For classifiers, we use the generative GMM-UBM-JFA [8] and the discriminative GMM-SVM approaches with KL-divergence kernel [21] and the Bhattacharyya kernel [22]. We also include another method, feature transformation [23], as an alternative supervector for SVM. All of the methods are grounded on the *universal background model* (UBM) paradigm and share similar form of subspace channel compensation, though the training methods differ. We used data from the NIST SRE 2004, SRE 2005 and SRE 2006 corpora to train the UBM and the session variability subspaces, and additional data from the Switchboard corpus to train the speaker-variability subspace for the JFA systems. Each base classifier has its own score normalization prior to score warping and fusion. To this end, we use T-norm and Z-norm with NIST SRE 2004 and SRE 2005 data as the background and cohort training data.

[3] `http://www.lnc.ens.fr/~jdrugowi/code_vb.html`

[4] `http://www.itl.nist.gov/iad/mig//tests/sre/`

[5] Female trials are somewhat more difficult than males. Similar rationale was taken, for instance, in [8].

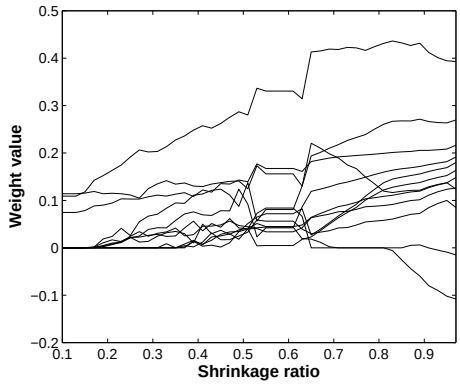


Figure 2: Weight evolution of the Elastic-net regularization, with $\alpha = 0.7$, as a function of normalized t.

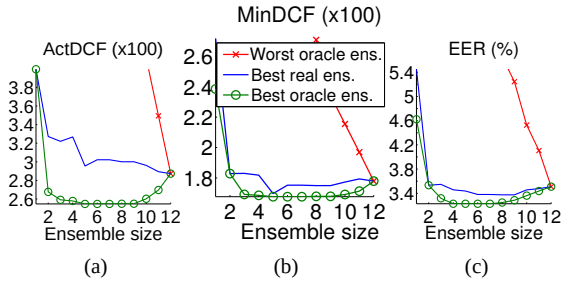


Figure 3: Effect of ensemble size to accuracy Evalset using VB logistic regression. For a fixed ensemble size (K), the lowest (green) and highest (red) lines show the best and worst possible selections out from the $\binom{12}{K}$ choices from Evalset (NIST SRE 2010). The middle (blue) line indicates the actual ensemble selected by cross-validation set.

5. Experiments

It is instructive to show the evolution of the individual classifier weights as the function of threshold parameter t. In Fig. 2 we observe the fusion weights as a function of normalized shrinkage threshold $\hat{t} = t/\|\hat{\boldsymbol{w}}\|$, where $\hat{\boldsymbol{w}}$ is the unregularized solution. We see that $\hat{t}$ will tell how much of the unregularized norm is left after shrinkage. Regularization path of the elastic-net so-

lutions shows grouping effect, it appears to group classifiers into 4 different groups with the $\beta = 0.7$ selection. Only two classifiers are zeroed out when shrinkage ratio is set to 0.66.

Table 1: Variational Bayes logistic regression compared to maximum likelihood trained logistic regression.

	Fusion	EER (%)	MinDCF ($\times$100)	ActDCF ($\times$100)	Ensemble size
itv-itv	Log. Regr	3.55	1.8072	**2.8420**	12
	VB	3.51	1.7789	2.8728	10
	VB-ARD	**3.48**	**1.7621**	2.9289	10
itv-tel	Log. Regr	**2.40**	0.98	**1.74**	12
	VB	2.50	**0.9683**	2.0020	12
	VB-ARD	2.50	0.9924	2.0112	12
mic-mic	Log. Regr	**5.10**	.2.35	**4.14**	12
	VB	**5.10**	2.2273	4.8788	9
	VB-ARD	5.67	**2.1127**	5.6405	9
tel-tel	Log. Regr	2.33	**1.12**	**1.18**	12
	VB	**2.23**	1.1396	3.0361	12
	VB-ARD	2.27	1.1746	3.1334	12

5.1. Variational Bayes approaches

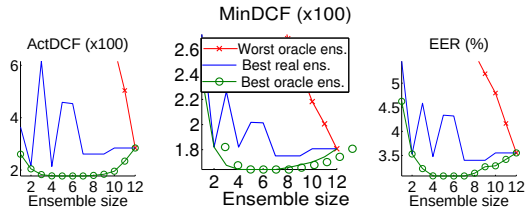
Variational Bayesian (VB) logistic regression with automatic relevance determination (ARD) prior results are shown in Table 1. In logistic regression, C_{wlr} was optimized, but in both VB approaches observed target/non-target ratios did not play any role nor did we use global effective prior bias. It was assumed instead that in terms of DCF, logistic regression would be the winner. However, in all but tel-tel condition VB approaches win in terms of MinDCF. Showing that there is no need to weight the cost function by target/non-target ratios. By not applying the global effective prior bias our VB approaches did not obtain well calibrated scores.

In terms of EER, VB works better for the itv-itv and tel-tel conditions. Relative improvement over logistic regression baseline in the tel-tel condition is 5.6%.

It is interesting to note that both VB and VB-ARD approaches do, in fact, zero out some base classifiers, but only for the case of itv-itv and mic-mic conditions. Table 1 also shows that difference between VB and VB-ARD is small. We do not consider ARD further.

As noted from Table 1, the VB approaches are not aggressive in finding sparse solutions. For instance, in itv-itv condition, only two base classifiers are zeroed out. However, we can also utilize subset selection methodology, where VB solution is found for all subsets of base classifiers. Results for this are shown in Fig. 3, where the solution chosen based on the cross-validation set is shown in blue. The best and the worst bounds are also shown. As a comparison, we show also the same experiment when logistic regression was used instead of VB in Fig. 4. We notice that VB provides much more stable performance as function of subset size.

In Table 2, subset size is also selected from cross-validation set. There is an improvement over the full ensemble methods except in ActDCF. Using subset selection, the ensemble size was further reduced from 10 to 6.



(a) (b) (c)

Figure 4: Effect of ensemble size to accuracy (Evalset 2) using logistic regression. For a fixed ensemble size (K), the lowest (green) and highest (red) lines show the best and worst possible selections out from the $\binom{12}{K}$ choices from Evalset 2 (NIST SRE 2010). The middle (blue) line indicates the actual ensemble selected by cross-validation Evalset 1.

Table 2: Variational Bayes using subset selection and without subset selection applied to itv-itv portion of Evalset. Subset is selected from cross-validation set.

	Fusion	EER (%)	MinDCF ($\times$100)	ActDCF ($\times$100)	Ensemble size
full set	Log. Regr	3.55	1.8072	**2.8420**	12
	VB	3.51	1.7789	2.8728	10
	VB-ARD	3.48	1.7621	2.9289	10
subset	subset Log. Regr	3.40	**1.7506**	**2.6119**	6
	subset VB	3.38	1.7524	3.0221	6
	subset VB-ARD	**3.37**	1.7532	3.0322	6

5.2. Summarization of results

In Table 3 we show the recognition results for different NIST SRE 2010 sub-conditions (itv-itv, itv-tel, mic-mic and tel-tel). Here, baseline method refers to the unregularized solution (i.e. $\lambda = 0$), equivalent to the implementation of the FoCal toolkit. Best single classifier is selected based on the performance on the cross validation set, so all the methods are directly, and fairly, comparable in Table 3. We notice that, for the itv-tel and mic-mic subconditions, elastic-net and subset selection achieve similar and the best results. It is interesting to note that improvement in the ActDCF is because scores are better calibrated.

General trend, when comparing minDCF over all conditions seems to be that there are no large differences except in the mic-mic condition where no regularization clearly fails. Differences in ActDCF are mostly the product of different calibrations. Note that the bias is *not* regularized.

It is interesting to note that predicting the β value using cross validation set is not a trivial problem. It is clear that in the case when either LASSO or ridge wins over elastic-net in terms of ActDCF, the prediction of β was unsuccesful. Especially interesting is the itv-itv case, where prediction gave $\beta = 0$ (i.e. ridge) and for NIST SRE 2010, LASSO was clearly better.

Regularization, however, does not bring improvement in the tel-tel condition in terms of ActDCF. For the tel-tel condition, designers of base classifiers had a very large and extensively used corpora available for tuning up their systems. In addition, selection of data sets for the estimation of session compensation parameters is more straightforward. But the interview and microphone data conditions did not have such a wealth of material backing their classifier design. It is thus expected that regu-

larization will hurt the classification performance in the tel-tel condition. In the other conditions, significant improvement over the baseline can be achieved by any of the regularization methods.

In Table 3, cross-validation was used to select the classifier not to be regularized for restricted LASSO. We notice that leaving one base classifier out of the LASSO regularization does not necessarily lead to an increase in ensemble size by one classifier, as one would expect. For example, in the case of itv-tel ensemble size was actually decreased from 8 to 7. In other conditions, increase in ensemble size is observed, extreme being tel-tel condition where ensemble size was increased from 5 to a full ensemble.

Restricted LASSO wins in the itv-tel condition, where EER improves from 2.40% to 2.25%. Lowest EER (2.25%) in itv-tel and minDCF (1.1074) in tel-tel condition are obtained using this configuration.

6. Conclusions

We have studied regularized logistic regression fusion on the NIST SRE 2010 core test conditions. We find that regularization brings improvement over unregularized variant in all other sub-conditions and measures (EER, MinDCF, ActDCF) except ActDCF in tel-tel condition, even there our proposed restricted LASSO achieves practically same result as no regularization.

Regularization techniques need a separate tuning for the λ parameter. Here we have studied how to automatically obtain sparse solutions using variational Bayesian logistic regression. We obtained sparse solutions on 2 out of 4 sub-conditions. As a result, the obtained EER is same or lower in 3 out 4 sub-conditions and MinDCF is lower in 3 out of 4 sub-conditions than baseline logistic regression.

As a future work we plan to extend the variational Bayes approach used in this paper to the different, and more aggressive priors such as elastic-net.

7. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[2] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media, LLC, 2006.

[3] S. Pigeon, P. Druytsa, and P. Verlinde, "Applying logistic regression to the fusion of the nist'99 1-speaker submissions," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 237–248, January 2000.

[4] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grézl, M. Karafiát, D. Leeuwen, P. Matějka, P. Schwartz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, September 2007.

[5] L. Ferrer, K. Sönmez, and E. Shriberg, "An anticorrelation kernel for subsystem training in multiple classifier systems," *J. of Machine Learning Research*, vol. 10, pp. 2079–2114, 2009.

[6] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[7] W. Campbell, J. Campbell, D. Reynolds, E. Singer, and P. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[8] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE T. Audio, Speech & Lang. Proc.*, vol. 16, no. 5, pp. 980–988, July 2008.

[9] A. Solomonoff, W. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP 2005*, Philadelphia, Mar. 2005, pp. 629–632.

[10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2008.

[11] A. Hoerl and R. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[13] F. Sedlak, T. Kinnunen, V. Hautamäki, K. Lee, and H. Li, "Classifier subset selection and fusion for speaker verification," in *ICASSP 2011*.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of Royal Statistical Society, Series B*, vol. 67, pp. 301–320, 2005.

[15] V. Hautamäki, K. Lee, T. Kinnunen, B. Ma, and H. Li, "Regularized logistic regression fusion for speaker verification," in *Interspeech 2011*, Florence, Italy, August, pp. 2745–2748.

[16] G. Kubin, C. Lainscsek, and E. Rank, "Identification of nonlinear oscillator models for speech analysis and synthesis," in *Nonlinear Speech Modeling and Applications*, ser. Lecture Notes in Computer Science, G. Chollet, A. Esposito, M. Faundez-Zanuy, and M. Marinaro, Eds. Springer Berlin / Heidelberg, 2005, vol. 3445, pp. 1–3.

[17] W. Campbell, D. Sturim, W. Shen, D. Reynolds, and J. Navratil, "The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition," in *Proc. ICASSP 2007*, vol. IV, 2007, pp. 217–220.

[18] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for L1 regularization: A comparative study and two new approaches," in *ECML 2007*, Warsaw, Poland, September 2007.

[19] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, August 2007.

[20] R. M. Neal, *Bayesian Learning for Neural Networks*. New York: Springer-Verlag, 1996.

[21] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, May 2006.

[22] C. You, K. Lee, and H. Li, "GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1300–1312, August 2010.

[23] D. Zhu, B. Ma, and H. Li, "Joint MAP adaptation of feature transformation and gaussian mixture model for speaker recognition," in *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP 2009)*, Taipei, Taiwan, April 2009, pp. 4045–4048.

Table 3: Comparison of fusion methods for NIST SRE 2010 set, all tuning parameters have been cross validated using NIST SRE 2008 development set.

	Training method	**EER (%)**	**MinDCF (×100)**	**ActDCF (×100)**	$\frac{\|\boldsymbol{w}_{\text{reg}}\|_1}{\|\boldsymbol{w}\|_1}$	Ensemble size
itv-itv	Best Single (GSV-MFCC)	5.45	2.72	3.65		1
	no regularization	3.55	1.81	2.84	1	12
	VB	3.51	1.78	2.87		10
	subset sel.	3.40	1.75	2.61		6
	ridge	3.40	1.70	2.51	0.96	12
	LASSO	**3.33**	**1.69**	**2.23**	0.96	6
	Restricted LASSO	3.40	1.71	2.52		8
	E-net $\beta = 0$	3.40	1.70	2.50	0.96	12
itv-tel	Best Single (JFA-PLP)	3.03	1.39	1.75		1
	no regularization	2.40	0.98	1.74	1.0	12
	VB	2.50	**0.97**	2.00		12
	subset sel.	2.31	1.06	**1.34**		7
	ridge	2.40	**0.97**	1.65	0.86	12
	LASSO	2.40	0.99	1.63	0.71	8
	Restricted LASSO	**2.25**	2.36	3.45		7
	E-net $\beta = 0.7$	2.37	**0.97**	1.47	0.66	10
mic-mic	Best Single (JFA-PLP)	6.52	3.04	3.14		1
	no regularization	5.10	2.35	4.14	1.0	12
	VB	5.10	**2.22**	4.88		9
	subset sel.	**4.80**	2.30	3.08		8
	ridge	5.10	2.30	3.04	0.66	12
	LASSO	5.62	2.44	3.23	0.56	3
	Restricted LASSO	5.67	2.36	3.45		4
	E-net $\beta = 0.7$	4.82	2.30	**3.03**	0.51	6
tel-tel	Best Single (JFA-PLP)	3.62	1.58	1.74		1
	no regularization	2.33	1.12	**1.18**	1.0	12
	VB	**2.23**	1.14	3.04		12
	subset sel.	2.43	1.25	1.27		6
	ridge	2.33	1.14	1.28	0.91	12
	LASSO	2.25	1.19	1.27	0.91	5
	Restricted LASSO	2.27	**1.11**	1.19		12
	E-net $\beta = 0.1$	2.42	1.15	1.32	0.81	12