Variational Bayes logistic regression as regularized fusion for NIST SRE 2010

Ville Hautamäki, Kong Aik Lee, Anthony Larcher, Tomi Kinnunen, Haizhou Li

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, Finland.
[2]Human Language Technology Department, Institute for Infocomm Research, Singapore.
{villeh, tkinnu}@cs.joensuu.fi
{kalee, alarcher, mabin, hli}i2r.a-star.edu.sg

Abstract

Fusion of the base classifiers is seen as the 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 it. 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 the subset of classifiers that has a good generalization performance. We approach the problem from the sparse learning point of view. We assume that the true, but unknown, fusion weights are actually sparse. As a practical solution we regularize the weighted logistic regression loss function by the Elastic-Net constraint. Though sparse solutions can be easily obtained using the so-called least absolute shrinkage and selection operator (LASSO), but it does not take into account high correlation between classifiers. Elastic-Net, on the other hand, is a compromise between LASSO and ridge regression constraints. While ridge regression cannot produce sparse solutions, Elastic-Net can. By using sparseness enforcing constraint we are able to improve over the un-regularized solution in all but tel-tel condition.

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 then called the *classifier fusion*. In fusion, binary classifier is trained on the base classifier scores to make the accept or reject decision. The base classifiers might utilize, for instance, different speech parameterizations (e.g. spectral, prosodic or high-level features), classifiers (e.g. Gaussian mixture models [2] or support vector machines [3]) or channel compensation techniques (e.g. joint factor analysis [4] or nuisance attribute projection [5]).

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 learned classifier generalizes to an unseen data [6]. It is well known that 0/1-loss, where classification error is directly optimized, can lead to a serious overfit. In addition, finding the global optimum of 0/1-loss is an NP-complete computational problem [7]. The *hinge loss*, also known as maximum margin, and *logistic regression* have been proposed to tackling these deficiencies, by optimizing the upper bound of the 0/1-loss instead of the classification error itself.

Logistic regression loss defines an unconstrained convex programming problem, meaning that the global optimum can be found easily by iterative schemes [6]. In addition, logistic regression loss has similar generalization features as the maximum margin in the SVM. Logistic regression has been applied to the speaker verification score fusion task [8]. Later it was popularized by the *fusion and calibration* (FoCal) toolkit. It has subsequently been found to be usefull linear fusion training methodology by a number of independent studies (e.g. [9, 10, 11]) and is taken here as a reference method.

Overfitting on the training data is still possible, even though upper bound is optimized instead of 0/1-loss. To avoid overfit, a regularization is required. Most common one is the quadratic regularization $\frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$ also known as the *ridge regression* [12]. Regularization forces parameter shrinkage, where the greater the Lagrange coefficient λ is, smaller the norm $\|\boldsymbol{w}\|_1$ will be. Smaller norm implies better generalizability. Reason for it is also 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 the evaluation data.

When the ensemble has a large number of classifiers, it is expected that some of them will not play any role in a successful ensemble. So, it would be beneficial to remove some badly performing classifiers from the ensemble and thus reduce the prediction variance [13]. We have recently studied whether FoCal-based fusion can be improved by computing optimal weights and bias for all subsets and then selecting the one subset that gives best performance based on the training set [14]. We noticed, in oracle experiments, that classifier selection can significantly improve performance if suitable selection criterion is utilized. Our proposal was to use 0/1-loss as the selection criterion, this turned out not to generalize well. In addition, selection does not necessarily shrink the norm of the weight vector.

In contrast to the ridge regression, other approach is to regularize via the sum of absolute values $\lambda \sum_i |w_i|_1$, which is called *least absolute shrinkage and selection operator* (LASSO) [13]. It shrinks all coefficients, where some are forced to exactly zero. By regularizing weighted logistic regression with LASSO constraint, we can simultaneously optimize fusion weights and perform classifier subset selection. The convex combination (by parameter α) of ridge regression and LASSO leads to a regularization technique known as the Elastic-Net [15], which is believed to be sharp on the zeroing capability and at the same time 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 influence of Ridge constraint.

The works of T. Kinnunen and V. Hautamäki were supported by the Academy of Finland.

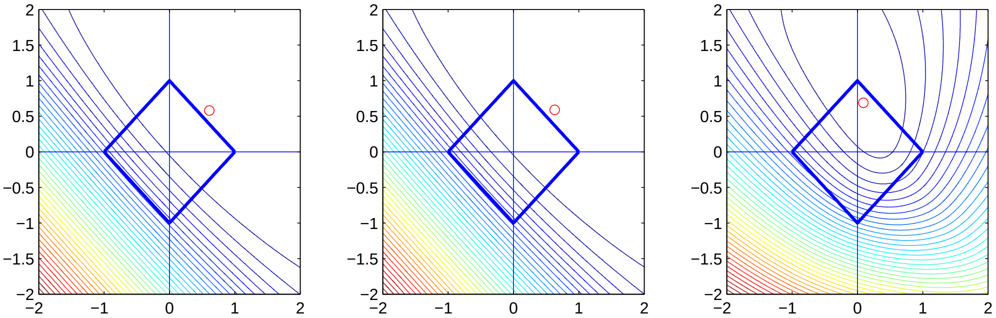


Figure 1: The C_{wlr} objective plotted as a function of two weights for systems $\{3, 11\}$. Included in the plots are the LASSO constraint (diamond) and the minimum (circle). Training set (right), Evalset 1 (center) and Evalset 2 (left). **This figure is a place-holder right now, Tomi will find a better one.**

In our previous work [16], we proposed to use LASSO and Elastic-Net regularization techniques to simultaneously achieve generalizable fusion device and classifier subset selection. By doing so we have proposed a method to train the subset selector by optimizing the weighted logistic regression loss.

and in this work we ...

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 $(\mathbf{w}^*, \theta^*)$ of a linear combiner $f_{\mathbf{w},\theta}(\mathbf{s}) = \mathbf{w}^t\mathbf{s} + \theta$ so that a classification error measure is minimized. We adopt the *detection cost function* (DCF) used in the NIST speaker recognition evaluations,

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

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 $\text{logit}\, P = \log P - \log(1 - P)$. It is possible to minimize DCF directly (e.g. [17]) or to optimize a surrogate cost such as effective prior weighted logistic regression cost [18].

2.2. Logistic regression

Here shortly about logistic regression model, where important point is that quantity $\mathbf{w}^t\mathbf{s}_i + w_0$ is log-odds and that optimizing the negative log-likelihood leads to the cross-entropy objective.

2.3. Weighted cross-entropy objective

In the speaker verification applications, we are usually interested in a specific set of DCF parameters, by so doing we are operating in a cost-sensitive learning. 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 **[correct ref]**, indirect optimization of said parameters is achieved by modifying the cross-entropy objectiveC_{wlr} . Modification weights cost by effective prior and the observed ratio of positive and negative examples.

$$\begin{aligned} C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) &= \frac{P}{N_t} \sum_{i=1}^{N_t} \log\left(1 + e^{-\boldsymbol{w}^t\boldsymbol{s}_i + \theta'}\right) \\ &+ \frac{1 - P}{N_f} \sum_{j=1}^{N_f} \log\left(1 + e^{\boldsymbol{w}^t\boldsymbol{s}_j - \theta'}\right), \quad (3) \end{aligned}$$

where the two sums go through the N_t target score vectors $\boldsymbol{s}_i$ and the 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. Here, P is the effective prior defined in subsection 2.1 and $\theta' = -\,\text{logit}(P)$ is the decision threshold which is determined from the pre-set cost parameters P_{tar}, C_{miss} and C_{fa}.

Due to the cross-entropy being convex in which the C_{wlr} loss function is optimized. We use iterative gradient descent method

3. Regularized Logistic Regression

We extend the weighted logistic regression in Eq. (3), by adding a regularization term. It leads to minimizing [6],

$$C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) \quad \text{s.t.} \quad J(\boldsymbol{w}) \leq t, \quad (4)$$

where $J(\boldsymbol{w})$ is either $\frac{1}{2}\|\boldsymbol{w}\|_2^2$, which is called ridge regression or $\|\boldsymbol{w}\|_1$, and LASSO as well. 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,

$$C_{\text{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda\|\boldsymbol{w}\|_1. \quad (5)$$

It is known that the larger λ, the more norm $\|\boldsymbol{w}\|$ will be shrunk [13]. If the optimization is based on the Eq. (5), 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 which 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,

$$C_{\mathrm{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda \left(\alpha \|\boldsymbol{w}\|_1 + (1-\alpha)\|\boldsymbol{w}\|_2^2\right). \qquad (6)$$

As can be seen, Eq. (6) is a generalized variant of both LASSO and ridge regression, we can always find such a α 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 crossvalidate over a 2-d space. In this work we use the methodology, where α parameter is first fixed and then shrinkage factor can be cross validated as in LASSO and Ridge. In practice, α will also be cross validated in such a way that best α and shrinkage factor will be selected based on cross validation set to be applied on the evaluation set.

Depending on the chosen regularization method, there are different strategies to optimize (4). Since logistic regression using quadratic regularization is differentiable, it can be efficiently optimized using standard packages [6]. Situation is not so simple for LASSO regularization. In [13], a *quadratic programming* (QP) solution was proposed to it by rewriting the constraints in (4) to a more convenient form. However, more recent techniques are faster in practice, for that reason we apply *projectionL1* algorithm [19] that optimizes the Lagrangian form Eq. (5). We apply the same method to Elastic-Net, as, sum of two convex functions is still convex, we can minimize $C_{\mathrm{wlr}}(\boldsymbol{w}, \boldsymbol{s}) + \lambda(1-\alpha)\|\boldsymbol{w}\|_2^2$, given $\lambda\alpha\|\boldsymbol{w}\|_1$ as the constraint.

3.1. Bayesian interpretation

Regularized logistic regression can be seen as a MAP estimate [13] of logistic regression, where prior is Gaussian, in the case of Ridge regression, and Laplacian, in the case of LASSO. **add more details**

3.2. Variational Bayes fusion

Using *automatic relevance determination* (ARD) in the fully Bayesian logistic regression, we can generate sparse solutions *without* cross-validating any regularization parameters. **add more details**

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[1]. 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 was used as pre-calibration method), fusion weights and bias. The second part, *cross validation set*, is used to estimate shrinkage parameter (λ) and tradeoff between LASSO and Elastic-Net (α). Parameters are then applied to NIST 2010 data, which serves as 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 effective prior in Eq. (2). And we are interested in comparing the application dependent classification error as measured in actual DCF (ActDCF).

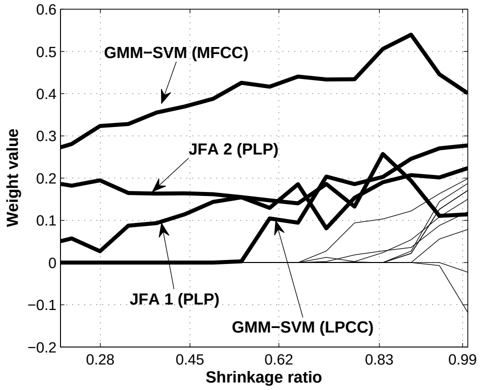


Figure 2: Weight evolution of the LASSO regularization as a function of normalized t.

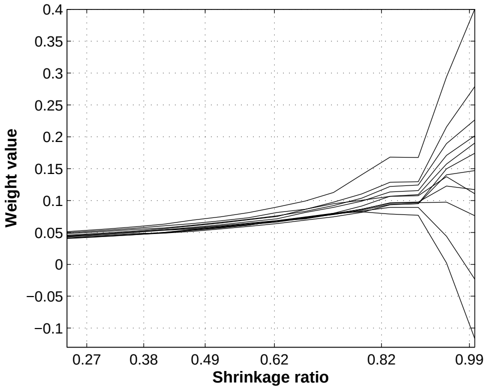


Figure 3: Weight evolution of the ridge regression regularization as a function of normalized t.

In this study we use the same ensemble setup as in our previous work [14]. We have twelve subsystems in total, all are based on different cepstral features and four different classifiers, as part of the of the I4U system. When subsystems share the same classifier and features, it means that the systems are independent implementations. For classifiers, we use the generative GMM-UBM-JFA [4] and the discriminative GMM-SVM approaches with KL-divergence kernel [20] and the recently proposed Bhattacharyya kernel [21]. We also include another recent method, feature transformation [22], 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.

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

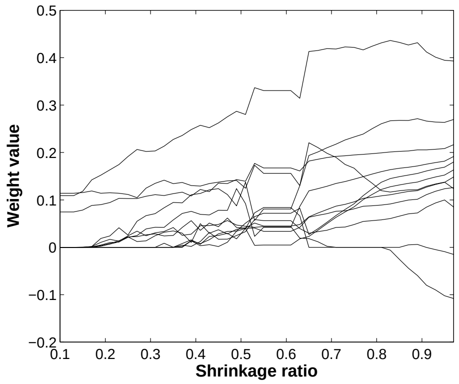
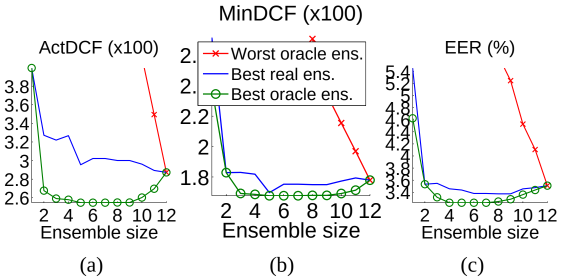


Figure 4: Weight evolution of the Elastic-net regularization, with $\alpha = 0.7$, as a function of normalized t.



(a) (b) (c)

Figure 5: Effect of ensemble size to accuracy (Evalset 2) 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 2 (NIST SRE 2010). The middle (blue) line indicates the actual ensemble selected by cross-validation Evalset 1.

5. Experiments

It is instructive to show the evolution of the individual classifier weights as the function of threshold parameter t. In Figs. 2, 3 and 4 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. It can be noticed immediately that ridge regression tends to group all classifiers to similar weights as the norm is shrunk towards zero. The grouping effect and the lack of it in the LASSO is known in the general regression literature [15]. Ridge regression tends to group together classifiers that are correlated. LASSO on the other hand tends to select few classifers per group. Selection is evident in Fig. 2, as very quickly only four classifiers are left in the ensemble, namely GMM-SVM using MFCC and LPCC front-end and two JFA systems with PLP front-end (base classifiers $\{1, 2, 6, 7\}$). It is notable that even though both JFA base classifers use same features, they are different implementations, even using different programming languages, also data sets used for learning factor loading matrices are different.

Regularization path of the Elastic-Net solutions on the other hand shows grouping effect, it appears to group classifiers into 4 different groups with the $\alpha = 0.7$ 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 (×100)	ActDCF (×100)	Ensemble size
itvitv	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
itvtel	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
micmic	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
teltel	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 is shown in Table 1. As in logistic regression weighting was used but in VB approaches effective prior information was not used it was assumed that in terms of DCF logistic regression would be the winner. However, in all but tel-tel condition VB approaches won in terms of minDCF. Our way of using VB logistic regression did not output well calibrated scores (at least in contrast to the logistic regression).

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

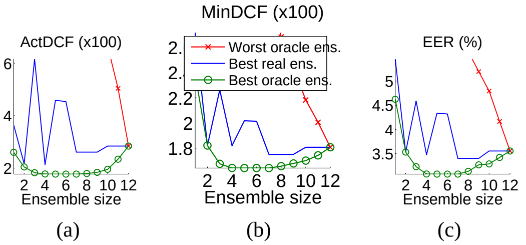
It is interesting to note that both VB and VB-ARD approaches do infact zero out some base classifiers, but only for the case of itv-itv and mic-mic conditions. Table 1 also shows that difference between VB and VB-ARD is negligible. We do not consider ARD further.

As noted in the Table 1, VB approaches underfit, in terms of finding the sparse solution. However, we can utilize subset selection methodology, where VB solution is found for all subsets of base classifiers. Results for it are shown in In Fig. 5, where best real solution is chosen based on the Evalset 1 (shown in blue) and applied to Evaset 2. Best and worst oracle bound are also shown. As a comparison we show also same experiment when logistic regression was used instead of VB in Fig. 6. We notice that VB provides much more stabile performance as function of subset size.

In Table 2, subset size is also selected by cross-validation from Evalset 1. There is an improvement over full ensemble methodsm except in ActDCF. Using subset selection ensemble size was further reduced from 10 to 6.

Table 2: Variational Bayes using subset selection applied to Evalset 2. Cross-validation is performed on Evalset 1.

	Fusion	EER (%)	MinDCF (×100)	ActDCF (×100)	Ensemble size
itvitv	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



(a) (b) (c)

Figure 6: 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.

5.2. Results with all core conditions

In Table 4 we show recognition results for the NIST SRE 2010 sub-conditions (itv-itv, itv-tel, mic-mic and tel-tel). We define baseline method to signify an unregularized solution (i.e. $\lambda = 0$), equivalent to the implementation of the FoCal toolkit, but with a different optimizer. Best single classifier is selected based on the performance on the cross validation set, so all methods are directly, and fairly, comparable in Table 4. We notice that for itv-tel and mic-mic subcondition (for both males and females) Elastic-Net achieves the best results, in terms of Actual DCF. It is interesting to note that improvement in Actual DCF is because scores are better calibrated after fusion with Elastic-Net than with other methods.

General trend, when comparing minDCF over all conditions seems to be that there are no large differences in except in the mic-mic condition where FoCal clearly fails. Differences in Actual DCF are the mostly the product of different calibrations. Noting here that the bias is *n*ot regularized.

It is interesting to note that predicting the α value using cross validation set is not a trivial problem. It is clear that in the case when either LASSO or Ridge won over Elastic-Net in terms of Actual DCF the prediction of α was unsuccessful. Especially interesting is the case itv-itv female, where prediction gave $\alpha = 0$ (i.e. Ridge) and in the NIST SRE 2010 LASSO was clearly better.

Regularization, however, does not bring improvement in tel-tel condition. For the tel-tel condition, designers of base classifiers had a very large and extensively used corpora available where to tune their systems. In addition, selection of data sets for the estimation of session compensation parameters is more straight forward. But the interview and microphone data conditions did not have such a wealth of material backing their classifier design. Then it is expected that regularization will hurt the classification performance in the tel-tel condition. In other conditions, significant improvement over the baseline can be achieved by all regularization methods. Ridge regression and Elastic-Net obtain in general best performance, where best one according to cross validation results come from Elastic-Net (in five out of eight sub conditions).

5.3. Avoiding sparsity

As λ is increased LASSO tends to zero out large number of base classifiers. In this section we are interested to regularize the LASSO regularizer, in so words to make it less harsh. We can exclude any of the base classifiers from being zeroed out by the optimizer by adding an extra constraint $w_j \neq 0$ to the (4). Taking Lagrange formulation of the constrained optimization problem, we give λ for $w_j \neq 0$ constraint. We assume that said λ is the same as one for the LASSO, then λ equals to zero for base classifier j.

In Table 3, cross-validation was used to select which classifier not to regularize per condition. We notice that not regularizing one classifier does not lead to ensemble size being increased by one classifier, in the case of itv-tel ensemble size 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 full ensemble.

Not regularizing one classifier helps in the itv-tel condition, where EER improves from 2.40% to 2.25%. Lowest minDCF in tel-tel condition is obtained using this configuration.

Table 3: Restricting LASSO by not regularizing one base classifier on Evalset 2. Regularization parameter and non-regularizing base classifier selected using Evalset 1.

Condition	EER	MinDCF	ActDCF	Ensemble
	(%)	($\times$100)	($\times$100)	size
itv-itv	3.40	1.7135	2.5198	8
itv-tel	2.25	1.1631	1.4407	7
mic-mic	5.67	2.3601	3.4493	4
tel-tel	2.27	1.1074	1.1922	12

6. Conclusions

We have studied regularized logistic regression fusion on the NIST SRE 2010 core test sub conditions. We find that regularization brings improvement over unregularized variant when the development set and evaluation set (NIST SRE 2010) are so closely matched.

As a future work we plan to extend the variational Bayes approach used in this paper to the Elastic-Net regularized logistic regression.

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] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] S. Ben-David, N. Eiron, and P. Long, “On the difficulty in approximately maximizing agreements,” *Journal of Computer and System Sciences*, vol. 66, no. 3, pp. 496–514, 2003.

Table 4: 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 ($\times$100)**	**ActDCF ($\times$100)**	$\frac{\|\boldsymbol{w}_{\mathrm{reg}}\|_1}{\|\boldsymbol{w}\|_1}$	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
	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
	E-net $\alpha = 0$	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
	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
	E-net $\alpha = 0.7$	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
	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
	E-net $\alpha = 0.7$	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
	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
	E-net $\alpha = 0.1$	2.42	1.15	1.32	0.81	12

[8] 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.

[9] 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.

[10] 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.

[11] T. Kinnunen, J. Saastamoinen, V. Hautamäki, M. Vinni, and P. Fränti, “Comparative evaluation of maximum *a Posteriori* vector quantization and Gaussian mixture models in speaker verification,” *Pattern Recognition Letters*, vol. 30, no. 4, pp. 341–347, March 2009.

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

[13] 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.

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

[15] 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.

[16] 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.

[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] N. Brümmer and J. Preez, “Application-independent evaluation of speaker detection,” *Computer Speech and Language*, vol. 20, pp. 230–275, April-July 2006.

[19] 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.

[20] 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.

[21] 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.

[22] 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.