ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MAXIMUM A POSTERIOR LINEAR REGRESSION WITH ELLIPTICALLY SYMMETRIC MATRIX VARIATE PRIORS

Wu Chou

Bell-Labs – Lucent Technologies
600 Mountain Ave., Murray Hill, NJ 07974, USA
wuchou@research.bell-labs.com

10.21437/Eurospeech.1999-4

ABSTRACT

In this paper, elliptic symmetric matrix variate distribution is proposed as the prior distribution for maximum a posterior linear regression (MAPLR) based model adaptation. The exact close form solution of MAPLR with elliptically symmetric matrix variate priors is obtained. The effects of the proposed prior in MAPLR are characterized and compared with conventional maximum likelihood linear regression (MLLR). The proposed priors are significant informative priors, through which a well-founded Bayesian theoretical framework is formulated to incorporate prior information in model adaptation. Moreover, an efficient approach of hyperparameter estimation in MAPLR is described. Experimental results indicate that significant gain can be obtained when adaptation data are sparse.

1. INTRODUCTION

Channel and speaker variations are most noticeable factors which affect the performance of speech recognition systems. Many methods are proposed to compensate the variations in the signal. Maximum likelihood linear regression (MLLR) based transformation is an effective method for speaker and channel adaptation[1]. MLLR is based on a group of linear transformations which transform the model parameters and maximize the model likelihood on the adaptation data. The group of transformations in MLLR are derived from the training data and each transformation can be designed to control certain portion of the model. This is because the adverse effects from acoustic mismatch may not be uniform and exhibit significant variations among different model units. MLLR is one of the most widely used techniques in speech recognition and applied in various applications with significant performance improvements.

One of the problems associated with adaptation is the sparseness of the adaptation data. When adaptation data are sparse, maximum likelihood estimation often gives biased inaccurate estimate. Maximum A Posterior (MAP) based adaptation is a powerful approach with its root deep in the fundamental Bayesian formalism and has been applied successfully in speech recognition[7]. The conventional maximum likelihood (ML) estimation is to find model parameter λ such that

$$\lambda_{ML} = \underset{\lambda}{\operatorname{argmax}}\, g(\lambda \mid x) = \underset{\lambda}{\operatorname{argmax}}\, f(x \mid \lambda), \qquad (1)$$

where $f(x \mid \lambda)$ is the likelihood of observing data x. In MAP estimation, an appropriate prior is used, and the MAP estimate is given by

$$\lambda_{MAP} = \underset{\lambda}{\operatorname{argmax}}\, f(x \mid \lambda) g(\lambda). \qquad (2)$$

where the prior distribution $g(\lambda)$ characterizes the distribution of the model parameter λ. MAP estimation provides a way to incorporate prior knowledge into the model parameter estimation. The relation between ML and MAP estimation is based on the Bayes' theorem where the posterior distribution $p(\lambda \mid x) \propto f(x \mid \lambda) g(\lambda)$.

In this paper, we apply MAP estimation to linear regression (MAPLR) based model transformation for robust model adaptation. Unlike the conventional MAP based HMM parameter estimation, many known and widely used prior distributions, such as normal-Washart priors for HMM parameters, do not have a close form solution in MAPLR under their EM formulation[4]. Therefore, the exact effects of priors in linear regression based model adaptation need to be derived. In this paper, we first prove that there is a very rich class of prior distributions, i.e. elliptically symmetric matrix variate distribution, in which MAPLR has a close form solution. From this fundamental result, the effects of MAPLR with elliptically symmetric priors are derived for HMMs with mixture Gaussian observation densities. The novel contributions of this paper are:

- The elliptically symmetric matrix variate distributions is proposed as the prior distribution for maximum a posterior linear regression (MAPLR) in model adaptation.

- A close form exact solution of MAPLR under the proposed prior is derived from its EM (Expectation-Maximization) equation for both full matrix and diagonal based transforms.
- The effects of prior distribution in MAPLR are characterized and compared with the conventional MLLR solution.
- An efficient approach of hyperparameter estimation in MAPLR is proposed and experimental results are given which indicate the efficacy of the proposed approach.

2. MAPLR WITH ELLIPTIC SYMMETRIC PRIORS

In MAPLR framework, the group of linear transformations are estimated according to equation (2). Instead of maximizing the likelihood function $f(x \mid \lambda)$ as in MLLR, MAPLR is to estimate the group of linear transformations which maximize the posterior probability. Although MAP solution is often more desirable because of its posterior nature, the use of prior makes the problem much more difficult. This is because MAP solution is strongly dependent on the form of priors being used. The common practice in Bayesian analysis is to use conjugate priors so that the posterior distribution $p(\lambda \mid x)$ has the same functional form as priors, and the problem is reduced to estimate the parameters in a known form of density. Unfortunately, there is no joint conjugate prior densities which can be specified for mixture Gaussian distributions[7]. Therefore, finding a class of informative and yet solvable prior distributions for MAPLR becomes crucial.

In order to state the results, we need to fix some notations, and we leave details of the proof in the future full paper. We adopt the same notation used to derive MLLR for the purpose of easy comparison [2]. For linear transformation based adaptation of HMMs with mixture Gaussian observation densities, the adapted mean in mixture component s in the mixture Gaussian density is given by

$$\hat{\mu_s} = W_s \xi_s \tag{3}$$

where W_s is an $n \times (n+1)$ matrix, and ξ_s is an extended mean vector (we consider general offset case) $\xi_s = [1, \mu_{s_1}, ..., \mu_{s_n}]'$.

The auxiliary Q-function under EM algorithm for MAPLR with prior matrix distribution $p(W_s)$ is given by

$$Q(\lambda, \bar{\lambda}) = a + logp(W_s)$$

$$+ \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_t(n,m) log(o_t \mid W_s, \mu_{n,m}, \Sigma_{n,m}^{-1}), \tag{4}$$

where a is a term not relevant to W, $\gamma_t(n,m) = P(s_t = n, l_t = m \mid O, \lambda)$ is the probability of being in state n and mixture m at time t given observation sequence $O = \{o_1, ..., o_T\}$. To reduce the complexity of notation, we consider the following relevant part in Q-function:

$$Q(\lambda, \bar{\lambda}) = logp(W_s)$$

$$+ \sum_{t,n,m} \gamma_t(n,m)(o_t - W_s\xi_s)'\Sigma_s^{-1}(o_t - W_s\xi_s), \tag{5}$$

where s is a generic index which is a function of (t, n, m).

Let $X = (X_{i,j}) = (X_1, ..., X_n)'$, where $X_i = (x_{i,1}, ..., x_{i,p})$, be a random matrix with $E(X_i) = \mu_i = (\mu_{i,1}, ..., \mu_{i,p})'$.

Definition: A $N \times P$ random matrix X with values $x \in R^{N \times P}$ is said to have a distribution belonging to the family of elliptically symmetric matrix variate distribution with location parameter $\mu = (\mu_1, ..., \mu_N)'$ and scale matrix $\Delta = diag(\Sigma_1, ..., \Sigma_N)$ if its probability density function can be written as

$$f_X(x) = (det\Delta)^{-1/2} q(\sum_{i=1}^{N} (x_i - \mu_i)'\Sigma_i^{-1}(x_i - \mu_i)) \tag{6}$$

where q is a function on $[0, \infty)$ of the sum of N quadratic forms $(x_i - \mu_i)'\Sigma_i^{-1}(x_i - \mu_i)$, $i = 1, ..., N$.

For a given matrix W, we denote the i-th row of W by $W(\underline{i})$ and the j-th column of W by $W(\mid j)$. Following the notation in MLLR solution, let $Z^{(i)}$ and $G^{(i)}$ be Z and G matrices related to the solution of $W(\underline{i})$. The following results are proved.

Theorem 1: If the prior distribution of the transformation matrix W_s is elliptically symmetric with q being an exponential function, then MAPLR has a close form solution and the transformation matrix W_s is given by

$$W_s(\underline{i})' = (\hat{G}^{(i)})^{-1}\hat{Z}(\underline{i})' \tag{7}$$

$$= (G^{(i)} + \Sigma_i^{-1})^{-1}(Z(\underline{i}) + \mu_s^i\Sigma_i^{-1}) \tag{8}$$

where

$$\hat{G}^{(i)} = (G^{(i)} + \Sigma_i^{-1}) \tag{9}$$

and

$$\hat{Z}(\underline{i}) = Z(\underline{i}) + \mu_s^i\Sigma_i^{-1} \tag{10}$$

are corresponding G and Z matrixes for the i-th row $W_s(\underline{i})$ in MAPLR solution, μ_s^i and Σ_i are the corresponding mean vector and the covariance matrix in the elliptic symmetric matrix variate distribution of (6).

Theorem 1 establishes the close form relation between MLLR and MAPLR with elliptic symmetric matrix variate priors. Although there is no conjugate prior distribution for mixture Gaussian distribution, the linear transformation based adaptation for continuous density HMM with mixture

Gaussian observation densities still has a close form solution under elliptic symmetric matrix variate priors. Moreover, the family of elliptic symmetric matrix variate priors is a very rich class of priors, and these priors are meaningful and informative. In fact, based on (9), elliptic symmetric priors did exactly what we would hope, namely to improve the condition of $G^{(i)}$ matrices in MLLR. These $G^{(i)}$ matrices in MLLR can become singular or very close to be singular, when the adaptation data are sparse. This is a problem which plagues the application of MLLR for small amount of adaptation data. This close form solution also formulates a well founded Bayesian theoretical framework to incorporate prior knowledge in linear transformation based adaptation.

Corresponding results are also derived for MAPLR with diagonal linear transformations. As a special case of the full matrix transformation, it also has a close form solution in MAPLR under elliptic symmetric matrix variate priors. The approach used in the proof of *Theorem 1* also extends to other variants of elliptic symmetric matrix variate distributions.

3. PRIOR HYPERPARAMETER ESTIMATION

When an informative prior distribution is used, additional parameters are needed to describe the prior distribution. These additional parameters are called hyperparameters. In a strict Bayes approach, hyperparameters in the prior density is assumed known based on a common or subjective knowledge about the stochastic process. But in most cases, these hyperparameters cannot be derived from the subjective knowledge and alternative approaches are needed. One popular solution to adopt is based on empirical Bayes (EB) approach in which the prior parameters are also estimated from the data. However in EB approach, additional data points are often required, and ML estimation based on the EB moment estimator can be rather difficult to solve. In linear regression transform based model adaptation, data points for each regression class can be extremely sparse and computationally expensive to obtain.

In this section we describe an approach to derive prior hyperparameters which is quite efficient and works with sparse adaptation data. This approach is tailored to the functional form of the elliptic symmetric matrix variate priors and integrated with the structure of MAPLR transformations. The group of linear transformation in MAPLR is closely related to the model structure, since it is trying to cope with non-uniform acoustic-phonetic variations caused by the environment. A typical approach is to form a group of linear transformations based on a tree of broad phonetic classes, where each linear transformation is associated with a tree node and it transforms all states on that tree node during MAPLR adaptation[2]. One of the advantages of using elliptic symmetric matrix variate priors is that they are related to so called location-scale family. Once the location parameters are determined and the scale factors σ are sufficiently large, they are quite robust and the likelihood function will quickly become dominant when more adaptation data are available. In order to avoid estimate location parameter matrices directly, we use the global MLLR transform from the adaptation data as the location parameter and only estimate scale parameter matrices from the data. This procedure consists of the following steps.

1. Generate a global transform for the location parameters for elliptic symmetric priors.
2. Generate a group of MLLR transform using a lower sample count threshold.
3. For each tree node with sample count above the threshold of having a linear transform in MAPLR, estimate the scale parameter matrices of the prior if the number of lower sample count transforms under this tree node is above a threshold. Otherwise, a scale parameter matrix based on all lower sample count transforms is used in the prior.

Since we would like adaptation become quickly dominated by data, this reduces the requirement on the scale factor matrices and it can be estimated from just a few transforms.

4. EXPERIMENTS

The speech recognition experiments were performed on a controlled environment using an internal database of non-native speaker telephone band speech of RM (resource management) sentences. It was collected from five non-native speakers recording simultaneously through two channels: a close talking microphone and a telephone line. For each speaker, the database contains 300 sentences for training and enrollment purposes from each speaker. This database were used extensively for earlier adaptation experiments[3]. The speaker independent model was trained on the standard DARPA RM speaker independent training corpus. The training data were based on the bandpass filtered 4KHz speech signal obtained from the original 8KHz wideband speech training data. The bandpass filtered signal was down sampled at 8KHz and passed through a pre-emphasis filter to flatten the signal spectrally. A 30 msec Hamming window with 10 msec shift was used. A 10-th order auto-correlation analysis was performed to compute the LPC derived cepstral analysis. The feature vector used in recognition has 39 parameters, including 12 liftered cepstral coefficients, 12 delta cepstral coefficients, 12 delta-delta cepstral coefficients, and the energy, delta and delta-delta energy features.

# of Adaptation Sentence	MLLR	MAP_LR	Relative Err Change %
0	18.6	18.6	0
20	17.7	15.4	-13.5
30	13.9	12.5	-10.1
40	13.4	12.7	-5.2
50	11.3	10.8	-4.4
75	10.6	10.2	-3.7
100	9.6	9.4	-2.1
150	7.9	7.8	-1.5
300	5.8	5.7	-1.7

Table 1: Batch supervised adaptation on microphone database

The acoustic model was built from the bandpass filtered data using decision tree based state tying[8]. A broad phonetic class tree was derived from the decision tree state tying and used to construct a group of linear transforms [2]. Two sample counts were used. One was for generating a true MAPLR linear transform and another one was significantly lower for prior scale factor estimation. In order to reduce the number of hyperparameters, diagonal scale factor matrices Σ_i were used. The prior estimation was based on the procedure described in the previous section, and the number of data sample points needed for prior scale factor can be set very low. A lower bound threshold of two was used in the experiments. If the number of lower sample count transformations under that tree node is below two, a back-off scale factor matrix based on all lower sample count transforms was used. The sample count for a MAPLR transform was set to 600 and lower sample count to generate extra transforms for prior estimation was set to 200. The number of MAPLR transforms were determined by the distribution of adaptation data, and there was no fixed hard limit. Table 1 depicts the experimental results of batch supervised adaptation using the enrollment sentences in the training portion of the non-native speaker database. The training adaptation sentences are very short with an average duration of about 4 seconds. The baseline SI model performance is about 18.6%, and after 300 sentence adaptation, the performance is around 5.7%. It is interesting to note that the effect of prior in linear transformation based adaptation was quite significant when the number of training samples was low, and the effects of prior tapered off smoothly as more adaptation data were added. For very sparse adaptation data, MAPLR can make a significant performance difference, and with more training data its performance did not degrade. This may due part to the very conservative priors used in our approach.

5. SUMMARY

In this paper, we studied maximum a posterior linear transformation (MAPLR) for model adaptation. The elliptic symmetric matrix variate distribution was proposed as the prior distribution for MAPLR. The close form solution of MAPLR under the proposed prior distribution were derived. The effects of the proposed priors in MAPLR were characterized and compared with the conventional MLLR solution. An efficient approach of prior hyperparameter estimation in MAPLR was described. Prior selection and hyperparameter estimation are perhaps two most tricky issues in Bayesian analysis. Since they are related to the subjective knowledge, they can vary from case to case for sparse data samples, but for large data samples, a transformation obtained from MAPLR solution will converge to its MLLR solution. We will address these properties in the future paper.

6. REFERENCES

[1] C. J. Leggetter and P. C. Woodland (1995), "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Speech Communication*, **Vol. 9**, pp. 171-185.

[2] C. J. Leggetter and P. C. Woodland (1994), "Speaker Adaptation of HMMs Using Linear Regression" *CUED/F-INFENG/TR.181*

[3] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," IEEE Trans. on Speech and Audio Processing, 1996.

[4] C. Chesta, O. Siohan and C.-H. Lee, "Maximum A Posterior Linear Regression for Hidden Markov Model Adaptation", to appear*Eurospeech'99*.

[5] M. DeGroot, "Optimal Statistical Decision", em McGraw-Hill, 1970.

[6] N. Giri, "Multivariate Statistical Analysis", em Marcel Dekker, 1996.

[7] J.-L. Gauvain, C.-H. Lee, *'Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Ma rkov Chains'*, IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, 1994.

[8] W. Reichl, W. Chou, *'Decision Tree State Tying Based on Segmental Clustering for Acoustic Modeling'*,ICAS SP 98, Seattle, pp. 801-804, May 1998.