Bayesian Speaker Verification with Heavy-Tailed Priors

Patrick Kenny

Odyssey Speaker and Language Recognition Workshop

June 29, 2010

Introduction

- The low dimensional i-vector representation of speech segments opens up new possibilities for using Bayesian methods in speaker recognition.
- You can go a long way with Bayesian methods under the assumptions that speaker and channel effects are **Gaussian** and **statistically independent**.
- You can do better by relaxing the Gaussian assumption. In particular, it seems to be possible to do away with score normalization in speaker verification.
- The directional scattering behavior which appears to explain the success of cosine distance scoring in speaker recognition can be modeled by relaxing the statistical independence assumption. (This part of the talk is speculative.)

Joint Factor Analysis with i-vectors

Each recording is represented by a single vector of dimension F known as an i-vector (e.g. $F = 400$).

Given a speaker and i-vectors $D_1, \ldots, D_R$ we assume

$$D_r = S + C_r$$

where

- R is the number of recordings of the speaker, indexed by r
- S depends on the speaker, C_r depends on the channel
- S and C_r are statistically independent (**?**)
- S and C_r are multivariate Gaussian (**?**)

To begin with, we assume both statistical independence and Gaussianity.

Probabilistic Linear Discriminant Analysis

PLDA

Under Gaussian assumptions, this model is known in face recognition as PLDA [Prince and Elder].

The between-speaker covariance matrix is $\mathrm{Cov}(S, S)$.

The within-speaker covariance matrix is $\mathrm{Cov}(C_r, C_r)$ (assumed to be independent of r).

If the feature dimension F is high, these matrices cannot be treated as full rank.

Hidden variable formulation of Gaussian PLDA

Assume that there are low dimensional, normally distributed hidden variables x_1 and x_{2r} such that

$$D_r = m + U_1 x_1 + U_2 x_{2r} + \epsilon_r .$$

The residual ϵ_r is normally distributed with mean 0 and precision matrix Λ (typically diagonal).

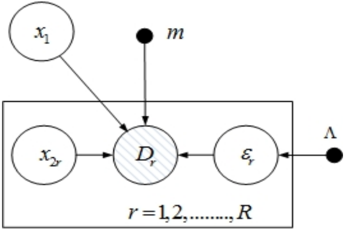
- m is the center of the acoustic space
- x_1 depends only on the speaker (speaker factors)
- the columns of U_1 are the eigenvoices

$$\mathrm{Cov}(S, S) = U_1 U_1^*$$

- x_{2r} varies from one recording to another (channel factors)
- the columns of U_2 are the eigenchannels

$$\mathrm{Cov}(C_r, C_r) = \Lambda^{-1} + U_2 U_2{}^* .$$

Graphical model for Gaussian PLDA



Including x_{2r} enables the decomposition

$$\mathrm{Cov}(C_r, C_r) = \Lambda^{-1} + U_2 U_2^*.$$

This is not needed in all cases and, under Gaussian assumptions, x_{2r} can **always** be eliminated at recognition time (later).

Working assumptions

Assume for the time being that

- we have succeeded in estimating the model parameters (m, U_1, U_2, Λ)
- given a collection $D = (D_1, \ldots, D_R)$ of i-vectors associated with a speaker, we have figured out how to evaluate the marginal likelihood ("the evidence")

  $$P(D) = \int P(D, h)\mathrm{d}h$$

  where h represents the entire collection of hidden variables associated with the speaker.

We will show how to do speaker recognition in this situation and how both of these problems can be tackled by using variational Bayes to approximate the posterior distribution $P(h|D)$.

PLDA speaker recognition

Given two i-vectors D_1 and D_2 suppose we wish to perform the hypothesis test

$$\begin{aligned} H_1 &: \text{The speakers are the same} \\ H_0 &: \text{The speakers are the different.} \end{aligned}$$

The likelihood ratio is

$$\frac{P(D_1, D_2|H_1)}{P(D_1|H_0)P(D_2|H_0)}.$$

Every term here is an evidence integral of the form

$$\int P(D, h)\mathrm{d}h.$$

The likelihood ratio for any type of speaker recognition or speaker clustering problem can be written down just as easily.

Approximating the evidence

The evidence $P(D)$ can be evaluated exactly in the Gaussian case but this involves inverting large (sparse) block matrices [Prince and Elder].

Even in the Gaussian case, it is more efficient to use a variational approximation: if $Q(h)$ is any distribution on h and

$$\mathcal{L} = \mathbb{E}\left[\ln \frac{P(D,h)}{Q(h)}\right]$$

then

$$\mathcal{L} \leq \ln P(D)$$

with equality iff $Q(h) = P(h|D)$ [Bishop].

Variational Bayes provides a principled way of finding a good approximation $Q(h)$ to $P(h|D)$.

Why are posteriors generally intractable?

There is nothing mysterious about the posterior $P(h|D)$. By Bayes rule,

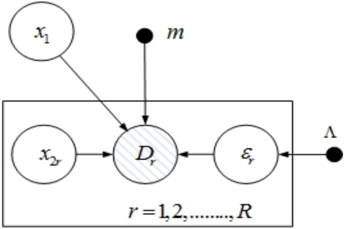
$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}.$$

The problem in practice is that the normalizing constant $1/P(D)$ — the evidence — cannot be evaluated:

$$P(D) = \int P(D|h)P(h)\mathrm{d}h.$$

Another way of stating the difficulty is that whatever factorizations (i.e. statistical independence assumptions) exist in the prior $P(h)$ are destroyed in the posterior by taking the product $P(D|h)P(h)$.

Example



x_1 and x_{2r} are independent in the prior:

$$P(x_{2r}|x_1) = P(x_{2r})$$

but this is not true if D_r is observed. Since D_r depends on both x_1 and x_{2r}, knowing x_1 changes the conditional distribution of x_{2r}:

$$P(x_{2r}|D_r, x_1) \neq P(x_{2r}|D_r).$$

Variational Bayes

Set $x_2 = (x_{21}, \ldots, x_{2R})$. x_1 and x_2 are independent in the prior but they are correlated in the posterior.

The idea in variational Bayes is to **force** independence in the posterior so that we seek an approximation to $P(x_1, x_2|D)$ of the form

$$Q(x_1, x_2) = Q(x_1)Q(x_2).$$

$Q(x_1)$ and $Q(x_2)$ are estimated by the standard update formulas

$$\begin{aligned} \ln Q(x_1) &= \mathbb{E}_{x_2}\left[\ln P(D, x_1, x_2)\right] + \text{const} \\ \ln Q(x_2) &= \mathbb{E}_{x_1}\left[\ln P(D, x_1, x_2)\right] + \text{const} \end{aligned}$$

which do not involve inverting large block matrices and are guaranteed to increase the value of $\mathcal{L}$ on each iteration [Bishop].

The variational posterior is also the key to estimating the model parameters

The model parameters (m, U_1, U_2, Λ) are estimated by maximizing the evidence criterion

$$\sum_s \mathcal{L}(s)$$

where s ranges over all of the speakers in a training set.

We defined

$$\mathcal{L} = \mathbb{E}\left[\ln \frac{P(D,h)}{Q(h)}\right].$$

It is convenient to decompose this as

$$\mathcal{L} = \mathbb{E}\left[\ln P(D|h)\right] - \mathrm{KL}\left(Q(h)||P(h)\right).$$

Note that the second term does not involve the model parameters at all.

Maximum likelihood estimation

As a first step towards maximizing the evidence, we sum the contributions from the first term over all training speakers

$$\sum_{s} \mathbb{E}\left[\ln P(D(s)|h(s))\right]$$

and maximize this with respect to the model parameters. We refer to this as maximum likelihood estimation.

This expression is formally the same as the EM auxiliary function in probabilistic principal components analysis [Bishop].

The only difference is that the expectations are evaluated with the variational posteriors rather than the true posteriors.

Minimum divergence estimation

Another way to increase the value of the objective function is to find affine transformations of the model parameters and hidden variables

$$\begin{aligned} (m, U_1, U_2, \Lambda) &\rightarrow (m', U_1', U_2', \Lambda') \\ h(s) &\rightarrow h'(s) \end{aligned}$$

which preserve the value of the EM auxiliary function but minimize the sum of divergences

$$\sum_s \mathrm{KL}\left(Q'(h'(s) \| P(h(s))\right).$$

We refer to this as minimum divergence estimation.

Example

Applying this to the speaker factors $x_1(s)$ amounts to finding an affine transformation such that the posterior moments of $x_1'(s)$ agree with those of the prior on average:

$$
\begin{aligned}
\frac{1}{S}\sum_s \mathrm{Cov}(x_1'(s), x_1'(s)) &= I \\
\frac{1}{S}\sum_s \mathbb{E}\left[x_1'(s)\right] &= 0
\end{aligned}
$$

where S is the number of speakers in the training set. The model parameters are then updated by applying the inverse transformation to m and U_1 so as to preserve the value of the EM auxiliary function.

Interleaving maximum likelihood and minimum divergence estimation helps to accelerate convergence.

Variational Bayes allows other possibilities to be explored . . .

These estimation procedures are adequate for speaker recognition but hard-core Bayesians would avoid point estimates of the model parameters altogether.

For example, it is possible to put a prior on U_1 and calculate a posterior with variational Bayes.

In theory, even the number of speaker factors could be treated as a hidden variable, rather than a parameter that has to be manually tuned. (Analogous to the the treatment of the number of mixture components in Bayesian estimation of Gaussian mixture models [Bishop].)

There is an extensive literature on Bayesian principal components analysis.

Replacing the Gaussian distribution by Student's *t*

The Student's *t* distribution is a heavy tailed distribution in the sense that the density $P(x)$ has the property that there is a positive exponent k such that

$$P(x) = O(\|x\|^{-k})$$

as $\|x\| \to \infty$. Compare this with the Gaussian distribution:

$$P(x) = O(e^{-\|x\|^2/2}).$$

Like the Gaussian distribution, the Student's *t* distribution is unimodal but it is less susceptible to some well known problems of Gaussian modeling:

- The Gaussian assumption effectively prohibits large deviations from the mean ("Black Swans")
- Maximum likelihood estimation of a Gaussian (i.e. least squares) can be thrown off by outliers.

Definition of Student's *t* suitable for variational Bayes

The Student's *t* distribution can be represented as a continuous mixture of Gaussians, using a construction based on the Gamma distribution.

The Gamma distribution $\mathcal{G}(a, b)$ is a unimodal distribution on the positive reals whose density is given by

$$P(u) \quad \propto \quad u^{a-1} e^{-bu} \quad (u > 0).$$

The parameters *a* and *b* enable the mean and the variance to be adjusted independently:

$$\begin{aligned} \mathbb{E}\left[u\right] &= a/b \\ \mathrm{Var}(u) &= a/b^2. \end{aligned}$$

A random vector x has a Student's t distribution with n degrees of freedom, mean μ and scale parameter Λ (roughly speaking, the precision matrix) if

$$x \sim \mathcal{N}(\mu, (u\Lambda)^{-1}), u \sim \mathcal{G}(n/2, n/2)$$

where $\mathcal{N}$ indicates the normal distribution and $\mathcal{G}$ the Gamma distribution.

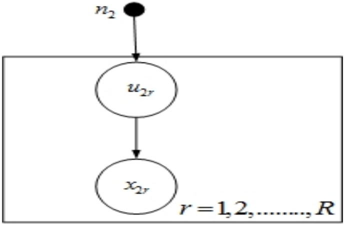
At one extreme ($n \to \infty$) the variance of u is 0 and this reduces to the Gaussian distribution.

At the other ($n = 1$), this reduces to the Cauchy distribution. This is so heavy-tailed that the variance and all higher order moments are infinite.

The term "degrees of freedom" comes from classical statistics. It doesn't mean anything in particular in this context.

Example

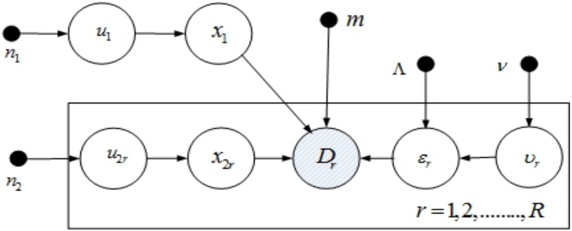
To make the distribution of the channel factors x_{2r} heavy-tailed, introduce a scalar hidden variable u_{2r}:



$$x_{2r} \sim \mathcal{N}(0, u_{2r}^{-1} I) \text{ where } u_{2r} \sim \mathcal{G}(n_2/2, n_2/2)$$

Graphical model for heavy-tailed PLDA

In heavy-tailed PLDA all of the hidden variables x_1, x_{2r} and ϵ_r have Student's *t* distributions:



Extra hidden variables: u_1, u_{2r} and υ_r.

Extra parameters: the numbers of degrees of freedom n_1, n_2 and ν.

Variational Bayes carries over straightforwardly

Heavy-tailed PLDA is fully diagonalizable — only diagonal matrices need to be inverted for variational Bayes (see the paper).

If $n_2 = \nu$, the channel factors x_{2r} can be eliminated at recognition time and variational Bayes converges very quickly.

Idea: Let P be the orthogonal matrix whose columns are the eigenvectors of the effective covariance matrix $\Lambda^{-1} + U_2 U_2{}^*$. Rotate both the model and the data by P to obtain an equivalent model with a diagonal residual precision matrix and no channel factors.

The numbers of degrees of freedom n_1, n_2 and ν can be estimated (by divergence minimization) using the evidence criterion.

Gaussian vs. Student's *t* on telephone data

	Gaussian EER/DCF	Student's *t* EER/DCF
short2-short3	3.6% / 0.014	**2.2%** / **0.010**
8conv-short3	3.7% / 0.009	**1.3%** / **0.005**
10sec-10sec	16.4% / 0.070	**10.9%** / **0.053**

- NIST 2008 SRE **English** language **female** data
- EER = equal error rate, DCF = 2008 detection cost function
- unnormalized likelihood ratios

The effect of score normalization

	Gaussian EER/DCF	Student's *t* EER/DCF
short2-short3	2.7% / 0.013	**2.4%** / **0.012**
8conv-short3	1.5% / 0.009	**0.8%** / **0.007**
10sec-10sec	13.3% / 0.063	**12.8%** / **0.066**

- likelihood ratios normalized with *s*-norm (see the paper)
- helpful in the Gaussian case, harmful for Student's *t*
- one exception (EER 0.8% for 8conv-short3), not statistically significant
- even with normalization, Student's *t* is better than Gaussian

Score normalization is usually needed in order to set a trial-independent decision threshold for speaker verification. It is typically fragile and computationally expensive.

A good generative model for speech should produce likelihood ratios which do not need to be normalized (or even calibrated).

Score normalization is needed in practice because outlying recordings tend to produce exceptionally low scores for all of the trials in which they are involved.

In the Student's *t* case, the hidden variables u_1, u_{2r} and υ_r seem to model these outliers adequately.

The curious case of microphone speech

For **telephone speech**

- Gaussian PLDA with score normalization gives results which are comparable to cosine distance scoring
- Heavy-tailed PLDA without score normalization gives better results. The error rates are about 25% lower than for traditional Joint Factor Analysis.

For **microphone speech** heavy-tailed PLDA breaks down in an interesting way.

In a companion paper, we describe an i-vector extractor suitable for speaker recognition with both microphone and telephone speech ($F = 600$).

Using only telephone speech, we first trained a model without channel factors

$$D_r = m + U_1 x_1 + \epsilon_r$$

with a full precision matrix for the residual ϵ_r.

We augmented this with channel factors trained only on microphone speech:

$$D_r = m + U_1 x_1 + U_{2r} x_{2r} + \epsilon_r.$$

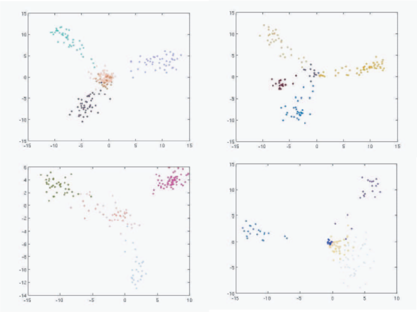
It turns out that the microphone channel factors are **Cauchy** distributed ($n_2 < 1$).

Speaker recognition breaks down unless the model is constrained (by flooring the number of degrees of freedom) so that microphone transducer effects have finite variance.

In this case, Student's *t* modeling is no better than Gaussian.

Perhaps the best course would be to project away the troublesome dimensions using linear discriminant analysis (classical LDA or PLDA).

How can this behavior be modeled probabilistically?



Plots of GMM supervectors for different speakers (Tang, Chu and Huang, ICASSP 2009) illustrating directional scattering. The *x* and *y* axes are essentially the first two i-vector components.

Caution: The directional scattering effect in these plots is probably exaggerated. The utterances are **conversation turns** (which may be very short), not conversation sides, and the authors use relevance MAP, not eigenvoice MAP, to estimate supervectors. An artifact of relevance MAP is that the longer an utterance, the further the supervector from the origin.

Still, directional scattering seems to be the only possible explanation for the success of cosine distance scoring in speaker recognition. (But does anybody know how to account for it?)

There appears to be a principal axis of session variability which varies from one speaker to another. This is **inconsistent** with the assumption that speaker and session effects are additive and statistically independent.

A proposal for modeling directional scattering

Instead of representing a speaker by a single point x_1 in the speaker factor space, we could represent the speaker by a **distribution** specified by a mean vector μ and a precision matrix Λ. (This is different from the $F \times F$ residual precision matrix previously denoted by Λ.)

i-vectors are generated by sampling "speaker factors" from this distribution:

$$D_r = m + U_1 x_{1r} + U_2 x_{2r} + \epsilon_r.$$

The trick is to choose a prior $P(\mu, \Lambda)$ in which μ and Λ are **not** statistically independent. Since point estimation of μ and Λ is hopeless, it is necessary to integrate over μ and Λ with respect to the prior $P(\mu, \Lambda)$. Do this with variational Bayes.

The Normal-Wishart prior

To generate observations for a speaker: **First**, generate an $N \times N$ precision matrix Λ (where N is the dimension of the speaker factors) by sampling from a standard Wishart prior

$$\Lambda \sim \mathcal{W}(I, \tau).$$

The Wishart distribution with parameters W and τ, $\mathcal{W}(W, \tau)$, is a generalization of the Gamma distribution. It is concentrated on positive definite $N \times N$ matrices Λ and its density is given by:

$$P(\Lambda) \propto |\Lambda|^{(\tau - N - 1)/2} \exp\left(-\frac{1}{2}\mathrm{Tr}\left(W^{-1}\Lambda\right)\right).$$

τ is the number of degrees of freedom: the larger τ, the more peaked the distribution. There is no loss in generality assuming a standard Wishart prior, $W = I$ (other possibilities could be accommodated by modifying U_1).

Next, generate the mean vector μ for the speaker by sampling from a Student's *t* distribution with scale parameter Λ and hidden variable *w*:

$$\mu \sim \mathcal{N}(0, (w\Lambda)^{-1}), w \sim \mathcal{G}(\alpha/2, \beta/2).$$

where, like τ, α and β are parameters to be estimated. We will get to the question of why a Student's *t* distribution is needed here in a moment. (Again, there is no loss of generality in taking the mean of this Student's *t* distribution to be 0.)

Since the distribution of μ depends on Λ and w, Λ and μ are not statistically independent in the prior:

$$P(\Lambda|\mu) \neq P(\Lambda)$$

so there is some hope of modeling speaker-dependent directional scattering.

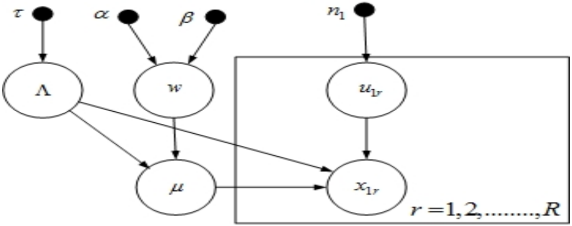
Finally, generate “speaker factors” x_{1r} **for each the speaker’s recordings** by sampling from a Student’s *t* distribution with mean μ and scale parameter Λ:

$$x_{1r} \sim \mathcal{N}(\mu, (u_{1r}\Lambda)^{-1}), u_{1r} \sim \mathcal{G}(n_1/2, n_1/2).$$

The difference between the “speaker factors” x_{1r} and the channel factors x_{2r} is that the distribution of x_{2r} is assumed to be **speaker-independent** whereas the model for x_{1r} includes **speaker-dependent** hidden variables Λ, μ and w.

Thus x_{1r} models both speaker variability and a particular type of session variability.

Graphical model for x_{1r}



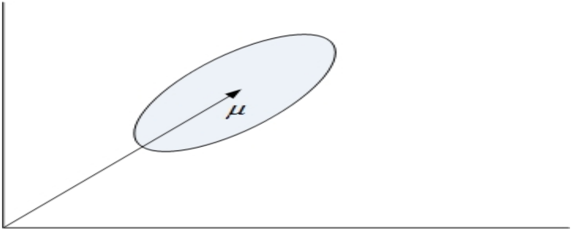
The values of the parameters α, β and τ determine whether this model exhibits speaker-dependent directional scattering or not.

How can this capture directional scattering?

Compare the speaker-dependent distribution of the covariance matrix Λ^{-1} with the speaker-independent distribution:

$$\mathbb{E}\left[\Lambda^{-1}|w,\mu\right] = \frac{\tau - N - 1}{\tau - N}\mathbb{E}\left[\Lambda^{-1}\right] + \frac{1}{\tau - N} w\mu\mu^{*}.$$

The effect of the second term (a rank 1 covariance matrix) is to augment the variance in direction of the speaker's mean vector μ:



Because $\mu\mu^*$ is weighted by $\frac{1}{\tau - N} w$, the strength of this effect depends on τ and the parameters α and β which govern the distribution of w:

$$\begin{aligned} \mathbb{E}\left[w\right] &= \alpha/\beta \\ \mathrm{Var}(w) &= 2\alpha/\beta^2. \end{aligned}$$

If α and β are such that the mean of w is large and the variance is small, there is marked directional scattering for most speakers. (This flexibility is achieved by taking $P(\mu|\Lambda)$ to be Students *t* rather than Gaussian.)

On the other hand if $\beta = \tau^{-1}$ and $\tau \rightarrow \infty$, this model can be shown to reduce to heavy-tailed PLDA and there is no speaker-dependent directional scattering.

How well this works remains to be seen . . .

Conclusion

Gaussian PLDA is an effective model for speaker recognition, even though it is based on questionable assumptions, namely that speaker and channel effects are additive, statistically independent and normally distributed.

These assumptions can be relaxed by adding hidden variables u_1, u_{2r}, υ_r to model outliers and μ, Λ, w to model directional scattering.

The derivation of the variational Bayes update formulas is mechanical and, because variational Bayes comes with EM-like convergence guarantees, implementations can be debugged.

Caveat: In practice, to do the variational Bayes derivations, the prior distributions of the hidden variables need to be in the exponential family. For example, if you try to do heavy-tailed PLDA with the text book definition of Student's *t*, you will get nowhere.

I believe that in order to get the full benefit of Bayesian methods, you need **informative** priors whose parameters can be elicited from training data. (This view is sometimes referred to as "empirical Bayes".)

For example, the priors on the additional hidden variables u_1, u_{2r}, υ_r and μ, Λ, w — namely the degrees of freedom n_1, n_2, ν and τ, α, β — can be estimated from training data using the evidence criterion.

References for Bayesian Methods

- S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.
- C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer Science+Business Media, LLC, 2006.
- P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors” in *Proc. Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010. `http://www.crim.ca/perso/patrick.kenny`