# TIED NORMAL VARIANCE–MEAN MIXTURES FOR LINEAR SCORE CALIBRATION

*Sandro Cumani, Pietro Laface*

sandro.cumani@polito.it, pietro.laface@polito.it

## ABSTRACT

A speaker verification system decides whether two voice segments belong to the same speaker based on a threshold. An optimal threshold can be set if the recognition scores are well calibrated, i.e., they represent Log–Likelihood Ratios. Logistic Regression (LogReg) is a standard approach for score calibration. While training this discriminative model requires labeled scores, Gaussian and non–Gaussian generative calibration models have been recently proposed. They not only have similar or better performance with respect to LogReg, but also allow for unsupervised or semi–supervised training of the models.

The goal of this work is to extend these models. In particular, we show that normal variance–mean mixture distributions are able to model well–calibrated non–Gaussian distributed scores, provided that their parameters for the target and non–target score distributions are properly tied. As for the Gaussian case, a linear calibration model can then be estimated by computing Maximum Likelihood estimates of the distributions parameters and of the score transformation. The quality of all these approaches has been compared on a dataset of segments of variable duration obtained by cutting the NIST 2010 evaluation test data.

***Index Terms*—** score calibration, likelihood ratio interpretation, linear score calibration models

## 1. INTRODUCTION

Score calibration aims at transforming the scores produced by a system so that each score can be interpreted as the Log–Likelihood Ratio (LLR) between the hypotheses that two voice segments belong to the same or to different speakers. If a score is a LLR, the optimal decision depends only on the score, and on the prior probability of the same–speaker (target) and different–speaker (non–target) trials.

The standard approach for score calibration is based on discriminative prior–weighted Logistic Regression (LogReg) [1], which optimizes the expected value of the logarithmic proper scoring rule [2] assuming a linear calibration model. LogReg has been widely investigated, and is routinely employed as a calibration tool in different fields [3, 4, 5, 6].

Alternative methods based on generative models have recently gained interest [7, 8]. The theoretical properties of log–likelihood ratios have been analyzed in [7] showing that, for a score to represent a LLR, it should verify the so called "LLR of the LLR is the LLR" property. This LLR property has strong implications on the admissible distributions that can generate calibrated scores. Additionally, a Maximum Likelihood (ML) approach for linear calibration has been proposed, referred to as Constrained Maximum–Likelihood Gaussian (CMLG) calibration, based on the assumption that the scores are Gaussian distributed. In contrast with discriminative approaches, generative calibration models allow for unsupervised or semi–supervised training of calibration models. The CMLG approach has been extended in [9] considering a set of unlabeled scores as samples of a two–components GMM, whose components satisfy the LLR property. The result is an unsupervised linear score calibration model that has shown to be effective for several tasks [10, 11, 12].

More recently, non–Gaussian generative models have been proposed for non–linear calibration [8, 13]. In particular, [8] analyzes different models based on non–Gaussian distributions, including T–student and Normal Inverse Gaussian (NIG) distributions [14]. Rather than estimating an explicit score mapping to fit theoretically calibrated score distributions as in [7], this approach estimates a probabilistic model in score space, and computes the LLRs by evaluating the likelihood ratio between the hypotheses that a score was generated by the target or by the non–target distribution, respectively. Although good results have been obtained using NIG score models, the approach of [8] does not guarantee that the NIG learned transformation is monotonic, as assured instead by CMLG.

The goal of this work is to extend CMLG [7]. We show that normal variance–mean mixture distributions, of which NIG is a subclass, are able to model non–Gaussian distributed scores, provided that the parameters for the target and non–target score distributions are properly tied. The parameters of a linear calibration model can be estimated, as for CMLG, by computing ML estimates of the distributions parameters and of the score transformation.

The rest of the paper is organized as follows. Section 2 recalls the CMLG calibration approach. Section 3 highlights the limits of CMLG, and proposes a mixture density model that is able to generate well–calibrated, non Gaussian–distributed, scores. The experimental results are illustrated in Section 4, and conclusions are given in Section 5.

## 2. CMLG SCORE CALIBRATION

Since our work extends the findings of [7], in this section we briefly recall the properties of LLRs that lead to the CMLG score normalization model.

We consider a verification system that computes the log–likelihood ratio of a trial as:

$$x = LLR(e) = \log \frac{P(e|H_T, \mathcal{M})}{P(e|H_F, \mathcal{M})} , \tag{1}$$

where $e$ is the evidence associated to the trial extracted by the recognizer (e.g., a pair of i–vector or speaker embeddings [15, 16]), $H_T$ and $H_F$ are the target and non–target trial hypotheses, respectively, and $\mathcal{M}$ is a statistical model for $e$, such as Probabilistic Linear Discriminant Analysis (PLDA) [17, 18].

It has been shown that the log–likelihood ratio between the likelihood of observing a particular value of $x$, given the target and the non–target hypotheses, is again $x$ [7, 19]:

$$x = LLR(e) = \log \frac{P(LLR(e) = x|H_T, \mathcal{M})}{P(LLR(e) = x|H_F, \mathcal{M})} . \tag{2}$$

The scores of a set of independent trials are well calibrated if they approximately satisfy this constraint [7]. In the following we will refer to (2) as the LLR property or constraint.

Let's consider the distribution of the LLRs of the target and non–target trials. We consider, thus, the scores as samples of a random variable $X$, whose conditional distributions given the target ($T$) and the non–target ($F$) classes are $f_{X|T}$ and $f_{X|F}$, respectively. The two distributions are closely related because from (2) we get:

$$f_{X|T}(x) = e^x \cdot f_{X|F}(x) . \tag{3}$$

If $X|F$ is Gaussian distributed, (3) implies that also $X|T$ is Gaussian distributed, and that the parameters of the two distribution are related as [7]:

$$f_{X|T}(x) = \mathcal{N}(x|\mu, 2\mu) , \quad f_{X|F}(x) = \mathcal{N}(x|-\mu, 2\mu) , \tag{4}$$

where $\mu$ is the mean of the target Gaussian. CMLG estimates a linear calibration model that transforms scores so that they fit the theoretical well–calibrated distributions (4).

## 3. MIXTURE DENSITIES FOR CALIBRATION

CMLG provides good results as long as the distributions of the target and non–target trial scores are approximately Gaussian, with common variance. This is often not the case. Consider a system that is able to produce well–calibrated, Gaussian–distributed, LLRs for each condition $k$ of a discrete set of homogeneous classes. A class may consist, for example, of trials including speech segments from the same channel. Due to the LLR constraint (2), target and non–target distributions are described by a single parameter $\mu_k$. It was shown in [7] that $\mu_k$ is directly related to the system EER. Since it is reasonable assuming that heterogeneous conditions may lead to different system accuracy, the distributions of the scores of each condition will be characterized by different values of $\mu_k$. Thus, even in this ideal case, we cannot expect that *pooled* scores of all $K$ distributions follow a Gaussian distribution.

The generative process for sampling a score with such a system can be described as: (i) sample an experimental condition $k = 1, \ldots, n$ from random variable $K$ with probability $P_K(k) = w_k$, and (ii) sample target and non–target scores from $X|(T, k)$ and from $X|(F, k)$, respectively.

Assuming that trials can be divided into a discrete set of homogeneous conditions is often not possible. For example, if we consider utterance durations and noise levels as the main factors that affect the recognizer accuracy, we expect that conditions vary continuously. Thus, modeling conditions with continuous, rather than with discrete, random variables is more accurate.
Let a continuous random variable $V$, with density $g_V(v)$, be responsible for the selection of sub–condition $v$. Let $f_{X|T,V}(x|v)$ and $f_{X|F,V}(x|v)$ denote the conditional distributions, parametrized by $v$, for target and non–target scores. The PDFs $f_{X|T}(x)$ and $f_{X|F}(x)$ are obtained by marginalizing over $v$ the joint densities:

$$f_{X|T}(x) = \int_v f_{X|T,V}(x|v) g_V(v) dv ,$$
$$f_{X|F}(x) = \int_v f_{X|F,V}(x|v) g_V(v) dv \tag{5}$$

If the conditional distributions $f_{X|T,V}(x|v)$ and $f_{X|F,V}(x|v)$ satisfy the LLR property, so that:

$$f_{X|T,V}(x|v) = e^x f_{X|F,V}(x|v) , \tag{6}$$

then $f_{X|T}$ and $f_{X|F}$ also satisfy the LLR property because

$$\log \frac{f_{X|T,V}(x|v)}{f_{X|F,V}(x|v)} = \log \frac{e^x \int_v f_{X|F,V}(x|v) g_V(v) dv}{\int_v f_{X|F,V}(x|v) g_V(v) dv} = x .$$

The distributions (5) are not directly tractable. Thus, we assume that the conditional distribution for $f_{X|T,V}$ and for $f_{X|F,V}$ is Gaussian. We also assumed that $f_{X|T,V}$ and $f_{X|F,V}$ satisfy the LLR property, thus:

$$f_{X|T,V} = \mathcal{N}(x|\mu(v), 2\mu(v)) , \; f_{X|F,V} = \mathcal{N}(x|-\mu(v), 2\mu(v)) , \tag{7}$$

where parameter $\mu$ is a function of the (unobserved) parameter $v$. Since we can freely specify the mixing distribution $g_V(v)$, without loss of generality, we set $\mu(v) = \frac{1}{2}v$, and we get the normal variance–mean mixture densities [20]:

$$f_{X|T}(x) = \int_v \mathcal{N}(x|\frac{1}{2}v, v) g(v) dv$$
$$f_{X|F}(x) = \int_v \mathcal{N}(x|-\frac{1}{2}v, v) g(v) dv \tag{8}$$

for the target and non–target distributions, respectively.

A generic univariate variance–mean mixture random variable $X$ can be represented as:

$$X = \mu + \beta V + \sqrt{V} Y \quad (9)$$

where $V$ and $Y$ are independent random variables, $Y$ has standard normal distribution, and the distribution of $V$ is $g_V(v)$. The PDF of $X$ is given by

$$f_X(x) = \int_v \mathcal{N}(x|\mu + \beta v, v) g_V(v) dv \,. \quad (10)$$

Let's consider the conditional distributions:

$$X|T = \mu_T + \beta_T V + \sqrt{V} Y \,, \; X|F = \mu_F + \beta_F V + \sqrt{V} Y \,. \quad (11)$$

To obtain the PDFs in (8) requires setting:

$$\mu_T = \mu_F = 0 \,, \quad \beta_T = \frac{1}{2} \,, \quad \beta_F = -\frac{1}{2} \,. \quad (12)$$

Normal variance–mean mixtures (11) with parameters defined by (12) satisfy the LLR constraint, and are thus suitable for representing well–calibrated scores. However, normal variance–mean mixtures are difficult to handle. Thus, we restrict our analysis to the class of the Generalized Hyperbolic distributions (GH) [21, 22, 23]. GH distributions are obtained as normal variance–mean mixtures where the mixing distribution $g_V(v)$ is the Generalized Inverse Gaussian (GIG) [24]. The result is a 5–parameter family of distributions, with PDF:

$$GH(x|\lambda, \alpha, \beta, \delta, \mu) = Z(\lambda, \alpha, \beta, \delta) \left[\delta^2 + (x-\mu)^2\right]^{\frac{\lambda - \frac{1}{2}}{2}} \cdot e^{\beta(x-\mu)} K_{\lambda - \frac{1}{2}} \left(\alpha \sqrt{\delta^2 + (x-\mu)^2}\right) \,, \quad (13)$$

with $\lambda$, $\delta > 0$, $\alpha > |\beta| \in \mathbb{R}$. $K_\nu$ denotes the modified Bessel function of the third kind with index $\nu$, and $Z(\lambda, \alpha, \beta, \delta)$ is the normalization constant. Please notice that a GH distribution with parameters $\alpha \to \infty$, $\delta \to \infty$, and $\frac{\alpha}{\delta} = \sigma^2$ converges to the Gaussian distribution $\mathcal{N}(\mu + \beta\sigma^2, \sigma^2)$ [25].

From (12), taking into account that the mixing density is the same for both the target and non–target scores, we obtain a suitable model for well–calibrated scores by setting the parameters of the GH distributions as:

$$f_{X|T}(x) = GH(x|\lambda, \alpha, \frac{1}{2}, \delta, 0)$$
$$f_{X|F}(x) = GH(x|\lambda, \alpha, -\frac{1}{2}, \delta, 0) \,, \quad (14)$$

where the parameters $\lambda$, $\alpha$ and $\delta$ are shared.

The subclass of GH, with $\lambda = -\frac{1}{2}$, is the Normal Inverse Gaussian (NIG) distribution, which was successfully used to model target and non–target scores in [8, 13]. In particular, in [8] independent NIG distributions are used to model score distributions, and likelihood ratios are then computed in score space from these distributions.

Our proposed approach is different: we show that GH (and thus NIG) distributions of (14) are a natural solution for obtaining well–calibrated scores, provided that the distribution parameters are tied. We calibrate the scores by transforming the observed score densities to fit theoretically well–calibrated GH distributed scores. This approach is similar to CMLG [7], but rather than targeting Gaussian distributions, we target the more flexible family of GH distributions.

As for CMLG, a linear calibration model can be obtained by ML, estimating the distributions that fit the original scores, assuming that the scores distributions are linear transformation of the theoretical well–calibrated distributions.

Since GH distributions are closed under affine transformations [26], if $X \sim GH(\lambda, \beta, \alpha, \delta, \mu)$, then:

$$aX + b \sim GH(\lambda, \frac{\alpha}{|a|}, \frac{\beta}{a}, \frac{\delta}{|a|}, a\mu + b) \,. \quad (15)$$

Assuming the linear calibration model:

$$x(s) = as + b \,, \quad a > 0 \,, \quad (16)$$

and defining the random variables that generated the observed scores, $S|T = \frac{1}{a}(X|T) - \frac{b}{a}$ and $S|F = \frac{1}{a}(X|F) - \frac{b}{a}$, we have that:

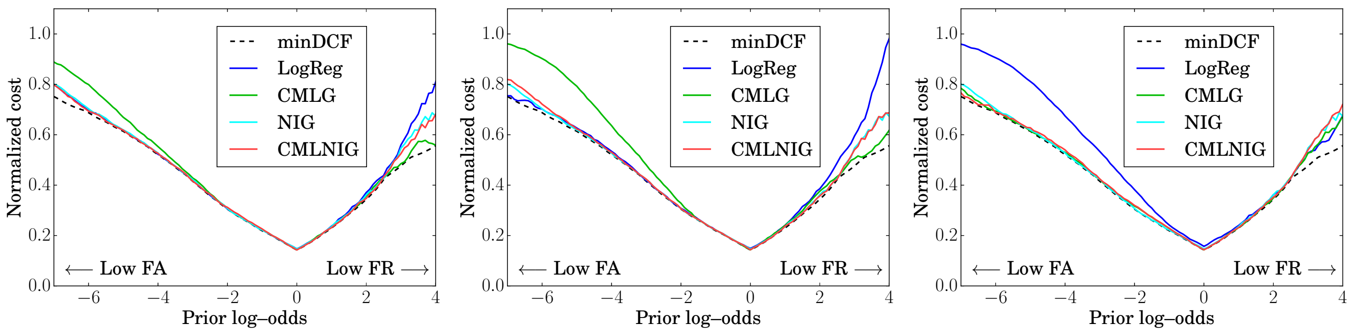
$$S|T \sim GH(\lambda, a\alpha, a\beta, a\delta, -\frac{b}{a}) \sim GH(\lambda, \overline{\alpha}, \overline{\beta}, \overline{\delta}, \overline{\mu}) \quad (17)$$
$$S|F \sim GH(\lambda, a\alpha, -a\beta, a\delta, -\frac{b}{a}) \sim GH(\lambda, \overline{\alpha}, -\overline{\beta}, \overline{\delta}, \overline{\mu}) \,,$$

where we set $\overline{\mu} = -\frac{b}{a}$, $\overline{\alpha} = a\alpha$, $\overline{\beta} = a\beta$, $\overline{\delta} = a\delta$. The parameters $\overline{\mu}$, $\overline{\alpha}$, $\overline{\beta}$, and $\overline{\delta}$ can be estimated by maximizing the weighted likelihood:

$$\mathcal{L} = \frac{\pi}{n_T} \sum_{i=1}^{n_T} GH(s_{T,i}|\lambda, \overline{\alpha}, \overline{\beta}, \overline{\delta}, \overline{\mu}) + \frac{1-\pi}{n_F} \sum_{i=1}^{n_F} GH(s_{F,i}|\lambda, \overline{\alpha}, -\overline{\beta}, \overline{\delta}, \overline{\mu}) \,, \quad (18)$$

where $n_T$ and $n_F$ denote the number of target and non–target training scores, $0 < \pi < 1$ is a tunable weight, and $s_{T,i}$ and $s_{F,i}$ denote the $i$–th target and non–target score, respectively. The ML solution can be found using a slight modification of the Expectation Maximization algorithm [27], which has to take into account weights and parameters relations during the M–step. The case of NIG distributions can be solved by modifying the simpler EM algorithm of [28]. Given the ML solution for (17), the calibration parameters can be finally obtained simply as:

$$a = 2\overline{\beta} \,, \quad b = -2\overline{\beta}\overline{\mu} \,. \quad (19)$$




(a) $\pi = 0.5$ (b) $\pi = 0.01$ (c) $\pi = 0.99$

**Fig. 1**: Normalized Bayes error rate for LogReg, CMLG, CMLNIG, and NIG methods as a function of a DCF operating point. The value of $\pi$ for the first three methods is reported in each figure.

## 4. EXPERIMENTS

In this section we compare our proposed method, restricted to NIG distributions (CMLNIG), with LogReg, CMLG and the NIG approach of [8]. Since we are interested in non–Gaussian distributed scores, we focus on variable duration utterances, and we analyze the scores of an i–vector based PLDA model that exploits the i–vector uncertainty [29, 30]. Indeed, in accordance with our analysis of Section 3, we observed that methods that consider the i–vector covariance have better accuracy, but their score distribution is "less Gaussian".

The system that has been used for the experiments is based on a gender independent, 2048 components, GMM with diagonal covariances, and on a gender–dependent i–vector extractor that produces 400–dimensional i–vectors and their covariances. The training set for the UBM consisted of NIST 04, 05 and 06 data. Switchboard 2 was added for training the i–vector extractor. The tests were performed on the female portion of SRE 2010 tel–tel extended condition, cutting short segments from 3 to 60 seconds. Calibration parameters were estimated on a subset of the NIST 08 female dataset, similarly cut.

Figure 1a shows the normalized Bayes error rate plot [31] for different systems. X–axis corresponds to different target prior log–odds $x = \log \frac{p}{1-p}$, where $p$ is a synthetic prior. Y–axis plots the corresponding normalized actual DCF. Parameter $\pi$ was set to $0.5$ for LogReg, CMLG and CMLNIG.

While our proposed approach achieves similar results to LogReg and NIG, CMLG–based calibration is less effective for low False Acceptance (FA) regions.
Since LogReg, CMLG and CMLNIG are sensible to the choice of $\pi$, a second set of experiments has been performed varying the values for $\pi$. Figures 1b and 1c show the results with $\pi = 0.01$ and $\pi = 0.99$, respectively. As expected, a low $\pi$ value, i.e., a low target prior, allows LogReg to slightly improve calibration for low FA regions, but it worsen calibration for low False Rejection region. On the contrary, for large $\pi$ values, LogReg fails in the low FA regions. Surprisingly, we observed the opposite behavior for CMLG. This is in contrast with the findings of [8], which showed that, for low $\pi$ values, CMLG provides better calibration in the low FA region. We believe that this is due to different distributions of the scores produced by our system and by the system used in [8]. Indeed, the target prior value is used in CMLG only to favor a better fitting of the target or of the non–target score distribution, rather than to give more weight the FAs, as it is the case for LogReg. This was confirmed by an experiment in which we artificially increased the variance of the target scores, and obtained the same behavior reported in [8].

Figures 1b and 1c show that our approach is less sensitive to proper tuning of parameter $\pi$, and that it provides similar results with respect to the NIG approach for a wide range of values of $\pi$.

## 5. CONCLUSIONS

We introduced tied normal variance–mean mixture distributions for modeling well–calibrated non–Gaussian score distributions. In this work we limited our experiments to the estimation of the parameters of tied NIG distributions so that they fit the original scores, assuming that the scores distributions are a linear transformation of the theoretical well–calibrated distributions. However, an alternative approach consists in learning a density transformation that directly transforms the score distributions to well–calibrated LLR distributions. This solution is appealing because it can be extended to define and estimate *monotonic non–linear* calibration models using the same approach that has been used in [32, 33] for transforming i–vectors. Furthermore, we believe that our approach is better suited to unsupervised training of the calibration models with respect to the NIG models of [8] because it has less parameters to estimate, and because tying helps identifying the small fraction of targets typically included in an unlabeled dataset.

## 6. REFERENCES


[1] N. Brummer, L. Burget, and al., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST Speaker Recognition Evaluation 2006," *Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 7, pp. 2072–2084, 2007.

[2] N. Brümmer and G. R. Doddington, "Likelihood-ratio calibration using prior-weighted proper scoring rules," in *Proceedings of Interspeech*, pp. 1976–1979, 2013.

[3] N. Brümmer, "Focal toolkit." Available at http://sites.google.com/site/nikobrummer/focal.

[4] N. Brümmer and J. A. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[5] D. Ramos Castro, *Forensic evaluation of the evidence using automatic speaker recognition systems*. PhD thesis, Autonomous University of Madrid, 2007.

[6] M. I. Mandasari, M. Gnther, R. Wallace, R. Saeidi, S. Marcel, and D. A. van Leeuwen, "Score calibration in face recognition," *IET Biometrics*, vol. 3, no. 4, pp. 246–256, 2014.

[7] D. van Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proceedings of Interspeech*, pp. 1619–1623, 2013.

[8] N. Brümmer, A. Swart, and D. van Leeuwen, "A comparison of linear and nonlinear calibrations for speaker recognition," in *Odyssey 2014: The Speaker and language Recognition Workshop*, pp. 14–18, 2014.

[9] N. Brümmer and D. Garcia-Romero, "Generative modelling for unsupervised score calibration," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1680–1684, 2014.

[10] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the Domain Adaptation Challenge using Deep Neural Networks," in *Proceedings of SLT 2014*, pp. 378–383, 2014.

[11] G. Sell and D. Garcia-Romero, "Speaker diarization with PLDA i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 413–417, 2014.

[12] D. Garcia-Romero, A. McCree, S. Shum, N. Brümmer, and C. Vaquero, "Unsupervised domain adaptation for i-vector speaker recognition," in *Proc. of Odyssey 2014, The Speaker and Language Recognition Workshop*, pp. 260–264, 2014.

[13] O. Plchot, P. Matějka, and al., "Analysis and description of ABC submission to NIST SRE 2016," in *Proc. Interspeech 2017*, pp. 1348–1352, 2017.

[14] O. E. Barndorff-Nielsen, "Normal inverse gaussian distributions and stochastic volatility modelling," *Scandinavian Journal of Statistics*, vol. 24, no. 1, pp. 1–13, 1997.

[15] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front–end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X–vectors: Robust DNN embeddings for speaker recognition," in *Proceedings of ICASSP 2018*, pp. 5329–5333, 2018.

[17] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the 9th European Conference on Computer Vision*, vol. Part IV of *ECCV'06*, pp. 531–542, 2006.

[18] P. Kenny, "Bayesian speaker verification with Heavy–Tailed Priors," in *Keynote presentation, Odyssey 2010, The Speaker and Language Recognition Workshop*, 2010.

[19] K. Slooten and R. Meester, "Forensic identification: Database likelihood ratios and familial DNA searching," in *arXiv:1201.4261 [stat.AP]*, 2012.

[20] O. Barndorff-Nielsen, J. Kent, and M. Sørensen, "Normal variance-mean mixtures and z distributions," *International Statistical Review / Revue Internationale de Statistique*, vol. 50, no. 2, pp. 145–159, 1982.

[21] O. Barndorff-Nielsen, "Exponentially decreasing distributions for the logarithm of particle size," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 353, no. 1674, pp. 401–419, 1977.

[22] K. Prause, "Modelling financial data using generalized hyperbolic distributions," tech. rep., University of Freiburg, 1997.

[23] K. Prause, *The Generalized Hyperbolic Model: Estimation, Financial Derivatives, and Risk Measures*. PhD thesis, University of Freiburg, 1999.

[24] B. Jørgensen, *Statistical properties of the Generalized Inverse Gaussian distribution*, vol. 9 of *Lecture notes in statistics*. Springer, 1982.

[25] E. Eberlein and E. A. von Hammerstein, "Generalized hyperbolic and inverse gaussian distributions: Limiting cases and approximation of processes," in *Seminar on Stochastic Analysis, Random Fields and Applications IV* (R. Dalang, M. Dozzi, and F. Russo, eds.), pp. 221–264, Birkhäuser Basel, 2004.

[26] P. Blæsild, "The two-dimensional hyperbolic distribution and related distributions, with an application to Johannsen's bean data," *Biometrika*, vol. 68, no. 1, pp. 251–263, 1981.

[27] R. P. Browne and P. D. McNicholas, "A mixture of generalized hyperbolic distributions," *Canadian Journal of Statistics*, vol. 43, no. 2, pp. 176–198, 2015.

[28] D. Karlis, "An EM type algorithm for maximum likelihood estimation of the normal–inverse Gaussian distribution," *Statistics & Probability Letters*, vol. 57, no. 1, pp. 43–52, 2002.

[29] S. Cumani, O. Plchot, and P. Laface, "On the use of i-vector posterior distributions in probabilistic linear discriminant analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 846–857, 2014.

[30] S. Cumani, "Fast scoring of full posterior plda models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2036–2045, 2015.

[31] N. Brümmer, *Measuring, refining and calibrating speaker and language information extracted from speech*. PhD thesis, Stellenbosch University, South Africa, 2010.

[32] S. Cumani and P. Laface, "Non–linear i–vector transformations for PLDA based speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 908–919, 2017.

[33] S. Cumani and P. Laface, "Joint estimation of PLDA and non–linear transformations of speaker vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1890–1900, 2017.