ESTIMATION OF MISSING LSF PARAMETERS USING GAUSSIAN MIXTURE MODELS

Rainer Martin, Carsten Hoelper, and Ingo Wittke

Institute of Communication Systems and Data Processing
Aachen University of Technology, D-52056 Aachen, Germany
Phone: +49 241 806984, Fax: +49 241 8888 186, E-mail: martin@ind.rwth-aachen.de

ABSTRACT

Speech transmission over packet networks has to cope with packet delays and packet losses. When a packet loss occurs the missing information must be estimated. In this contribution we focus on restoring the spectral parameters of a speech coder. A novel approach to estimating missing Line Spectral Frequency (LSF) parameters using Gaussian Mixture Models (GMM) is proposed. We present the estimation algorithm and study its performance when one or several LSF parameters are lost. We show that a GMM of a relatively low order is sufficient to achieve a substantial improvement in parameter SNR. Therefore, the new estimation procedure requires much less memory than histogram based estimation methods.

1. INTRODUCTION

The number of Internet users and the data traffic on the Internet have been rapidly increasing during the last years. The ubiquitous data networks and the emergence of new interactive applications make the integration of traditional speech services, such as telephony, into the packet networks of the Internet highly desirable. However, speech transmission imposes stringent real time demands on the network and network congestion is likely to lead to packet delays and packet loss. When a packet is delayed beyond acceptable limits or completely lost, the missing speech parameters must be extrapolated from the available information in order to satisfy real time constraints.

A versatile candidate for Voice over IP applications is the GSM Adaptive Multi-Rate speech coder [1]. It offers coding rates between 4.75 kbit/s and 12.2 kbit/s and allows switching of bitrates from one speech frame of 20 ms to the next. The AMR coder is based on the ACELP principle, i.e. it is a linear predictive (LP) coder with algebraic codebook excitation. The LP coefficients are converted to the Line Spectral Frequency (LSF) domain prior to transmission.

In this contribution we focus on the estimation of missing LSF parameters. The method presented here uses the intraframe correlation of the differentially encoded (residual) LSF parameters and *a priori* information to find a Minimum Mean Square Error (MMSE) estimate of missing LSF parameters given the present parameters. The *a priori* information which is required for MMSE estimation is the joint probability density function of the LSF parameters. In this work, a Gaussian Mixture Model (GMM) is used to model the joint density.

Optimal parameter estimation using *a priori* knowledge has been proposed before in the context of error concealment and *soft-bit* decoding, e.g. [2, 3]. In those studies the *a priori* knowledge has been stored in terms of histograms. The application of those concepts to intraframe estimation of LSF parameters is difficult as the dimension of the LSF vector leads to prohibitively large memory requirements. Suboptimal methods based on first order Markov modeling [4] require less memory but do not completely solve the memory problem. The GMM approach proposed in this paper, however, has very modest memory requirements.

The remainder of this paper is organized as follows: Section 2 summarizes the computation and the properties of differentially encoded (residual) LSF parameters in the AMR coder. Section 3 presents the MMSE estimator and Gaussian Mixture Models. In Section 4 we summarize the objective measurements for various GMM orders.

2. SPECTRAL PARAMETERS OF THE AMR SPEECH CODER

The GSM AMR speech coder implements eight different source coding modes at bitrates between 4.75 kbit/s and 12.2 kbit/s [1]. All modes use a filter of order 10 for linear prediction (LP) analysis. The 12.2 kbit/s mode (which is identical to the GSM enhanced fullrate coder) uses two different windows to calculate two sets of LSF parameters. All other modes use a single window to calculate one set of coefficients for each speech frame of 20 ms. The LP coefficients are converted to the LSF domain and differentially encoded. To remove correlation a mean vector is subtracted from each LSF vector and a first order linear prediction filter with fixed prediction coefficients is applied. For modes below 12.2 kbit/s the residual LSF vectors $\underline{L}_k$, where k denotes the frame index, are then Split Vector (SVQ) quantized. The 10-dimensional vectors are partioned into subsets of 3, 3, and 4 residual LSF coefficients. Each of these subsets is then vector quantized with 7 to 9 bits. E.g., for the 10.2 kbit/s mode the first, the second, and the third subset are quantized with 8, 9, and 9 bits, respectively. For the 10.2 kbit/s mode (which we used in this study) the joint histogram of the quantized LSF coefficients would require the storage of $2^8 \cdot 2^9 \cdot 2^9 \approx 67 \cdot 10^6$ histogram values. For a suboptimal Markov chain approach the transition matrices between subsets need to be stored. For the 10.2 kbit/s mode this still amounts to $2^8 \cdot 2^9 + 2^9 \cdot 2^9 = 393,216$ histogram values. These memory requirements exclude a histogram based estimation approach.

0-7803-7041-4/01/$10.00 ©2001 IEEE

Figure 1 shows an intensity plot of the correlation coefficient matrix of the residual LSF vectors, computed on a data set of 20,000 residual LSF vectors. Adjacent components of the LSF vector are strongly correlated. This intraframe correlation will be used in the estimation procedure outlined below.

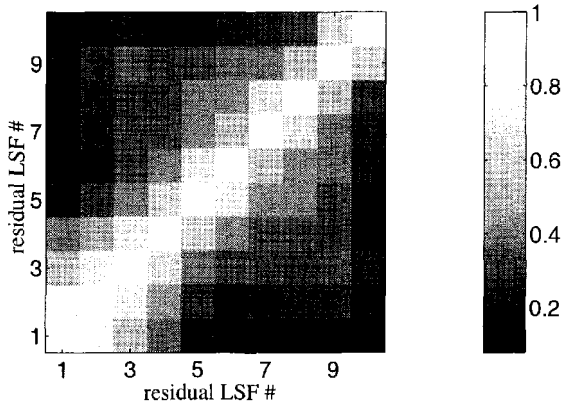

Fig. 1. Intensity plot of correlation coefficient matrix for residual LSF vectors.

3. ESTIMATION OF MISSING COMPONENTS

3.1. MMSE Estimators for Missing LSF Components

We assume that some components of the current residual LSF vector $\underline{L}_k$ are lost and that all other components and all other frames are received without error. The aim of the estimation procedure is to restore the missing components using MMSE estimation. The approach taken here is also known as "Missing Feature Theory" which has been employed in robust speech recognition [5].

We partition the current residual LSF vector $\underline{L}_k$ into a received (or present) part $\underline{L}_k^{(p)}$ and a lost (or missing) part $\underline{L}_k^{(m)}$,

$$\underline{L}_k = \begin{pmatrix} \underline{L}_k^{(m)} \\ \underline{L}_k^{(p)} \end{pmatrix} . \tag{1}$$

If quantization errors are negligible, the MMSE estimate of the missing components is given by the conditional expectation

$$\widehat{\underline{L}}_k^{(m)} = E\{\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}, \underline{Z}_{k-1}\} \tag{2}$$

as a function of the present components and the sequence of all previously received residual LSF vectors $\underline{Z}_{k-1}$. E.g., in case that $\underline{L}_k$ is governed by a first order Markov model and quantization errors are negligible, we obtain

$$\begin{aligned} \widehat{\underline{L}}_k^{(m)} &= E\{\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}, \underline{L}_{k-1}\} \\ &= \int_{\underline{L}_k^{(m)}} \underline{L}_k^{(m)} p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}, \underline{L}_{k-1}) d\underline{L}_k^{(m)} . \end{aligned} \tag{3}$$

If we neglect the correlation over time the expression can be further simplified

$$\widehat{\underline{L}}_k^{(m)} = \int_{\underline{L}_k^{(m)}} \underline{L}_k^{(m)} p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}) d\underline{L}_k^{(m)} . \tag{4}$$

To compute the optimal estimate the conditional probability density $p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)})$ must be known.

If all components of the current LSF vector are lost and correlation over time is neglected we have $p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}) = p(\underline{L}_k^{(m)})$. In this case the best estimate is the mean value. The substitution of the lost components by the (unconditional) mean will be termed *a priori* mean imputation.

3.2. Gaussian Mixture Models

Mixture models are frequently used in data classification and clustering problems with the aim of fitting the probability density function (pdf) of some given data. The mixture model can represent the statistics of the given data with a relatively small number of parameters. We approximate the joint probability density of the residual LSF vectors by a Gaussian Mixture Model [6], i.e., by a sum of M multivariate Gaussian densities

$$p(\underline{L}_k) = \sum_{i=1}^{M} \alpha_i \mathfrak{N}(\underline{L}_k, \underline{\mu}_i, \mathbf{C}_i) \tag{5}$$

where each N-dimensional mixture density is given by

$$\mathfrak{N}(\underline{L}_k, \underline{\mu}_i, \mathbf{C}_i) = \frac{1}{\sqrt{(2\pi)^N \mid \mathbf{C}_i \mid}} \tag{6}$$

$$\cdot \exp(-\frac{1}{2}(\underline{L}_k - \underline{\mu}_i)^T \mathbf{C}_i^{-1} (\underline{L}_k - \underline{\mu}_i)) \tag{7}$$

and α_i denotes the *a priori* probability of the mixture components $\mathfrak{N}_i = \mathfrak{N}(\underline{L}_k, \underline{\mu}_i, \mathbf{C}_i)$, i.e. $P(\mathfrak{N}_i) = \alpha_i$.

The mixture probabilities α_i, the mixture mean vectors $\underline{\mu}_i$, and the covariance matrices $\mathbf{C}_i$ are determined from training the model by means of the well known Estimate-Maximize (EM) algorithm [7]. To reduce the number of free parameters it is common to use covariance matrices with non-zero elements on the main diagonal only [8, 6].

3.3. MMSE Estimation Using GMM

In order to compute an approximate MMSE estimate using the GMM, we partition all parameters of the GMM with respect to present and missing components.

Analogous to the LSF parameter vector in (1) the mean vectors $\underline{\mu}_i$ and the covariances $\mathbf{C}_i$ of all mixture component can be then written as follows

$$\underline{\mu}_i = \begin{pmatrix} \underline{\mu}_i^{(m)} \\ \underline{\mu}_i^{(p)} \end{pmatrix} , \tag{8}$$

$$\mathbf{C}_i = \begin{pmatrix} \mathbf{C}_i^{(m,m)} & \mathbf{C}_i^{(m,p)} \\ \mathbf{C}_i^{(p,m)} & \mathbf{C}_i^{(p,p)} \end{pmatrix} . \tag{9}$$

The conditional pdf of present and missing components can be now expressed in terms of a GMM

$$\begin{aligned} p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}) &= \frac{p(\underline{L}_k^{(m)}, \underline{L}_k^{(p)})}{p(\underline{L}_k^{(p)})} \\ &= \sum_{i=1}^{M} \frac{\alpha_i}{p(\underline{L}_k^{(p)})} \mathfrak{N}(\underline{L}_k, \underline{\mu}_i, \mathbf{C}_i) . \end{aligned} \tag{10}$$

Since the conditional pdf and any marginal pdf of jointly Gaussian random variables are (multivariate) Gaussian densities, the joint probability $\mathfrak{N}(\underline{L}_k, \underline{\mu}_i, \mathbf{C}_i)$ of present and missing components can be factored into a conditional Gaussian pdf and a marginal Gaussian pdf (e.g. [9]). Therefore, $p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)})$ can be written as

$$p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}) = \frac{\sum_{i=1}^{M} \alpha_i \mathfrak{N}(\underline{L}_k^{(m)}, \underline{\mu}_i^{(m|p)}, \mathbf{C}_i^{(m|p)}) \mathfrak{N}(\underline{L}_k^{(p)}, \underline{\mu}_i^{(p)}, \mathbf{C}_i^{(p,p)})}{\sum_{i=1}^{M} \alpha_i \mathfrak{N}(\underline{L}_k^{(p)}, \underline{\mu}_i^{(p)}, \mathbf{C}_i^{(p,p)})} \tag{11}$$

with

$$\underline{\mu}_i^{(m|p)} = \underline{\mu}_i^{(m)} + \mathbf{C}_i^{(p,m)} (\mathbf{C}_i^{(p,p)})^{-1} (\underline{L}_k^{(p)} - \underline{\mu}_i^{(p)}) \tag{12}$$

and

$$\mathbf{C}_i^{(m|p)} = \mathbf{C}_i^{(m,m)} - \mathbf{C}_i^{(p,m)} (\mathbf{C}_i^{(p,p)})^{-1} \mathbf{C}_i^{(m,p)} . \tag{13}$$

We define the *a posteriori* probabilities

$$\alpha_i^{(m|p)} = \frac{\alpha_i \mathfrak{N}(\underline{L}_k^{(p)}, \underline{\mu}_i^{(p)}, \mathbf{C}_i^{(p,p)})}{\sum_{\ell=1}^{M} \alpha_\ell \mathfrak{N}(\underline{L}_k^{(p)}, \underline{\mu}_\ell^{(p)}, \mathbf{C}_\ell^{(p,p)})} \tag{14}$$

and obtain

$$p(\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}) = \sum_{i=1}^{M} \alpha_i^{(m|p)} \mathfrak{N}(\underline{L}_k^{(m)}, \underline{\mu}_i^{(m|p)}, \mathbf{C}_i^{(m|p)}) . \tag{15}$$

Using (4), (12), and (15) an approximate MMSE estimate of $\underline{L}_k^{(m)}$ is given by

$$\widehat{\underline{L}}_k^{(m)} = E\{\underline{L}_k^{(m)} \mid \underline{L}_k^{(p)}\} = \sum_{i=1}^{M} \alpha_i^{(m|p)} \underline{\mu}_i^{(m|p)} . \tag{16}$$

If the GMM models the pdf of the current residual LSF vector $\underline{L}_k$ by means of *diagonal* covariance matrices the off-diagonal matrices $\mathbf{C}_i^{(m,p)}$ and $\mathbf{C}_i^{(p,m)}$ are all zero and the MMSE estimate is given by

$$\widehat{\underline{L}}_k^{(m)} = \sum_{i=1}^{M} \alpha_i^{(m|p)} \underline{\mu}_i^{(m)} \tag{17}$$

where the *a posteriori* probabilities can be now easily computed using products of univariate normal densities

$$\alpha_i^{(m|p)} = \frac{\alpha_i \prod_{j=1}^{N_p} \mathfrak{N}(L_{k,j}^{(p)}, \mu_{i,j}^{(p)}, (\sigma_{i,j}^{(p,p)})^2)}{\sum_{l=1}^{M} \alpha_l \prod_{j=1}^{N_p} \mathfrak{N}(L_{k,j}^{(p)}, \mu_{l,j}^{(p)}, (\sigma_{l,j}^{(p,p)})^2)} = \tag{18}$$

$$\frac{\alpha_i \exp(-0.5 \sum_j (L_{k,j}^{(p)} - \mu_{i,j}^{(p)})^2 / (\sigma_{i,j}^{(p,p)})^2) \prod_j 1/\sigma_{i,j}^{(p,p)}}{\sum_{\ell=1}^{M} \alpha_i \exp(-0.5 \sum_j (L_{k,j}^{(p)} - \mu_{\ell,j}^{(p)})^2 / (\sigma_{\ell,j}^{(p,p)})^2) \prod_j 1/\sigma_{\ell,j}^{(p,p)}} . \tag{19}$$

$L_{k,j}^{(p)}$ denotes the j-th component of the k-th vector $\underline{L}_k^{(p)}$ of present components, and $\mu_{i,j}^{(p)}$ and $(\sigma_{i,j}^{(p,p)})^2$ are the mean and the variance of the j-th vector component and the i-th mixture component, respectively. $\sum_j = \sum_{j=1}^{N_p}$ and $\prod_j = \prod_{j=1}^{N_p}$ denote sums and products over the present components where N_p denotes the number of present components.

The memory requirements of the GMM approach are directly proportional to the dimension of the LSF vector and the GMM order, i.e for LSF vectors of size 10 and GMM's with diagonal covariance matrices $(10 + 10 + 1) \cdot M$ values must be stored. Since M is typically much smaller than 100 the GMM approach offers significant memory advantages with respect to the histogram approach.

4. EXPERIMENTAL RESULTS

The AMR coder groups the residual LSF parameters into three subsets. Each of these subsets is then vector quantized and transmitted. To reduce the probability that all subsets of a signal frame get lost we interleave the subsets of successive frames. When a single transmitted frame gets lost at least one out of three subsets is available for the reconstruction of the LSF vector.

In our experiments we therefore considered three different scenarios:

- Only one LSF coefficient is lost. The remaining 9 coefficients are present. This is of little practical importance for the above transmission scheme. It gives, however, an indication of how much can be achieved and might be useful in other speech enhancement applications.
- One of the three subsets is lost.
- Two of the three subsets are lost.

The experimental results were obtained using GMM's with 2, 4, 8, 16, 32, and 64 mixture components. The GMM's were trained by means of the EM algorithm and one million residual LSF vectors. A data base of (modified) IRS filtered male and female speech and the AMR coder (10.2 kbit/s mode) was used to generate the residual LSF vectors. The estimation algorithm was evaluated using 20,000 residual LSF vectors which were not part of the training data base.

4.1. Loss of a single LSF coefficient

The improvement in parameter SNR with respect to *a priori* mean imputation are shown in Table 1. The gain depends mainly on the position within the LSF vector and the correlation of adjacent LSF's. For $M = 64$ an average gain of about 4.1 dB is obtained. Further doubling the GMM order did result in small improvements but also in a significantly increased computational complexity.

4.2. Loss of LSF subsets

When a single subset of LSF coefficients is lost the results depend on the position within the subset. Table 2 summarizes the results. For $M = 64$ the average improvement is now 2.0 dB.

Table 3 presents the results for the case of two missing subsets. Again the improvement is best for those LSF coefficients which are adjacent to the present coefficients. The average improvement of a single LSF coefficient with respect to *a priori* mean imputation is now only 1.26 dB, however, six (subsets 1, 2) or seven (subsets 2, 3 or 1 ,3) coefficients are estimated. A comparison with Table 2 reveals that the subsets adjacent to a missing set contribute most to the improvement. A first order Markov approach is therefore close to optimal.

GMM order / lost LSF #	2	4	8	16	32	64
1	0.77	1.91	2.98	3.69	3.93	4.34
2	1.54	2.72	4.23	5.22	5.40	5.84
3	2.58	3.53	4.45	4.93	5.04	5.45
4	2.10	2.26	2.83	3.50	3.65	4.10
5	1.48	1.93	2.90	3.70	4.19	4.47
6	1.24	2.29	3.07	3.52	3.88	4.17
7	1.02	1.85	2.10	2.33	2.63	3.05
8	1.11	1.69	1.99	2.21	2.52	3.09
9	1.13	1.31	1.75	2.65	3.19	3.47
10	0.59	0.71	1.09	1.88	2.45	2.80
∅	1.35	2.02	2.74	3.36	3.69	4.08

Table 1. Improvement with respect to *a priori* mean imputation of parameter SNR in dB for estimating a single LSF coefficient using a GMM of order M. The average for order M is denoted by ∅.

lost set	M / LSF #	2	4	8	16	32	64
1	1	0.21	0.49	0.82	0.89	0.94	1.00
	2	0.67	1.01	1.28	1.42	1.44	1.50
	3	1.75	2.19	2.72	3.09	3.22	3.40
2	4	1.79	1.91	2.19	2.63	2.70	2.97
	5	0.95	1.24	1.44	1.69	1.80	1.90
	6	0.92	1.78	2.12	2.30	2.39	2.54
3	7	0.74	1.47	1.72	1.92	2.07	2.17
	8	0.75	1.14	1.32	1.40	1.47	1.53
	9	0.77	0.94	1.06	1.21	1.37	1.39
	10	0.33	0.41	0.49	0.79	0.99	1.05
	∅	0.91	1.29	1.56	1.78	1.88	2.0

Table 2. Improvement with respect to *a priori* mean imputation of parameter SNR in dB for the estimation of a single lost LSF subset using a GMM of order M.

lost sets	M / LSF #	2	4	8	16	32	64
1	1	-0.00	0.11	0.21	0.25	0.26	0.30
	2	0.31	0.44	0.49	0.55	0.58	0.62
	3	0.77	0.87	0.96	1.05	1.10	1.11
2	4	0.59	0.65	0.76	0.88	0.91	0.95
	5	0.91	1.05	1.17	1.27	1.28	1.34
	6	1.32	1.75	1.87	1.96	2.03	2.12
2	4	1.72	2.12	2.30	2.58	2.66	2.80
	5	0.50	0.96	0.94	1.08	1.13	1.16
	6	0.15	0.82	1.01	1.08	1.12	1.17
3	7	0.19	0.69	0.78	0.82	0.84	0.89
	8	0.25	0.56	0.69	0.75	0.76	0.79
	9	0.36	0.57	0.68	0.69	0.76	0.78
	10	0.14	0.22	0.29	0.39	0.50	0.53
1	1	0.32	0.67	0.83	0.84	0.91	0.92
	2	0.70	1.05	1.21	1.28	1.32	1.34
	3	1.90	2.37	2.77	3.00	3.06	3.18
3	7	1.10	1.71	1.79	1.92	2.02	2.10
	8	1.01	1.23	1.30	1.35	1.38	1.41
	9	0.88	0.93	0.93	1.04	1.10	1.12
	10	0.31	0.31	0.46	0.59	0.71	0.76
	∅	0.67	0.95	1.06	1.16	1.21	1.26

Table 3. Improvement with respect to *a priori* mean imputation of parameter SNR in dB for the estimation of two lost LSF subsets using a GMM of order M.

5. CONCLUSIONS

The proposed LSF reconstruction scheme was implemented in an Voice over IP transmission scheme. The transmission scheme uses frame interleaving for the coded LSF parameters and a multiple description scheme for the transmission of the LP residual. Informal listening tests confirmed that the estimation scheme as outlined above enhances the quality of the received speech signals when frame losses occur. The scheme is especially useful when no more than one LSF subset gets lost. Due to the correlation properties of residual LSF vectors, increasing the GMM order beyond 20–30 is only helpful when coefficients next to the lost coefficient are present. Compared to histogram based approaches the GMM based approach consumes significantly less memory.

The improvements for missing LSF subsets can be increased to the values given in Table 1 when the LSF coefficients are grouped differently into the subsets. If we, for instance, group LSF # 1, 3, 5 into subset one and LSF # 2, 4, 6 into subset two we obtain improvements which equal or exceed the ones given in Table 1.

6. ACKNOWLEDGMENT

We thank Peter Jax for providing a fast implementation of the EM algorithm.

7. REFERENCES

[1] ETSI, *GSM 06.71: Digital cellular telecomunications system (Phase 2+); Adaptive Multi-Rate (AMR); Speech Processing Functions; General Description.* European Telecommunications Standards Institute, 1998.

[2] T. Fingscheidt and P. Vary, "Robust Speech Decoding: A Universal Approach to Bit Error Concealment," in *ICASSP*, pp. 1667–1670, 1997.

[3] T. Fingscheidt and P. Vary, "Softbit Speech Decoding: A New Approach to Error Concealment," *IEEE Trans. Speech and Audio Processing*, (to appear).

[4] M. Adrat, J. Spittka, S. Heinen, and P. Vary, "Error Concealment by Near Optimum MMSE-Estimation of Source Codec Parameters," in *IEEE Workshop on Speech Coding*, pp. 84–86, 2000.

[5] A. Morris, M. Cooke, and P. Green, "Some Solutions to the Missing Feature Problem in Data Classification, with Application to Noise Robust ASR," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 737–740, 1998.

[6] P. Hedelin and J. Skoglund, "Vector Quantization Based on Gaussian Mixture Models," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 385–401, 2000.

[7] A. Dempster, N. Laird, and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Roy. Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[8] D. Reynolds and R. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models," *SAP*, vol. 3, no. 1, pp. 72–83, 1995.

[9] S. Kotz, N. Balakrishnan, and N. Johnson, *Continuous Multivariate Distributions*, vol. 1. Wiley, 2 ed., 2000.