A note on the calculation of a probabilistic distance between Hidden Markov Models *

Mehmet Karan† Brian D. O. Anderson†

Robert C. Williamson†

† Department of Systems Engineering, Research School of Physical Sciences and Engineering, The Australian National University, Canberra, ACT 0200, Australia.

Tel : +61 6 249 5127

Fax : +61 6 249 2698

E-mail : mxk101@syseng.anu.edu.au

Abstract

Streit in [1] conjectured that the expected value of the logarithms of the output sequence probabilities of Hidden Markov Models (HMMs) can be calculated from the integer moments of these output sequence probabilities. Assuming the conjecture to be valid, we can show how to calculate a probabilistic distance, the Kullback-Leibler (KL) number, between Hidden Markov Models in a simple way from their parameters. Since the correctness of the conjecture of Streit is difficult to check for general HMMs, it is specialized to the case of Bernoulli processes, and the conjecture is shown to be incorrect. Using the Central Limit Theorem, the conjecture can be modified and verified to be correct for Bernoulli processes such that the KL number for Bernoulli processes can be easily calculated indirectly. Unfortunately the extension of the calculation of the KL number between two general HMMs using the modified conjecture is still computationally very difficult, nevertheless the results indicate that there may be an indirect way of calculating the KL number between HMMs.

1 *Introduction*

The Kullback-Leibler (KL) number known in the literature also as the discrimination information, the directed divergence or the I-divergence, is a useful tool in statistics [2] and signal processing (e.g. [3]) as a probabilistic distance between stochastic processes. Hidden Markov Models (HMMs) have wide applications [4] and it is desirable to find a simple formula for the KL number between HMMs. It can be employed in several areas, for example the approximation of HMMs by HMMs which have a lesser number of states. However, there is no simple way of calculating the KL number between discrete-time HMMs which have a finite number of states and outputs, except by simulations, as is discussed in [5].

In [1], Streit made the observation that the logarithms of the output sequence probabilities of HMMs are approximately Gaussian. Based on this observation, he conjectured that the expected value of the logarithms of the output sequence probabilities of HMMs can be calculated from the integer moments of these output sequence probabilities, in a manner described below. If his conjecture is true, then, as we show in the next section, it can be used to calculate the KL number between HMMs using a simple formula. Thus it is important to check the correctness of this conjecture, which in turn depends on a second conjecture with a Central Limit Theorem flavour. This second conjecture has also not been proven in the literature. In the third section, in order to check the correctness of the main conjecture it is specialized to Bernoulli processes where Central Limit Theorem ideas are applicable and it is shown that Streit's conjecture needs modification. The fourth section is devoted to the modification of Streit's conjecture so that the KL number between Bernoulli processes can be calculated indirectly. Although the extension of this modification to general HMMs is computationally difficult, it is still important to show the possible existence of an indirect way of calculating the KL number between HMMs.

2 *The calculation of the KL number between HMMs using Streit's conjecture*

In this section, we will show that it is possible to calculate the KL number between HMMs if the

*The authors wish to acknowledge the funding of the activities of the Cooperative Research Centre for Robust and Adaptive Systems by the Australian Commonwealth Government under the Cooperative Research Centres Program

conjecture made in [1] is true. For this purpose we will restate the results of the asymptotic analysis of Streit's algorithm to calculate the integer moments of output sequence probabilities, which is given in [6].

A discrete-time Hidden Markov Model which has a finite state set $S = \{1, \ldots, N\}$ and a finite output set $\mathcal{O} = \{1, \ldots, M\}$ can be parametrized by $\lambda = (A, B, \Pi)$ where the entries of the state transition probability matrix A are given by

$$a_{rs} = \mathrm{P}\{X(t+1) = s \mid X(t) = r\} \quad r, s = 1, \ldots, N. \tag{2.1}$$

Here $X(t)$ for $t = 1, 2, \ldots$ denotes the state of the HMM at time t. The output probability matrix $B = [b_{kr}]_{M \times N}$ is defined by

$$b_{kr} = \mathrm{P}\{O(t) = k \mid X(t) = r\} \quad r = 1, \ldots, N, \quad k = 1, \ldots, M \tag{2.2}$$

where $O(t)$ for $t = 1, 2, \ldots$ is the output of the HMM at time t. The initial state probability vector $\Pi = [\pi_r]_{1 \times N}$ is given by

$$\pi_r = \mathrm{P}\{X(1) = r\} \quad r = 1, \ldots, N. \tag{2.3}$$

If the state transition probability matrix A is irreducible, then the asymptotic behaviour of the HMM can be analysed by selecting the initial state probability vector Π as the left eigenvector of A corresponding to the eigenvalue 1 such that the entries of Π add to 1.

Given an output sequence $O_T = (O(1), \ldots, O(T))$ such that $O(t) \in 1, \ldots, M$ for $t = 1, \ldots, T$, the probability of this output sequence using the probability measure $P^{(T)}(\cdot)$ defined by the parameters of the HMM $\lambda = (A, B, \Pi)$ which has N states and M outputs on the set of all output sequences of length T, can be calculated as

$$P^{(T)}(O_T) = \Pi\, B(O(1))\, A \ldots A\, B(O(T-1)) \times A\, B(O(T))\, \underline{1}_N' \tag{2.4}$$

where $\underline{1}_N$ is a N dimensional row vector of ones, $(\cdot)'$ denotes the transpose of $(\cdot)$, and $B(O(t))$ is a diagonal matrix defined by

$$B(O(t)) = \mathrm{diag}\,\{\, b_{O(t),1}, \ldots, b_{O(t),N} \,\}, \tag{2.5}$$

for $O(t) = 1, \ldots, M$ and $t = 1, \ldots, T$.

For the sake of simplicity, the superscript (T) will be dropped in $P^{(T)}(\cdot)$ in the following sections.

Given two HMMs $\lambda_i = (A^{(i)}, B^{(i)}, \Pi^{(i)})$ and $\lambda_j = (A^{(j)}, B^{(j)}, \Pi^{(j)})$ which have the same output set $\mathcal{O}$, and which have N_i and N_j states, respectively, the KL number $I_T(\lambda_i \| \lambda_j)$ between the probability measures of these HMMs on the set of all output sequences of length T, is defined by

$$I_T(\lambda_i \| \lambda_j) = \begin{cases} E_i \left\{ \log \left(\frac{P_i(O_T)}{P_j(O_T)} \right) \right\} & P_j << P_i \\ +\infty & \text{Else} \end{cases} \tag{2.6}$$

where P_i and P_j are the probability measures defined by the parameters of the HMM λ_i and λ_j, respectively. Here E_i is the expectation with respect to the probability measure P_i and $P_j << P_i$ means the probability measure P_j is absolutely continuous with respect to P_i. A sufficient condition for absolute continuity is that $A^{(i)}, A^{(j)}, B^{(i)}$ and $B^{(j)}$ are all positive matrices. In fact, the KL number $I_T(\lambda_i \| \lambda_j)$ can be expressed as the difference between the entropy, $H_T(\lambda_i)$, of the HMM λ_i and the relative entropy, $H_T(\lambda_j \mid \lambda_i)$, of the HMM λ_j with respect to the HMM λ_i as

$$I_T(\lambda_i \| \lambda_j) = -H_T(\lambda_i) + H_T(\lambda_j \mid \lambda_i) \tag{2.7}$$

where the entropy, $H_T(\lambda_i)$, and the relative entropy, $H_T(\lambda_j \mid \lambda_i)$, are given by

$$H_T(\lambda_i) = -\sum_{O_T} P_i(O_T) \log(P_i(O_T)) \tag{2.8}$$

$$H_T(\lambda_j \mid \lambda_i) = -\sum_{O_T} P_i(O_T) \log(P_j(O_T)). \tag{2.9}$$

The asymptotic KL number between HMMs λ_i and λ_j per sample, which we will simply call the asymptotic KL number, $I(\lambda_i \| \lambda_j)$ is defined by

$$I(\lambda_i \| \lambda_j) = \lim_{T \to \infty} \frac{I_T(\lambda_i \| \lambda_j)}{T} \tag{2.10}$$

$$= -H(\lambda_i) + H(\lambda_j \mid \lambda_i) \tag{2.11}$$

where the entropy rate of the HMM λ_i , $H(\lambda_i)$ is given by

$$H(\lambda_i) = \lim_{T \to \infty} \frac{H_T(\lambda_i)}{T} \tag{2.12}$$

$$= \lim_{T \to \infty} \frac{1}{T} P_i(O_T) \quad \text{a.s. } P_i \tag{2.13}$$

where a.s. P_i means almost sure convergence with respect to the probability measure P_i. Here (2.13) is known as the Shannon-McMillan-Breiman Theorem and also as the Asymptotic Equipartition Property (AEP) [7]. The relative entropy rate of the HMM λ_j with respect to the HMM λ_i is given by

$$H(\lambda_j \mid \lambda_i) = \lim_{T \to \infty} \frac{H_T(\lambda_j \mid \lambda_i)}{T} \tag{2.14}$$

$$= \lim_{T \to \infty} \frac{1}{T} P_j(O_T) \quad \text{a.s. } P_i \tag{2.15}$$

where a sufficient condition for the limit in (2.15) to exist is that the matrices $A^{(i)}, A^{(j)}$, $B^{(i)}$ and $B^{(j)}$ are positive matrices, as shown in [8].

In [1], Streit observed that the random variables $\log(P_i(O_T))$ and $\log(P_j(O_T))$ are asymptotically Gaussian, and using this observation he conjectured that the expected values of these random variables can be calculated approximately from any two of the integer moments of the random variables $P_i(O_T)$ and $P_j(O_T)$, for finite T. In [1], an algorithm to calculate these integer moments and some examples which support the conjecture are given. Note that if his conjecture is true, then the entropy, $H_T(\lambda_i)$, and the relative entropy, $H_T(\lambda_j \mid \lambda_i)$, for finite time can be calculated approximately from (2.8) and (2.9). Then the KL number between HMMs can be calculated using the asymptotic analysis of Streit's algorithm to calculate the integer moments of the random variables $P_i(O_T)$ and $P_j(O_T)$ as given in [6].

In [6], the algorithm to calculate the integer moments $M_{ji}(k,T)$ of the random variable $P_j(O_T)$ given in [1], is reformulated in a matrix algebra framework as

$$\begin{aligned} M_{ji}(k,T) &= E_i\left\{P_j(O_T)^k\right\} && (2.16)\\ &= \underline{1}_{\bar{N}}\, z_T. && (2.17)\end{aligned}$$

where $\bar{N}$ is equal to $N_i N_j^k$ and z_T is computed from the recursive algorithm

$$z_{t+1} = F_{ji}(k)\ z_t, \quad t = 2, \ldots, T \qquad (2.18)$$

where $F_{ji}(k)$ is a $\bar{N}\times\bar{N}$ dimensional square matrix given by

$$F_{ji}(k) = C_{ji}(k)\, \bar{A}_{ji}(k)' \qquad (2.19)$$

The matrix $C_{ji}(k)$ is a diagonal matrix of the same dimension as $F_{ji}(k)$ and its diagonal entries are given by $\Gamma(r_i, r_{j_1}, \ldots, r_{j_k})$, where $r_i = 1, \ldots, N_i$ and $r_{j_\nu} = 1, \ldots, N_j$ for $\nu = 1, \ldots, k$ in lexicographic order. $\Gamma(r_i, r_{j_1}, \ldots, r_{j_k})$ is defined as

$$\Gamma(r_i, r_{j_1}, \ldots, r_{j_k}) = \sum_{l=1}^{M} b_{l,r_i}^{(i)} \left\{ \prod_{\nu=1}^{k} b_{l,r_{j_\nu}}^{(j)} \right\}. \qquad (2.20)$$

The matrix $\bar{A}_{ji}(k)$ is defined by

$$\bar{A}_{ji}(k) = A^{(i)} \otimes (A^{(j)})^{[k]} \qquad (2.21)$$

where $\otimes$ denotes the Kronecker product between matrices and $A^{[k]}$ denotes the k-fold Kronecker product of the matrix A by itself. The algorithm in (2.18) is initialized by

$$z_1 = C_{ji}(k)\ \left(\Pi^{(i)} \otimes (\Pi^{(j)})^{[k]}\right). \qquad (2.22)$$

If the matrices $A^{(i)}, A^{(j)}$, $B^{(i)}$ and $B^{(j)}$ are positive matrices, then, as shown in [6], the moments $M_{ji}(k,T)$ can be approximated as

$$M_{ji}(k,T) \approx \rho^T(F_{ji}(k))\ \underline{1}_{\bar{N}}\, x\, y'\, z_1 \qquad (2.23)$$

in the sense that $\frac{1}{T}\log(M_{ji}(k,T))$ goes to $\rho(F_{ji}(k))$ as T goes to infinity. In (2.23), $\rho(F_{ji}(k))$ denotes the eigenvalue of $F_{ji}(k)$ which has the maximum modulus and x and y are the corresponding right and left eigenvectors of $F_{ji}(k)$ such that $x'y = 1$. Under the assumption that the parameter matrices of the HMMs λ_i and λ_j are positive the matrices $C_{ji}(k)$ and $\bar{A}_{ji}(k)$ are positive matrices, hence the matrix $F_{ji}(k)$ is also positive. Thus the eigenvalue $\rho(F_{ji}(k))$ and the entries of the eigenvectors x and y are positive and real by the Perron-Frobenius Theorem [9].

Streit conjectured in [1] that the random variable $\log(P_j(O_T))$ is approximately Gaussian for finite time and hence the integer moments of $P_j(O_T)$ can be used to calculate the expected value of the random variable $\log(P_j(O_T))$ as

$$\begin{aligned} H_T(\lambda_j \mid \lambda_i) &= E_i\{\log(P_j(O_T))\}\\ &\approx 2\log(M_{ji}(1,T)) - \frac{1}{2}\log(M_{ji}(2,T)).\end{aligned} \qquad (2.24)$$

Then from (2.23) and (2.24), it follows that

$$\begin{aligned} H(\lambda_j \mid \lambda_i) &= \lim_{T\to\infty} \frac{1}{T} H_T(\lambda_j \mid \lambda_i)\\ &= \lim_{T\to\infty} \frac{1}{T} \left\{ T\log\left(\frac{\rho(F_{ji}(1))^2}{\sqrt{\rho(F_{ji}(2))}}\right) \right.\\ &\quad \left. + \text{ constant}\right\}\\ &= \log\left(\frac{\rho(F_{ji}(1))^2}{\sqrt{\rho(F_{ji}(2))}}\right). \end{aligned} \qquad (2.25)$$

The entropy rate of the HMM λ_i can be calculated similarly and the asymptotic KL number between the HMMs λ_i and λ_j can be found as

$$I(\lambda_i \| \lambda_j) = \log\left(\left[\frac{\rho(F_{ii}(1))}{\rho(F_{ji}(1))}\right]^2 \left[\frac{\rho(F_{ji}(2))}{\rho(F_{ii}(2))}\right]^{1/2}\right) \qquad (2.26)$$

using Streit's conjecture.

The validity of this derivation of the asymptotic KL number between HMMs depends on the correctness of Streit's conjecture. However this conjecture is a special case of a conjecture which has a Central Limit Theorem flavour. However the authors are aware of no Central Limit Theorem style of results for discrete time HMMs which have a finite number of states and outputs. Hence the validity of this conjecture can be understood better by specializing it to Bernoulli processes where there are sufficient tools to analyse this conjecture.

3 *The analysis of Streit's conjecture using Bernoulli processes*

Bernoulli processes are a special class of HMMs where the state set and the output set are the same and the state transitions do not depend on the previous state. This means that all columns of the matrix A are identical, and the matrix B is the identity matrix.

Given two Bernoulli processes $\lambda_i = (\Pi^{(i)})$ and $\lambda_j = (\Pi^{(j)})$ where

$$\begin{aligned} \Pi_r^{(i)} &= \mathrm{P}_i\{O(t) = r\} \qquad r = 1, \ldots, N, \quad \forall t, \\ \Pi_r^{(j)} &= \mathrm{P}_j\{O(t) = r\} \qquad r = 1, \ldots, N, \quad \forall t, \end{aligned} \tag{3.1}$$

the integer moments of $P_j(O_T)$ with respect to the probability measure P_i, are given by

$$M_{ji}(k,T) = E_i\{P_j(O_T)^k\} \tag{3.2}$$

$$= \left(\sum_{r=1}^{N} \Pi_r^{(i)} (\Pi_r^{(j)})^k \right)^T . \tag{3.3}$$

Then using (2.14) and (2.24), the relative entropy of the Bernoulli process λ_j derived using Streit's conjecture $\tilde{H}(\lambda_j \mid \lambda_i)$ can be found as

$$\tilde{H}(\lambda_j \mid \lambda_i) = \log \left(\frac{[\ \sum_{r=1}^{N} \Pi_r^{(i)}\ \ \Pi_r^{(j)}\ ]^2}{\sqrt{\sum_{r=1} \Pi_r^{(i)}\ \ (\Pi_r^{(j)})^2}} \right). \tag{3.4}$$

which is equivalent to (2.26) for Bernoulli processes. If Streit's conjecture were correct, then the direct calculation of the KL number for Bernoulli processes (which is possible),would have to yield a result equal to that obtained via the indirect calculation of the KL number using Streit's conjecture given in (3.4). However if the relative entropy is calculated directly from the definition for Bernoulli processes as

$$\begin{aligned} H(\lambda_j \mid \lambda_i) &= \frac{1}{T} \sum_{t=1}^{T} E_i \{\log (P_j(O(t)))\} \\ &= \sum_{r=1}^{N} \Pi_r^{(i)} \log \left(\Pi_r^{(j)}\right), \end{aligned} \tag{3.5}$$

then a different expression for the relative entropy is obtained. Thus this shows that Streit's conjecture is not correct. Hence the formula to calculate the KL number between HMMs in (2.26) is wrong.

To understand why the conjecture is wrong, we should look more closely at the Central Limit Theorem (CLT) for Bernoulli processes. This is done in the next section.

4 *Modification of Streit's conjecture*

Although Streit's conjecture seems reasonable, it yields the wrong answer for the calculation of KL number for Bernoulli processes. The CLT for Bernoulli processes [10] says that

$$\frac{1}{T} \log (P_j(O_T)) = \frac{1}{T} \sum_{t=1}^{T} \log (P_j(O(t))) \xrightarrow{\mathcal{L}} z \tag{4.1}$$

where z is a normal random variable with mean μ and variance σ^2/T such that μ and σ^2 are independent of T. Here $\xrightarrow{\mathcal{L}}$ means convergence in law, i.e. if $z_T \xrightarrow{\mathcal{L}} z$, then the distribution of z_T converges to the distribution of z asymptotically. In other words, (4.1) can be interpreted as saying $\log \left(P_j(O_T)^{1/T}\right)$ is approximately Gaussian. On the other hand, Streit's conjecture was

$$\log (P_j(O_T)) \xrightarrow{\mathcal{L}} y \tag{4.2}$$

where y is a Gaussian random variable whose mean is $\bar{\mu}\, T$ and variance $\bar{\sigma}^2\, T$ such that $\bar{\mu}$ and $\bar{\sigma}$ are independent of T. The problem with this interpretation of the CLT is that since the mean of y diverges asymptotically, its mean cannot be calculated using the integer moments of $P_j(O_T)$.

A natural question which arises in the light of these facts, is whether it is possible to calculate the KL number using the moments of $P_i(O_T)^{1/T}$ and $P_j(O_T)^{1/T}$. The answer is "Yes". The integer moments of the random variable $[P_j(O_T)]^{1/T}$ with respect to P_i can be found as

$$\begin{aligned} \tilde{M}_{ji}(k,T) &= E_i \left\{P_j(O_T)^{k/T}\right\} \\ &= \left(\sum_{r=1}^{N} \Pi_r^{(i)} (\ \Pi_r^{(j)}\ )^{k/T}\right)^T . \end{aligned} \tag{4.3}$$

By virtue of the near log-normality of $[P_j(O_T)]^{1/T}$ for finite but large T, it follows that the mean of $\log(P_j(O_T)^{1/T})$ can be calculated approximately as;

$$E_i\left\{\log\left(P_j(O_T)^{1/T}\right)\right\}$$
$$\approx 2\log(\bar{M}_{ji}(1,T)) - \frac{1}{2}\log(\bar{M}_{ji}(2,T))$$
$$\approx T\log\left(\frac{[\sum_{r=1}^{N}\Pi^{(i)}(\Pi_r^{(j)})^{1/T}]^2}{\sqrt{\sum_{r=1}^{N}\Pi^{(i)}(\Pi_r^{(j)})^{2/T}}}\right) \quad (4.4)$$

where the error goes to zero as T goes to infinity. It can be shown using L'Hopital's rule that the right hand side of (4.4) converges to the relative entropy rate of the Bernoulli process λ_j as

$$(\text{ RHS of } (4.4)\ ) \stackrel{T\to\infty}{\longrightarrow} H(\lambda_j \mid \lambda_i) = \sum_{r=1}^{N}\Pi_r^{(i)}\log\left(\Pi_r^{(j)}\right). \quad (4.5)$$

Hence the KL number between Bernoulli processes can be calculated from the integer moments of $P_i(O_T)^{1/T}$ and $P_j(O_T)^{1/T}$, but not from the integer moments of $P_i(O_T)$ and $P_j(O_T)$. Unfortunately, to extend this result for HMMs is computationally very difficult since the determination of a recursive formula to calculate the integer moments of $P_i(O_T)^{1/T}$ and $P_j(O_T)^{1/T}$ does not seem possible.

5 Examples

In this section, we will consider some Bernoulli processes to show the difference the between the relative entropy rate calculations using Streit's conjecture, the modified conjecture and the exact calculation. For Bernoulli processes, the Streit conjecture gives the approximate expression in (3.4) whereas the modified conjecture gives the approximate expression in (4.4) and the exact calculation of the relative entropy is (3.5). For the Bernoulli processes $\Pi^{(i)} = [0.3, 0.4, 0.3]$ and $\Pi^{(j)} = [0.7, 0.1, 0.2]$, these expressions become

$$(3.4) = -1.43536, \quad (4.4) = \begin{cases} -1.51299, & T=5 \\ -1.51088, & T=100 \end{cases}$$
$$(3.5) = -1.51087 \quad (5.1)$$

For the Bernoulli processes $\Pi^{(i)} = [0.3, 0.6, 0.1]$ and $\Pi^{(j)} = [0.7, 0.1, 0.2]$, they become

$$(3.4) = -1.54999, \quad (4.4) = \begin{cases} -1.65391, & T=5 \\ -1.64951, & T=100 \end{cases}$$
$$(3.5) = -1.6495. \quad (5.2)$$

As already remarked, the right hand side of (4.4) is asymptotically equal to (3.5), so that for a large T, (4.4) is very close to (3.5) On the other hand, the difference between (3.4) and the exact relative entropy in (3.5) depends on the parameters of the Bernoulli processes.

6 Conclusion

A simple way of calculating the Kullback-Leibler number between Hidden Markov Models using Central Limit Theorem type of results has been investigated. The motivation was a conjecture made in [1] which allows the calculation of the KL number using the integer moments of the output sequence probabilities of HMMs. However by specializing the conjecture it has been shown that the conjecture must be modified in order to calculate the KL number between Bernoulli processes. Unfortunately, using the new conjecture the calculation of the KL number between HMMs appears to be computationally very difficult.

References

[1] R. L. Streit, "The moments of matched and mismatched Hidden Markov Models," *IEEE Transactions on ASSP*, vol. 38, no. 4, pp. 610–622, 1990.

[2] S. Kullback, *Information Theory and Statistics*, vol. 10. John Wiley and Sons, Inc., 1959.

[3] R. M. Gray, A. H. Gray, G. R. Jr., and J. E. Shore, "Rate-distortion speech coding with a minimum discrimination information distortion measure," *IEEE Transactions on Information Theory*, vol. IT-27, no. 6, pp. 708–721, 1981.

[4] L. R. Rabiner and B. H. Juang, "An introduction to Hidden Markov Models," *IEEE ASSP Magazine*, pp. 4–16, 1986.

[5] H. H. Juang and L. R. Rabiner, "A probabilistic distance measure for Hidden Markov Models," *AT&T Technical Journal*, vol. 64, no. 2, pp. 391–408, 1985.

[6] M. Karan, B. D. O. Anderson, and R. C. Williamson, "A simple calculation of the joint moments of Hidden Markov Models," *Submitted to 1994 International Conference on Acoustics, Speech & Signal Processing, ICASSP-1994.*

[7] T. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991.

[8] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite Markov chains," *Ann. Math. Stat.*, vol. 37, pp. 1554–1563, 1966.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge University Press, 1985.

[10] C. R. Rao, *Linear statistical inference and its applications.* John Wiley & Sons, 1973.