EFFICIENT COMPUTATION OF STATISTICS FOR BANDED MULTIVARIATE NORMAL DISTRIBUTIONS *

KEVIN S. VAN HORN†

Abstract. We look at the problem of computing statistics for a large multivariate normal distribution whose *precision* matrix has limited bandwidth. In particular, we wish to randomly sample from the distribution, compute its mean vector, and compute a central band of its covariance matrix. We show how to accomplish these tasks in time linear in the dimensionality of the distribution, for fixed bandwidths. Our algorithms have applications in parameter estimation for certain maximum-entropy models.

Key words. maximum entropy, Monte Carlo methods, band matrix, multivariate normal, mean, covariance

AMS subject classifications. 65C10, 65C50, 65C60

1. Problem Statement.

We are given the following as input:

1. integers n, k, and κ with $0 \le k \le \kappa < n$;
2. an n-variate, bandwidth-k normal form $F = (A, b, c)$.

DEFINITION 1.1. *$F = (A, b, c)$ is an n-variate normal form if A is an $n \times n$, symmetric, negative-definite, real matrix; b is a real n-vector; and c is a real scalar. We say that F* has bandwidth *k if A has bandwidth k, i.e., $A[i,j] = 0$ whenever $|i - j| > k$.*

F defines the function

$$f(x; A, b, c) = \exp\left(x^{\mathrm{t}} A x + b^{\mathrm{t}} x + c\right), \tag{1.1}$$

which is proportional to a multivariate normal pdf over R^n whose precision (inverse covariance) matrix is $-2A$. Let x be a vector random variable whose pdf is proportional to $f(x; A, b, c)$. We wish to carry out the following tasks each in $O(n\, p(k, \kappa))$ time, where $p(k, \kappa)$ is a low-order polynomial in k and κ:

1. Compute the mean for x.
2. Compute a bandwidth-κ central band of the covariance matrix for x.
3. Randomly sample x.
4. Compute the integral $\int_{\mathsf{R}^n} f(x; A, b, c)dx$.

Tasks 1 and 4 are straightforward to accomplish within our time-complexity bounds. First, complete the square:

$$x^{\mathrm{t}} A x + b^{\mathrm{t}} x + c = -\frac{1}{2}(x - \mu)^{\mathrm{t}} \Pi (x - \mu) + c',$$

where $\Pi = -2A$, $\Pi\mu = b$, and $c' = c + \frac{1}{2}\mu^{\mathrm{t}} b$. This takes $O(nk^2)$ time, and produces the mean vector μ as the solution for Task 1. Do a Cholesky decomposition of Π ($O(nk^2)$ time) then take the product of the diagonal elements to obtain $\det \Pi$ ($O(n)$ time). The integral of Task 4 is then

$$\exp(c') \left(\frac{(2\pi)^n}{\det \Pi}\right)^{1/2}.$$

*This work supported in part by NSF EPSCoR. This work has been submitted to the SIAM for possible publication. Copyright may be transferred without notice, after which this version will be superseded.

†Dept. of Computer Science, North Dakota State University, Fargo, ND 58105 (Kevin.VanHorn@ndsu.nodak.edu)

Altogether, this takes $O(nk^2)$ time.

Unfortunately, the obvious solutions to tasks 2 and 3 do not meet our time complexity requirement. Task 2 would seem to require a matrix inversion, and the usual methods for matrix inversion require $\Theta(n^3)$ time. For a bandwidth-k matrix M one can reduce this to $\Theta(nk^2)$ (for an LU decomposition) plus $\Theta(n^2k)$ (use backsubstitution to solve $Mu_i = e_i$ for u_i, $0 \leq i < n$, where e_i is column i of the identity matrix); however, this is still too costly.

The obvious algorithm for Task 3 is

1. Compute for Π an orthonormal basis of eigenvectors u_i and their corresponding eigenvalues λ_i.
2. Draw n independent samples α_i from the 0-centered, unit-variance normal distribution.
3. Return $\mu + \sum_{i=1}^{n} \alpha_i \lambda_i^{-1/2} u_i$.

Step 3 alone takes $\Theta(n^2)$ time. Thus we must look further for efficient solutions to tasks 2 and 3.

2. Motivation. The tasks of §1 arise during maximum-likelihood parameter estimation for a class of maximum-entropy statistical models that have applications in speech recognition, handwriting recognition, and other sequential phenomena traditionally modeled by hidden Markov models, but for which the HMM assumption of conditional independence of observations is inappropriate [5]. We describe a simplified example from this class of models.

Suppose that we have a sequence s of n discrete *states*, and a parallel sequence x of continuous values. In speech recognition the states correspond to segments of phonemes and x to the temporal evolution of some acoustic feature produced from the raw acoustic signal by a signal-processing front end. Suppose also that for each state we have gathered statistics giving the mean and variance of the feature in that state, along with the mean and variance of various estimated time derivatives of the feature (to capture some of the dynamics of the signal). In particular, for each state q and $0 \leq r \leq R$, we have measured average values for

$$\sum_{t:s_t=q} (D^r x)_t \tag{2.1}$$

$$\sum_{t:s_t=q} ((D^r x)_t)^2 \tag{2.2}$$

over some training set of vectors x. D^r is a band matrix used to compute order-r estimated time derivatives, and R is the maximum derivative order. For example, one might define row t of D^1 to be the coefficients one would use to do a least-squares linear fit of the values x_u, $t - w \leq u \leq t + w$, for some window width w. D^0 is the identity matrix. $(D^r x)_t$ is the value of the estimated r-th derivative of the acoustic feature at time t.

We wish to construct a distribution over x whose statistics (expected values for (2.1) and (2.2)) match the empirical averages we have measured; these averages should, in some sense, be the entirety of the information expressed by the distribution. The principle of maximum entropy [2] prescribes that we should then choose the unique distribution that maximizes the entropy (a measure of the spread and uncertainty of a distribution) subject to the given statistical constraints. This maximum-entropy distribution has a pdf that is proportional to $f(x; A, b, c)$ if we define

1. $A = \sum_r (D^r)^{\mathrm{t}} \Lambda^r D^r$,

2. Λ^r is a diagonal matrix whose diagonal element t is $\lambda[s_t, r]$,
3. $b = \sum_k (D^r)^{\mathrm{t}} \tilde{\nu}^r$,
4. $\tilde{\nu}^r_t = \nu[s_t, r]$,
5. $\nu[q, r]$ and $\lambda[q, r]$, for arbitrary states q and $0 \le r \le R$, are parameters chosen to make the expected values of (2.1) and (2.2) equal to their empirical averages.

Note that the bandwidth of A is twice the maximum of the bandwidths of the matrices D^r.

In [5] we use a variant of the above model in which there are several hundred possible state sequences for each utterance, a dozen or so acoustic features, and the temporal sequences of values for distinct features are independent given the state sequence. We obtain maximum-likelihood parameter estimates by applying a standard gradient-based optimization procedure; thus, efficient computation of the log likelihood and its gradient are essential. This leads to the tasks of §1:

Task 4. The log of this integral is required in computing the log likelihood.

Tasks 1 and 2. The gradient of the log likelihood is proportional to the vector of differences between expected values of (2.1) or (2.2) and the corresponding empirical averages, for each feature, state s, and order r. We compute the needed expected values from the mean μ (Task 1) and a central band of the covariance matrix C (Task 2) for each feature, given a particular state sequence, as follows. Define $\mu' = D^r \mu$ and $C' = D^r C (D^r)^{\mathrm{t}}$. Then the expected value of (2.1) is the sum of μ'_t over all t for which $s_t = q$, and the expected value of (2.2) is the sum of $C'[t, t]$ over all t for which $s_t = q$. It is easily verified that computation of the diagonal of C' requires only a bandwidth-κ central band of C, where κ is twice the bandwidth of D^r.

Task 3. This random sampling is useful for Monte Carlo computation of expectations of arbitrary functions of the temporal sequence of feature values. This is needed if one applies the improved iterative scaling algorithm [1] in parameter estimation, or if one wishes to compute the observed information matrix as a measure of parameter uncertainty [4].

In this application we have $\kappa = k$, with typical values for k running from 1 to 8. Typical values for n run from 60 to 1000.

3. Marginals. Our solutions are based on analytically integrating $f(x; A, b, c)$ one variable at a time and storing information about the resulting log-quadratic functions. This approach was inspired by elimination algorithms for doing inference with Bayesian networks [3]. A forward pass to compute each of these n marginals produces the log integral of Task 4 as a side effect. The information from the forward pass gives us the conditional distribution for x_i given all x_j, $j > i$. A subsequent backward pass can then either sample x or compute the needed mean vector and central band of the covariance matrix.

DEFINITION 3.1. *If A is a matrix, then $A[i_1 : i_2)$ is the submatrix of A comprising rows and columns i_1 through $i_2 - 1$. If b is a vector, then $b[i_1 : i_2)$ is the vector comprising elements i_1 through $i_2 - 1$ of b. If $F = (A, b, c)$ then $F[i_1 : i_2) = (A[i_1 : i_2), b[i_1 : i_2), c)$.*

In this paper we number vector elements and matrix rows/columns starting with 0 instead of 1.

DEFINITION 3.2. *Let $F = (A, b, c)$ be an n-variate normal form, $n > 0$, and $0 \le l \le u \le n$. The* $[l, u)$ marginal of F, *written* $\mathrm{marg}(l, u, F)$, *is the unique triple $(\tilde{A}, \tilde{b}, \tilde{c})$ such that*

$$f(y; \tilde{A}, \tilde{b}, \tilde{c}) = \int_{\mathbf{R}^h} f(x; A, b, c) dx_0 \cdots dx_{l-1}\, dx_u \cdots dx_{n-1}, \tag{3.1}$$

for all $(u-l)$*-vectors* y*, where* $h = n - u + l$ *and* $x_i \stackrel{\text{def}}{=} y_{i-l}$ *for* $l \le i < u$*. The* lower marginal *is the* $[1, n)$ *marginal.*

It is well known that the marginals of a multivariate normal distribution are themselves normal distributions, hence the marginals of normal forms exist, are unique, and are themselves normal forms. Note that for the $[u, u)$ marginal we have a 0×0 matrix $\tilde{A}$, a 0-length vector $\tilde{b}$, and $f(y; \tilde{A}, \tilde{b}, \tilde{c}) = \tilde{c}$ where y is a 0-length vector. For convenience, we consider a 0×0 matrix to be normal.

LEMMA 3.3. *Let* $F = (A, b, c)$ *be an* n*-variate normal form,* $n > 0$*. Then* $\text{marg}(1, n, F) = (\tilde{A}, \tilde{b}, \tilde{c})$*, where*

$$\begin{array}{rclcrcl} \tilde{A} & = & A[1:n) + \omega\alpha\alpha^{\mathrm{t}} & & \alpha & = & \alpha'[1:n) \\ \tilde{b} & = & b[1:n) + \omega b_0 \alpha & & \alpha' & = & \textit{column } 0 \textit{ of } A \\ \tilde{c} & = & c + \omega b_0^2/4 + \frac{1}{2}\log(\pi\omega) & & \omega & = & -1/A[0,0]. \end{array}$$

Furthermore, $\text{marg}(1, n, F)$ *is a normal form.*

Proof. Let $\hat{A} = A[1:n)$, $\hat{b} = b[1:n)$, $y = x[1:n)$, and $a = A[0,0]$. Then

$$x^{\mathrm{t}}Ax + b^{\mathrm{t}}x + c = a(x_0 - m(y))^2 + c - am(y)^2 + y^{\mathrm{t}}\hat{A}y + \hat{b}^{\mathrm{t}}y,$$

where $m(y) = -(b_0 + 2\alpha^{\mathrm{t}}y)/(2a)$. Integrating out x_0, we have

$$\int_{-\infty}^{\infty} \exp\left(x^{\mathrm{t}}Ax + b^{\mathrm{t}}x + c\right) dx_0 = \left(\frac{\pi}{-a}\right)^{1/2} \exp\left(y^{\mathrm{t}}\hat{A}y + \hat{b}^{\mathrm{t}}y + c - am(y)^2\right). \tag{3.2}$$

($a < 0$ because A is negative definite.) Some more algebra gives

$$y^{\mathrm{t}}\hat{A}y + \hat{b}^{\mathrm{t}}y + c - am(y)^2 = y^{\mathrm{t}}\tilde{A}y + \left(\tilde{b}\right)^{\mathrm{t}}y + c - \frac{b_0^2}{4a} \tag{3.3}$$

Equations (3.2) and (3.3) give us (3.1) for $l = 1$, $u = n$; hence, $(\tilde{A}, \tilde{b}, \tilde{c})$ is the lower marginal of F.

If $n = 1$ then $\tilde{A}$ is a 0×0 matrix and hence is, by definition, normal. Now consider the case $n > 1$.

If $x_0 = \omega\alpha^{\mathrm{t}}y$, then a bit of algebra shows that $x^{\mathrm{t}}Ax = y^{\mathrm{t}}\tilde{A}y$. If $y \neq 0$ then $x \neq 0$, hence $y^{\mathrm{t}}\tilde{A}y = x^{\mathrm{t}}Ax < 0$. Thus $\tilde{A}$ is negative definite. In addition, since $\hat{A}$ and $\alpha\alpha^{\mathrm{t}}$ are symmetric, so is $\tilde{A}$. Thus $\tilde{A}$ is a normal matrix. □

COROLLARY 3.4. *Let* F *be a bandwidth-*k*,* n*-variate normal form* ($n > k \ge 0$)*. Then* $\text{marg}(1, n, F)$ *also has bandwidth* k*. Furthermore, we can carry out the assignment*

$$F[1:n) \leftarrow \text{marg}(1, n, F);$$

in $O(1 + k^2)$ *time.*

Proof. Referring to Lemma 3.3, we see that since A has bandwidth k, all but the first k elements of α are 0. Thus, $\alpha\alpha^{\mathrm{t}}$ is all zeroes except for its $k \times k$ upper-left submatrix, and has bandwidth $k - 1$. The corollary immediately follows. □

The algorithms in the following sections apply Corollary 3.4 repeatedly, producing the $[0, 0)$ marginal as a byproduct in $O(n(1+k^2))$ time. The scalar component of this marginal is exactly the logarithm of the integral to be computed for Task 4.

4. Sampling. To sample x from the full multivariate normal distribution corresponding to a normal form F, we successively sample the conditional distribution for x_p given $x[p+1:n)$, as p runs from $n-1$ to 0. We find that this requires only the first element of $\tilde{b}$ and the first row/column of $\tilde{A}$, where $(\tilde{A}, \tilde{b}, \tilde{c})$ is the $[p, n)$ marginal of F.

LEMMA 4.1. *Let $F = (A, b, c)$ be an n-variate normal form. Consider the multivariate normal distribution over R^n proportional to $f(x; A, b, c)$. The conditional distribution for x_p $(0 \le p < n)$ given $x[p+1:n)$ is the normal distribution with mean μ and precision λ, where*

$$\begin{aligned} \lambda &= -2\tilde{A}[p,p] \\ \mu &= \lambda^{-1}\left(\tilde{b}_p + 2\sum_{i>p}\tilde{A}[i,p]x_i\right) \end{aligned} \tag{4.1}$$

and $(\tilde{A}, \tilde{b}, \tilde{c}) = \mathrm{marg}(p, n, F)$. Furthermore, if F has bandwidth k, then μ and λ can be computed in $O(1+k)$ time.

Proof. The conditional pdf is proportional to the marginal pdf for $y = x[p:n)$, which itself is proportional to $f(y; \tilde{A}, \tilde{b}, \tilde{c})$. Substituting in constant (known) values for $x[p+1:n)$, all of the exponent *terms* that don't involve x_p become constant *factors* that we can ignore. The conditional pdf for x_p is then proportional to

$$\exp\left(\tilde{A}[p,p]x_p^2 + 2\sum_{i>p}\tilde{A}[i,p]x_i x_p + \tilde{b}_p x_p\right).$$

Completing the square, this is proportional to

$$\exp\left(\tilde{A}[p,p]\left(x_p + \frac{\tilde{b}_p + 2\sum_{i>p}\tilde{A}[i,p]x_i}{2\tilde{A}[p,p]}\right)^2\right).$$

Matching this with the form $\exp(-\frac{1}{2}\lambda(x_p - \mu)^2)$ for a normal pdf gives us the first part of the lemma. The time-complexity bound follows from the fact that all but the first k terms of the sum for computing μ are zero. □

This now gives us our sampling algorithm.

THEOREM 4.2. *Let $F = (A, b, c)$ be a bandwidth-k, n-variate normal form, where $n > k \ge 0$. Then the algorithm of Figure 4.1 samples from the multivariate normal distribution corresponding to F in $O(n(1+k^2))$ time.*

Proof. The first `for` loop computes all $[p, n)$ marginals of F as p runs from 0 to n, nesting them inside of each other in a fashion that retains only the information we need for efficient sampling (Figure 4.2). Let $F^{(j)} = (A^{(j)}, b^{(j)}, c^{(j)}) = \mathrm{marg}(j, n, F)$. We have the following as an invariant of the first `for` loop:

1. $\tilde{F}[p:n) = F^{(p)}$.
2. For $0 \le j < p$ and $j \le i < n$, $\tilde{A}[i,j] = A^{(j)}[i,j]$.
3. For $0 \le j < p$, $\tilde{b}_j = b_j^{(j)}$.

As a result, the following are established by the end of the first `for` loop and hold throughout the second `for` loop (lines 5–10):

1. For all $0 \le j \le i < n$, $\tilde{A}[i,j] = A^{(j)}[i,j]$.
2. For all $0 \le j < n$, $\tilde{b}_j = b_j^{(j)}$.

Require: $n > k \geq 0$, $F = (A, b, c)$ is a bandwidth-k, n-variate normal form.
Ensure: $x[0:n)$ randomly drawn from normal distribution corresponding to F.
1: $\tilde{F} \leftarrow F$; $\{(\tilde{A}, \tilde{b}, \tilde{c}) = \tilde{F}$ by definition$\}$
2: **for** $p \leftarrow 0$ to $n-1$ **do**
3: $\quad \tilde{F}[p+1:n) \leftarrow \mathrm{marg}(1, n-p, \tilde{F}[p:n))$;
4: **end for**
5: **for** $p \leftarrow n-1$ to 0 **do**
6: $\quad \lambda \leftarrow -2\tilde{A}[p,p]$;
7: $\quad m \leftarrow \max(n-1, p+k)$;
8: $\quad \mu \leftarrow \lambda^{-1}\left(\tilde{b}_p + 2\sum_{p<i\leq m} \tilde{A}[i,p]x_i\right)$;
9: $\quad x_p \leftarrow$ sample from normal with mean μ and precision λ;
10: **end for**

FIG. 4.1. *Random generation of vectors from pdf corresponding to F*

Thus, lines 6 and 8 are equivalent to

1. $\lambda \leftarrow A^{(p)}[p,p]$ and
2. $\mu \leftarrow \lambda^{-1}\left(b_p^{(p)} + 2\sum_{p<i\leq m} A^{(p)}[i,p]x_i\right)$;

combining this with Lemma 4.1, this means that in line 9 we are sampling from the conditional distribution over x_p given $x[p+1:n)$, based on the multivariate normal distribution corresponding to F. Since each x_p is sampled only after $x[p+1:n)$ has been sampled, the resulting vector $x[0:n)$ is a sample from the desired normal distribution.

Copying F to $\tilde{F}$ takes $O(n(1+k))$ time, assuming we use any efficient band-matrix representation. By Corollary 3.4, line 3 takes $O(1+k^2)$ time, and so the first `for` loop takes $O(n(1+k^2))$ time. Lines 6–9 take $O(1+k)$ time altogether, and so the second `for` loop takes $O(n(1+k))$ time. Thus the entire algorithm takes $O(n(1+k^2))$ time. Note that if we wish to draw multiple samples, the setup in lines 1–4 need not be repeated, so we can draw additional samples in $O(n(1+k))$ time each. □

5. Means and Variances. We now present a solution for Task 2 that also solves Task 1 with little additional work. We begin with a corollary to Lemma 4.1.

COROLLARY 5.1. *Let $F = (A, b, c)$ be an n-variate normal form. Let μ_i and $\sigma_{i,j}^2$, $0 \leq i, j < n$, be the means and covariances for the corresponding multivariate normal distribution. Let $(\tilde{A}, \tilde{b}, \tilde{c}) = \mathrm{marg}(p, n, F)$ and $\lambda = -2\tilde{A}[p,p]$. Then for $j > p$,*

$$\mu_p = \lambda^{-1}\left(\tilde{b}_p + 2\sum_{i>p} \tilde{A}[i,p]\mu_i\right)$$
$$\sigma_{p,p}^2 = \lambda^{-1} + 4\lambda^{-2}\sum_{i,h>p} \tilde{A}[i,p]\tilde{A}[h,p]\sigma_{i,h}^2$$
$$\sigma_{p,j}^2 = 2\lambda^{-1}\sum_{i>p} \tilde{A}[i,p]\sigma_{i,j}^2.$$

Proof. The formula for μ_p follows directly from equation (4.1) of Lemma 4.1, and the linearity of the expectation operator. The formula for $\sigma_{p,j}^2$ follows from equation (4.1) and the fact that the covariance operator is linear in each argument: $C[ax, y] = aC[x, y]$ and $C[x + x', y] = C[x, y] + C[x', y]$. Lemma 4.1 tells us that we can write $x_p = e_p + \mu_p^*$, where μ_p^* is the conditional mean of x_p given $x[p+1:n)$

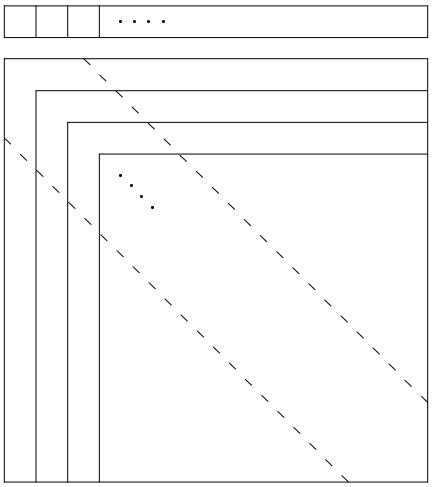

FIG. 4.2. *Nesting of marginals in the algorithm of Figure 4.1*

from equation (4.1), and e_p is a normal random variable independent of $x[p+1:n)$, with variance λ^{-1}. The formula for $\sigma^2_{p,p}$ then follows from the fact that the covariance operator is linear in each argument. □

This now gives us an algorithm to compute the mean vector and central band of the covariance matrix for the normal distribution corresponding to a normal form.

THEOREM 5.2. *Let $F = (A, b, c)$ be a bandwidth-k, n-variate normal form, where $n > \kappa \geq k \geq 0$. Then the algorithm of Figure 5.1 computes the mean vector and and a central band of the covariance matrix of bandwidth κ in $O(n(1+k)(1+\kappa))$ time.*

Proof. Let $F^{(j)} = (A^{(j)}, b^{(j)}, c^{(j)}) = \mathrm{marg}(j, n, F)$. The first `for` loop of Figure 5.1 is identical to the first `for` loop of Figure 4.1, and as in Theorem 4.2 it establishes the following condition that holds throughout the remainder of the algorithm:

1. For all $0 \leq j \leq i < n$, $\tilde{A}[i,j] = A^{(j)}[i,j]$.
2. For all $0 \leq j < n$, $\tilde{b}_j = b^{(j)}_j$.

Thus, we may replace every use of $\tilde{A}$ and $\tilde{b}$ in lines 6, 8, 9, and 11 with $A^{(p)}$ and $b^{(p)}$ respectively. Combining this with Corollary 5.1, it follows that we have the following as an invariant of the second `for` loop (lines 5–14):

1. μ_i is element i of the mean vector for all $i > p$;
2. $C[i,j]$ is the covariance of x_i and x_j for all $i, j > p$ and $|i-j| \leq \kappa$.

Substituting $p = -1$ upon termination of the second `for` loop into the above invariant gives us that $\mu[0:n)$ holds the mean vector and C holds the desired central band of the covariance matrix.

As in Theorem 4.2, lines 1–4 take $O(n(1+k^2))$ time. Lines 6–9 take $O(1+k^2)$ time, and the loop of lines 10–13 takes $O((1+k)(1+\kappa))$ time. Since $k \leq \kappa$, $O(1+k^2) = O((1+k)(1+\kappa))$. Thus, the entire loop in lines 5–14 takes $O(n(1+k)(1+\kappa))$ time; combined with the time bound for the first loop, this gives the time complexity bound of the theorem. □

REFERENCES

[1] S. DELLA PIETRA, V. DELLA PIETRA, AND J. LAFFERTY, *Inducing features of random fields*, IEEE Transactions on Patterna Analysis and Machine Intelligence, 19 (1997), pp. 1–13.

Require: $n > \kappa \geq k \geq 0$, $F = (A, b, c)$ is a bandwidth-k, n-variate normal form.
Ensure: μ_i is the mean of x_i and $C[i,j]$ is the covariance of x_i and x_j for $0 \leq i, j < n$ and $|i - j| \leq \kappa$, where x is a vector random variable whose pdf corresponds to F.

$\tilde{F} \leftarrow F$; $\{(\tilde{A}, \tilde{b}, \tilde{c}) = \tilde{F}$ by definition$\}$
for $p \leftarrow 0$ to $n - 1$ **do**
  $\tilde{F}[p + 1 : n) \leftarrow \text{marg}(1, n - p, \tilde{F}[p : n))$;
end for
for $p \leftarrow n - 1$ to 0 **do**
  $\lambda \leftarrow -2\tilde{A}[p, p]$;
  $m \leftarrow \max(n - 1, p + k)$;
  $\mu_p \leftarrow \lambda^{-1} \left(\tilde{b}_p + 2 \sum_{p < i \leq m} \tilde{A}[i, p] \mu_i \right)$;
  $C[p, p] \leftarrow \lambda^{-1} + 4\lambda^{-2} \sum_{p < i, h \leq m} \tilde{A}[i, p] \tilde{A}[h, p] C[i, h]$;
  for $j \leftarrow p + 1$ to $\max(n - 1, p + \kappa)$ **do**
    $C[p, j] \leftarrow 2\lambda^{-1} \sum_{p < i \leq m} \tilde{A}[i, p] C[i, j]$;
    $C[j, p] \leftarrow C[p, j]$;
  end for
end for

FIG. 5.1. *Computation of mean vector and central band of covariance matrix*

[2] E. T. JAYNES, *Information theory and statistical mechanics I*, Physical Review, 106 (1957), pp. 620–630.
[3] Z. LI AND B. D'AMBROSIO, *Efficient inference in bayes networks as a combinatorial optimization problem*, International Journal of Approximate Reasoning, 11 (1994), pp. 55–81.
[4] M. A. TANNER, *Tools for Statistical Inference*, Springer-Verlag, New York, NY, third ed., 1996.
[5] K. S. VAN HORN, *A maximum-entropy acoustic model for speech recognition*, tech. report, Dept. of Computer Science, North Dakota State University, Fargo, ND, Nov. 2001.