# An Introduction to Probabilistic Graphical Models

Michael I. Jordan
*University of California, Berkeley*

June 30, 2003

# Chapter 13

# The Multivariate Gaussian

In this chapter we present some basic facts regarding the multivariate Gaussian distribution. We discuss the two major parameterizations of the multivariate Gaussian—the *moment parameterization* and the *canonical parameterization*, and we show how the basic operations of marginalization and conditioning are carried out in these two parameterizations. We also discuss maximum likelihood estimation for the multivariate Gaussian.

## 13.1 Parameterizations

The multivariate Gaussian distribution is commonly expressed in terms of the parameters $\mu$ and $\Sigma$, where $\mu$ is an $n \times 1$ vector and $\Sigma$ is an $n \times n$, symmetric matrix. (We will assume for now that $\Sigma$ is also positive definite, but later on we will have occasion to relax that constraint). We have the following form for the density function:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \tag{13.1}$$

where $x$ is a vector in $\Re^n$. The density can be integrated over volumes in $\Re^n$ to assign probability mass to those volumes.

The geometry of the multivariate Gaussian is essentially that associated with the quadratic form $f(x) = \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)$ in the exponent of the density. Recall our discussion in Chapter 6, where we showed that a quadratic form $f(x)$ is a paraboloid with level surfaces, i.e, surfaces of the form $f(x) = c$ for fixed $c$, being ellipsoids oriented along the eigenvectors of the matrix $\Sigma$. Now note that the exponential, $\exp(\cdot)$, is a scalar function that leaves the geometrical features of the quadratic form intact. That is, for any $x$ lying on an ellipsoid $f(x) = c$, we obtain the value $\exp\{-c\}$. The maximum value of the exponential is 1, obtained at $x = \mu$ where $f(x) = 0$. The paraboloid $f(x)$ increases to infinity as we move away from $x = \mu$; thus we obtain a "bump" in $(n+1)$–dimensional space centered at $x = \mu$. The level surfaces of the Gaussian bump are ellipsoids oriented along the eigenvectors of $\Sigma$.

The factor in front of the exponential in Eq. 13.1 is the normalization factor that ensures that the density integrates to one. To show that this factor is correct, we make use of the diagonalization of $\Sigma^{-1}$. Diagonalization yields a product of $n$ univariate Gaussians whose standard deviations are the eigenvalues of $\Sigma$. When we integrate, each of these univariate Gaussians contributes a factor $\sqrt{2\pi}\lambda_i$ to the normalization, where $\lambda_i$ is the $i$th eigenvalue of $\Sigma$. Recall that the determinant of a matrix is the product of its eigenvalues to obtain the result. (We ask the reader to fill in the details of this derivation in Exercise **??**).

As in the univariate case, the parameters $\mu$ and $\Sigma$ have a probabilistic interpretation as the *moments* of the Gaussian distribution. In particular, we have the important result:

$$\mu = E(x) \tag{13.2}$$
$$\Sigma = E(x-\mu)(x-\mu)^T. \tag{13.3}$$

We will not bother to derive this standard result, but will provide a hint: diagonalize and appeal to the univariate case.

Although the moment parameterization of the Gaussian will play a principal role in our subsequent development, there is a second parameterization—the canonical parameterization—that will also be important. In particular, expanding the quadratic form in Eq. 13.1, and defining *canonical parameters*:

$$\Lambda = \Sigma^{-1} \tag{13.4}$$
$$\eta = \Sigma^{-1}\mu, \tag{13.5}$$

we obtain:

$$p(x|\eta,\Lambda) = \exp\left\{a + \eta^T x - \frac{1}{2}x^T\Lambda x\right\}, \tag{13.6}$$

where $a = -1/2(n\log(2\pi) - \log|\Lambda| + \eta^T\Lambda\eta)$ is the normalizing constant in this representation. The canonical parameterization is also sometimes referred to as the *information parameterization.*

We can also convert from canonical parameters to moment parameters:

$$\mu = \Lambda^{-1}\eta \tag{13.7}$$
$$\Sigma = \Lambda^{-1}. \tag{13.8}$$

Moment parameters and canonical parameters are useful in different circumstances. As we will see, different kinds of transformations are more readily carried out in one representation or the other.

## 13.2 Joint distributions

Suppose that we partition the $n \times 1$ vector $x$ into a $p \times 1$ subvector $x_1$ and a $q \times 1$ subvector $x_2$, where $n = p + q$. Form corresponding partitions of the $\mu$ and $\Sigma$ parameters:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{13.9}$$

We can write a joint Gaussian distribution for $x_1$ and $x_2$ using these partitioned parameters:

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{(p+q)/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \right\} \tag{13.10}$$

This partitioned form of the joint distribution raises a number of questions. In particular, we can equally well form partitioned versions of $\eta$ and $\Lambda$ and express the joint distribution in the canonical parameterization; is there any relationship between the partitioned form of these two representations? Also, what do the blocks in the partitioned forms have to do with the marginal and conditional probabilities of $x_1$ and $x_2$?

These questions all involve the manipulation of the quadratic forms in the exponents of the Gaussian densities; indeed, the underlying algebraic problem is that of "completing the square" of quadratic forms. In the next section, we discuss an algebra that provides a general solution to the problem of "completing the square."

## 13.3 Partitioned matrices

Our first result in this section is to show how to block diagonalize a partitioned matrix. A number of useful results flow from this operation, including an explicit expression for the inverse of a partitioned matrix.

Consider a general partitioned matrix:

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix}, \tag{13.11}$$

where we assume that both $E$ and $H$ are invertible. (Our results can be generalized beyond this setting). To invert this matrix, we follow a similar procedure to that of diagonalization. In particular, we wish to *block diagonalize* the matrix. We wish to put a block of zeros in place of $G$ and a block of zeros in place of $F$.

To zero out the upper-right-hand corner of $M$, note that it suffices to *premultiply* the second block column of $M$ by a "block row vector" having elements $I$ and $-FH^{-1}$. Similarly, to zero out the lower-left-hand corner of $M$, if suffices to *postmultiply* the second row of $M$ by a "block column vector" having elements $I$ and $-H^{-1}G$. The magical fact is that these two operations do not interfere with each other; thus we can block diagonalize $M$ by doing both operations. In particular, we have:

$$\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix}. \tag{13.12}$$

The correctness of this decomposition can be verified directly.

We define the *Schur complement* of the matrix $M$ with respect to $H$, denoted $M/H$, as the term $E - FH^{-1}G$ that appears in the block diagonal matrix. It is not difficult to show that $M/H$ is invertible.

We now take the inverse of both sides of Eq. 13.12. Note that in a matrix expression of the form $XYZ = W$ inverting both sides yields $Y^{-1} = ZW^{-1}X$; this implies that we don't need to explicitly invert the block triangular matrices in Eq. 13.12 (although this is easily done). Note also that the inverse of a block diagonal matrix is the diagonal matrix of the inverse of its blocks. Thus we have:

$$\begin{aligned} \begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} &= \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} && (13.13) \\ &= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix}, && (13.14) \end{aligned}$$

which expresses the inverse of a partitioned matrix in terms of its blocks.

We can also apply the determinant operator to both sides of Eq. 13.12. The block triangular matrices clearly have a determinant of one; thus we obtain another important result:

$$|M| = |M/H||H|. \tag{13.15}$$

(This result makes the choice of notation for the Schur complement seem quite natural!)

We are not yet finished. Note that we could alternatively have decomposed the matrix $M$ in terms of $E$ and $M/E$, yielding the following expression for the inverse:

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix}. \tag{13.16}$$

These two expressions for the inverse of $M$ (Eq. 13.14 and Eq. 13.16) must be the same, thus we can set the corresponding blocks equal to each other. This yields:

$$(E - FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1} \tag{13.17}$$

and

$$(E - FH^{-1}G)^{-1}FH^{-1} = E^{-1}F(H - GE^{-1}F)^{-1} \tag{13.18}$$

Both Eq. 13.17, which is generally referred to as the "matrix inversion lemma," and Eq. 13.18 are quite useful in transformations involving Gaussian distributions. They allow expressions involving the inverse of $E$ to be converted into expressions involving the inverse of $H$ and vice versa.

## 13.4 Marginalization and conditioning

We now make use of our block diagonalization results to develop general formulas for the key operations of marginalization and conditioning in the multivariate Gaussian setting. We present results for both the moment parameterization and the canonical parameterization.

Our goal is to split the joint distribution Eq. 13.10 into a marginal probability for $x_2$ and a conditional probability for $x_1$ according to the factorization $p(x_1, x_2) = p(x_1|x_2)p(x_2)$. Focusing first on the exponential factor, we make use of Eq. 13.12:

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}^T \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} \begin{bmatrix} (\Sigma/\Sigma_{22})^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right\} \\
&= \exp\left\{-\frac{1}{2}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))^T(\Sigma/\Sigma_{22})^{-1}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))\right\} \\
&\qquad \cdot \exp\left\{-\frac{1}{2}(x_2-\mu_2)^T\Sigma_{22}^{-1}(x_2-\mu_2)\right\}. \qquad (13.19)
\end{aligned}$$

We next exploit Eq. 13.15 to split the normalization into two factors:

$$\frac{1}{(2\pi)^{(p+q)/2}|\Sigma|^{1/2}} = \frac{1}{(2\pi)^{(p+q)/2}(|\Sigma/\Sigma_{22}||\Sigma_{22}|)^{1/2}} \qquad (13.20)$$

$$= \left(\frac{1}{(2\pi)^{p/2}|\Sigma/\Sigma_{22}|^{1/2}}\right)\left(\frac{1}{(2\pi)^{q/2}|\Sigma_{22}|^{1/2}}\right) \qquad (13.21)$$

To see that we have achieved our goal of factorizing the joint distribution into the product of a marginal distribution and a conditional distribution, note that if we group the first factor in Eq. 13.19 with the first factor in Eq. 13.21 we obtain a normalized Gaussian in the variable $x_1$. Integrating with respect to $x_1$, these factors disappear and the remaining factors must therefore represent the marginal distribution of $x_2$:

$$p(x_2) = \frac{1}{(2\pi)^{q/2}|\Sigma/\Sigma_{22}|^{1/2}} \exp\left\{-\frac{1}{2}(x_2-\mu_2)^T\Sigma_{22}^{-1}(x_2-\mu_2)\right\}. \qquad (13.22)$$

Given this result, we are now licensed to interpret the factors that were integrated over as the conditional probability $p(x_1|x_2)$:

$$p(x_1|x_2) = \frac{1}{(2\pi)^{p/2}|\Sigma_{22}|^{1/2}} \exp\left\{-\frac{1}{2}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))^T(\Sigma/\Sigma_{22})^{-1}(x_1-\mu_1-\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2))\right\} \qquad (13.23)$$

To summarize our results, let $(\mu_2^m, \Sigma_2^m)$ denote the moment parameters of the marginal distribution of $x_2$, and let $(\mu_{1|2}^c, \Sigma_{1|2}^c)$ denote the moment parameters of the conditional distribution of $x_1$ given $x_2$. Eq. 13.22 and Eq. 13.23 yield the following expressions for these parameters:

**Marginalization**:

$$\mu_2^m = \mu_2 \qquad (13.24)$$

$$\Sigma_2^m = \Sigma_{22} \qquad (13.25)$$

**Conditioning**:

$$\mu_{1|2}^c = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) \qquad (13.26)$$

$$\Sigma_{1|2}^c = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad (13.27)$$

We can also express the marginalization and conditioning operations in the canonical parameterization. The results can be obtained directly from the joint distribution, or indirectly by converting from the results for the moment parameterization. In any case, the results—which we ask the reader to derive in Exercise **??**—are as follows:

**Marginalization**:

$$\begin{aligned} \eta_2^m &= \eta_2 - \Lambda_{21}\Lambda_{11}^{-1}\eta_1 && (13.28) \\ \Lambda_2^m &= \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12} && (13.29) \end{aligned}$$

**Conditioning**:

$$\begin{aligned} \eta_{1|2}^c &= \eta_1 - \Lambda_{12}x_2 && (13.30) \\ \Lambda_{1|2}^c &= \Lambda_{11} && (13.31) \end{aligned}$$

Note that the marginalization operation is simple in the moment parameterization and conditioning is complicated, whereas the opposite holds in the canonical parameterization.

## 13.5 Maximum likelihood estimation

Let us now suppose that we have an independently, identically distributed data set $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ sampled from a multivariate Gaussian distribution. We want to estimate the parameters of the distribution via maximum likelihood.

We form the log likelihood function by taking the logarithm of the product of $N$ Gaussians:

$$l(\mu, \Sigma \,|\, \mathcal{D}) = -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i - \mu)^T\Sigma^{-1}(x_i - \mu). \tag{13.32}$$

Taking the derivative with respect to $\mu$ is straightforward:

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N}(x_i - \mu)^T\Sigma^{-1}, \tag{13.33}$$

and setting to zero we obtain a pleasant result:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{13.34}$$

That is, the maximum likelihood estimate of the mean is the sample mean.

We now turn to the more challenging problem of computing the maximum likelihood estimate of the covariance matrix. Although the tools that we require to solve this problem are inevitably somewhat technical, they are important, and will reappear on several occasions later in the text.

Letting $l(\Sigma \,|\, \mathcal{D})$ denote those terms in the log likelihood that are a function of $\Sigma$, we obtain:

$$l(\Sigma \,|\, \mathcal{D}) = -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_n (x_n - \mu)^T \Sigma^{-1} (x_n - \mu). \tag{13.35}$$

We need to take the derivative with respect to $\Sigma$ and set to zero. To calculate this derivative we require a short detour.

### 13.5.1 Traces and derivatives

The trace of a square matrix $A$ is defined to be the sum of the diagonal elements $a_{ii}$ of $A$:

$$\mathrm{tr}[A] \triangleq \sum_i a_{ii} \tag{13.36}$$

An important property possessed by the trace is its invariance under cyclical permutations of matrix products:

$$\mathrm{tr}[ABC] = \mathrm{tr}[CAB] = \mathrm{tr}[BCA], \tag{13.37}$$

where $A$, $B$ and $C$ are arbitrary matrices whose dimensions are conformal and are such that the product $ABC$ (and therefore the other two products) is a square matrix. This result is easily established by writing out the matrix products explicitly.

Our interest in the trace is due to its usefulness in calculating derivatives of quadratic forms. In particular, we need to take the derivative of an expression of the form $x^T A x$ with respect to the matrix $A$. There is a "trick" involving the trace that makes such calculations easy. We write:

$$x^T A x = \mathrm{tr}[x^T A x] = \mathrm{tr}[x x^T A] \tag{13.38}$$

where the first equality follows from the fact that $x^T A x$ is a scalar. We see that we can calculate derivatives of quadratic forms by calculating derivatives of traces.

Let us calculate the derivative of $\mathrm{tr}[BA]$ with respect to $A$. We write out the matrix product explicitly and calculate the partial with respect to a matrix element $a_{ij}$:

$$\frac{\partial}{\partial a_{ij}}\mathrm{tr}[AB] = \frac{\partial}{\partial a_{ij}} \sum_k \sum_l a_{kl} b_{lk} \tag{13.39}$$

$$= b_{ji}, \tag{13.40}$$

which shows that

$$\frac{\partial}{\partial A}\mathrm{tr}[BA] = B^T. \tag{13.41}$$

We can now use this result to calculate the derivative of the quadratic form $x^T A x$:

$$\frac{\partial}{\partial A} x^T A x = \frac{\partial}{\partial A}\mathrm{tr}[x x^T A] = [x x^T]^T = x x^T, \tag{13.42}$$

which is the outer product of the vector $x$ with itself.

### 13.5.2 Determinants and derivatives

From Eq. (13.35) we see that we also need to take the derivative of the logarithm of a determinant. In this section we establish the following result:

$$\frac{\partial}{\partial A} \log |A| = A^{-T}, \tag{13.43}$$

which together with the result on traces will allow us to solve our estimation problem.

To establish the result, note that:

$$\frac{\partial}{\partial a_{ij}} \log |A| = \frac{1}{|A|} \frac{\partial}{\partial a_{ij}} |A| \tag{13.44}$$

and recall the formula for the matrix inverse:

$$A^{-1} = \frac{1}{|A|} \tilde{A}, \tag{13.45}$$

where $\tilde{A}$ is the matrix of cofactors. This shows that we need only establish the following result:

$$\frac{\partial}{\partial a_{ij}} |A| = \tilde{A}. \tag{13.46}$$

The determinant of a square matrix $A$ can be expanded as follows (this formula is often taken as the definition of the determinant):

$$|A| = \sum_j (-1)^{i+j} a_{ij} M_{ij} \tag{13.47}$$

where $M_{ij}$ is the minor associated with matrix element $a_{ij}$ (the determinant of the matrix obtained by removing the $i$th row and $j$th column of $A$). Note that this formula holds for arbitrary $i$.

Given that $a_{ij}$ does not appear in any of the minors in the sum, the derivative of $|A|$ with respect to $a_{ij}$ is just $(-1)^{i+j} M_{ij}$. But the matrix of these values is simply the transpose of the matrix of cofactors.

### 13.5.3 Maximum likelihood estimate of $\Sigma$

With these two results in hand, we return to the problem of calculating the maximum likelihood estimate of the covariance matrix $\Sigma$.

We use the trace trick to cope with the quadratic form:

$$\begin{aligned} l(\Sigma \,|\, \mathcal{D}) &= -\frac{N}{2} \log |\Sigma| - \frac{1}{2} \sum_n (x_n - \mu)^T \Sigma^{-1} (x_n - \mu) && (13.48) \\ &= \frac{N}{2} \log |\Sigma^{-1}| - \frac{1}{2} \sum_n \mathrm{tr}[(x_n - \mu)^T \Sigma^{-1} (x_n - \mu)] && (13.49) \\ &= \frac{N}{2} \log |\Sigma^{-1}| - \frac{1}{2} \sum_n \mathrm{tr}[(x_n - \mu)(x_n - \mu)^T \Sigma^{-1}], && (13.50) \end{aligned}$$

where we have also used the fact that the determinant of the inverse of a matrix is the inverse of the determinant.

We now take the derivative with respect to the matrix $\Sigma^{-1}$:

$$\frac{\partial l}{\partial \Sigma^{-1}} = \frac{N}{2}\Sigma - \frac{1}{2}\sum_n (x_n - \mu)(x_n - \mu)^T. \tag{13.51}$$

Finally, setting to zero yields the maximum likelihood estimator:

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_n (x_n - \hat{\mu}_{ML})(x_n - \hat{\mu}_{ML})^T, \tag{13.52}$$

which is the expected result.

This result also allows us to obtain maximum likelihood estimates for the canonical parameters:

$$\hat{\Lambda} = \hat{\Sigma}_{ML}^{-1} \tag{13.53}$$

$$\hat{\eta} = \hat{\Sigma}_{ML}^{-1}\hat{\mu}_{ML}; \tag{13.54}$$

where we have used the fact that maximum likelihood estimates are invariant to changes in parameterization (see Exercise **??**).

There is much more to say about maximum likelihood estimation for the multivariate Gaussian. In particular, it is significantly more interesting to obtain estimates of $\Sigma$ when we have constraints on $\Lambda$, for example a constraint that requires that certain elements of $\Lambda$ are zero. Indeed, as we will see in Chapter **??**, this is a rather natural constraint in the world of graphical models.

## 13.6 Historical remarks and bibliography

[section yet to be written].