7 Gaussian Network Models

Although much of our presentation focuses on discrete variables, we mentioned in chapter 5 that the Bayesian network framework, and the associated results relating independencies to factorization of the distribution, also apply to continuous variables. The same statement holds for Markov networks. However, whereas table CPDs provide a general-purpose mechanism for describing any discrete distribution (albeit potentially not very compactly), the space of possible parameterizations in the case of continuous variables is essentially unbounded. In this chapter, we focus on a type of continuous distribution that is of particular interest: the class of multivariate Gaussian distributions. Gaussians are a particularly simple subclass of distributions that make very strong assumptions, such as the exponential decay of the distribution away from its mean, and the linearity of interactions between variables. While these assumptions are often invalid, Gaussians are nevertheless a surprisingly good approximation for many real-world distributions. Moreover, the Gaussian distribution has been generalized in many ways, to nonlinear interactions, or mixtures of Gaussians; many of the tools developed for Gaussians can be extended to that setting, so that the study of Gaussian provides a good foundation for dealing with a broad class of distributions.

In the remainder of this chapter, we first review the class of multivariate Gaussian distributions and some of its properties. We then discuss how a multivariate Gaussian can be encoded using probabilistic graphical models, both directed and undirected.

7.1 Multivariate Gaussians

7.1.1 Basic Parameterization

We have already described the univariate Gaussian distribution in chapter 2. We now describe its generalization to the multivariate case. As we discuss, there are two different parameterizations for a joint Gaussian density, with quite different properties.

The univariate Gaussian is defined in terms of two parameters: a mean and a variance. In its most common representation, a multivariate Gaussian distribution over $X_1, \ldots, X_n$ is characterized by an n-dimensional *mean vector* $\boldsymbol{\mu}$, and a symmetric $n \times n$ *covariance matrix* Σ; the density function is most often defined as:

mean vector

covariance matrix

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{7.1}$$

where $|\Sigma|$ is the determinant of Σ.

standard Gaussian

We extend the notion of a *standard Gaussian* to the multidimensional case, defining it to be a Gaussian whose mean is the all-zero vector $\mathbf{0}$ and whose covariance matrix is the identity matrix I, which has 1's on the diagonal and zeros elsewhere. The multidimensional standard Gaussian is simply a product of independent standard Gaussians for each of the dimensions.

positive definite

In order for this equation to induce a well-defined density (that integrates to 1), the matrix Σ must be *positive definite*: for any $\boldsymbol{x} \in \mathbb{R}^n$ such that $\boldsymbol{x} \neq 0$, we have that $\boldsymbol{x}^T \Sigma \boldsymbol{x} > 0$. Positive definite matrices are guaranteed to be nonsingular, and hence have nonzero determinant, a necessary requirement for the coherence of this definition. A somewhat more complex definition

positive semi-definite

can be used to generalize the multivariate Gaussian to the case of a *positive semi-definite* covariance matrix: for any $\boldsymbol{x} \in \mathbb{R}^n$, we have that $\boldsymbol{x}^T \Sigma \boldsymbol{x} \geq 0$. This extension is useful, since it allows for singular covariance matrices, which arise in several applications. For the remainder of our discussion, we focus our attention on Gaussians with positive definite covariance matrices.

information matrix

Because positive definite matrices are invertible, one can also utilize an alternative parameterization, where the Gaussian is defined in terms of its inverse covariance matrix $J = \Sigma^{-1}$, called *information matrix* (or *precision matrix*). This representation induces an alternative form for the Gaussian density. Consider the expression in the exponent of equation (7.1):

$$\begin{aligned} -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T J (\boldsymbol{x} - \boldsymbol{\mu}) \\ &= -\frac{1}{2}\left[\boldsymbol{x}^T J \boldsymbol{x} - 2\boldsymbol{x}^T J \boldsymbol{\mu} + \boldsymbol{\mu}^T J \boldsymbol{\mu}\right]. \end{aligned}$$

The last term is constant, so we obtain:

$$p(\boldsymbol{x}) \propto \exp\left[-\frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x} + (J\boldsymbol{\mu})^T \boldsymbol{x}\right]. \tag{7.2}$$

information form

This formulation of the Gaussian density is generally called the *information form*, and the vector $\boldsymbol{h} = J\boldsymbol{\mu}$ is called the *potential vector*. The information form defines a valid Gaussian density if and only if the information matrix is symmetric and positive definite, since Σ is positive definite if and only if Σ^{-1} is positive definite. The information form is useful in several settings, some of which are described here.

Intuitively, a multivariate Gaussian distribution specifies a set of ellipsoidal contours around the mean vector $\boldsymbol{\mu}$. The contours are parallel, and each corresponds to some particular value of the density function. The shape of the ellipsoid, as well as the "steepness" of the contours, are determined by the covariance matrix Σ. Figure 7.1 shows two multivariate Gaussians, one where the covariances are zero, and one where they are positive. As in the univariate case, the mean vector and covariance matrix correspond to the first two moments of the normal distribution. In matrix notation, $\boldsymbol{\mu} = \boldsymbol{E}[\boldsymbol{X}]$ and $\Sigma = \boldsymbol{E}[\boldsymbol{X}\boldsymbol{X}^T] - \boldsymbol{E}[\boldsymbol{X}]\boldsymbol{E}[\boldsymbol{X}]^T$. Breaking this expression down to the level of individual variables, we have that μ_i is the mean of X_i, $\Sigma_{i,i}$ is the variance of X_i, and $\Sigma_{i,j} = \Sigma_{j,i}$ (for $i \neq j$) is the *covariance* between X_i and X_j: $\boldsymbol{Cov}[X_i; X_j] = \boldsymbol{E}[X_i X_j] - \boldsymbol{E}[X_i]\boldsymbol{E}[X_j]$.

Example 7.1

Consider a particular joint distribution $p(X_1, X_2, X_3)$ over three random variables. We can

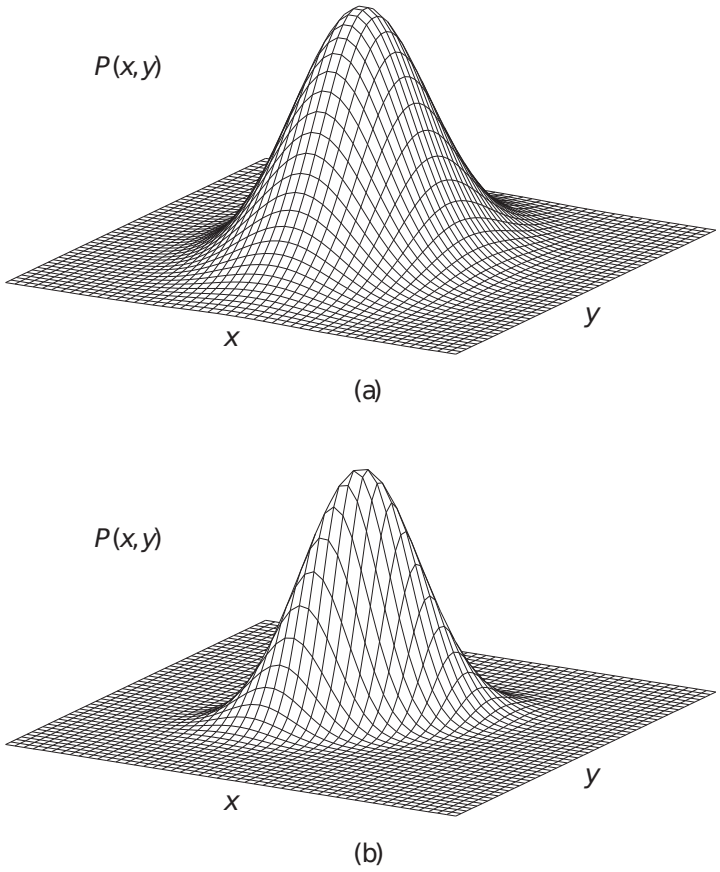

Figure 7.1 Gaussians over two variables X and Y. (a) X and Y uncorrelated. (b) X and Y correlated.

parameterize it via a mean vector $\boldsymbol{\mu}$ and a covariance matrix Σ:

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ -3 \\ 4 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & -5 \\ -2 & -5 & 8 \end{pmatrix}$$

As we can see, the covariances $\mathbf{Cov}[X_1; X_3]$ and $\mathbf{Cov}[X_2; X_3]$ are both negative. Thus, X_3 is negatively correlated with X_1: when X_1 goes up, X_3 goes down (and similarly for X_3 and X_2). ■

7.1.2 Operations on Gaussians

There are two main operations that we wish to perform on a distribution: compute the marginal distribution over some subset of the variables $\boldsymbol{Y}$, and conditioning the distribution on some assignment of values $\boldsymbol{Z} = \boldsymbol{z}$. It turns out that each of these operations is very easy to perform in one of the two ways of encoding a Gaussian, and not so easy in the other.

Marginalization is trivial to perform in the covariance form. Specifically, the marginal Gaussian distribution over any subset of the variables can simply be read from the mean and covariance matrix. For instance, in example 7.1, we can obtain the marginal Gaussian distribution over X_2 and X_3 by simply considering only the relevant entries in both the mean vector the covariance matrix. More generally, assume that we have a joint normal distribution over $\{\boldsymbol{X}, \boldsymbol{Y}\}$ where $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^m$. Then we can decompose the mean and covariance of this joint distribution as follows:

$$p(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu_X} \\ \boldsymbol{\mu_Y} \end{pmatrix}; \begin{bmatrix} \Sigma_{\boldsymbol{XX}} & \Sigma_{\boldsymbol{XY}} \\ \Sigma_{\boldsymbol{YX}} & \Sigma_{\boldsymbol{YY}} \end{bmatrix}\right) \tag{7.3}$$

where $\boldsymbol{\mu_X} \in \mathbb{R}^n$, $\boldsymbol{\mu_Y} \in \mathbb{R}^m$, $\Sigma_{\boldsymbol{XX}}$ is a matrix of size $n \times n$, $\Sigma_{\boldsymbol{XY}}$ is a matrix of size $n \times m$, $\Sigma_{\boldsymbol{YX}} = \Sigma_{\boldsymbol{XT}}^T$ is a matrix of size $m \times n$ and $\Sigma_{\boldsymbol{YY}}$ is a matrix of size $m \times m$.

Lemma 7.1 *Let $\{\boldsymbol{X}, \boldsymbol{Y}\}$ have a joint normal distribution defined in equation (7.3). Then the marginal distribution over $\boldsymbol{Y}$ is a normal distribution $\mathcal{N}(\boldsymbol{\mu_Y}; \Sigma_{\boldsymbol{YY}})$.*

The proof follows directly from the definitions (see exercise 7.1).

On the other hand, conditioning a Gaussian on an observation $\boldsymbol{Z} = \boldsymbol{z}$ is very easy to perform in the information form. We simply assign the values $\boldsymbol{Z} = \boldsymbol{z}$ in equation (7.2). This process turns some of the quadratic terms into linear terms or even constant terms, and some of the linear terms into constant terms. The resulting expression, however, is still in the same form as in equation (7.2), albeit over a smaller subset of variables.

In summary, although the two representations both encode the same information, they have different computational properties. **To marginalize a Gaussian over a subset of the variables, one essentially needs to compute their pairwise covariances, which is precisely generating the distribution in its covariance form. Similarly, to condition a Gaussian on an observation, one essentially needs to invert the covariance matrix to obtain the information form. For small matrices, inverting a matrix may be feasible, but in high-dimensional spaces, matrix inversion may be far too costly.**

7.1.3 Independencies in Gaussians

For multivariate Gaussians, independence is easy to determine directly from the parameters of the distribution.

Theorem 7.1 *Let $\boldsymbol{X} = X_1, ..., X_n$ have a joint normal distribution $\mathcal{N}(\boldsymbol{\mu}; \Sigma)$. Then X_i and X_j are independent if and only if $\Sigma_{i,j} = 0$.*

The proof is left as an exercise (exercise 7.2).

Note that this property does not hold in general. In other words, if $p(X, Y)$ is not Gaussian, then it is possible that $\boldsymbol{Cov}[X; Y] = 0$ while X and Y are still dependent in p. (See exercise 7.2.)

At first glance, it seems that conditional independencies are not quite as apparent as marginal independencies. However, it turns out that the independence structure in the distribution is apparent not in the covariance matrix, but in the information matrix.

Theorem 7.2 *Consider a Gaussian distribution $p(X_1, \ldots, X_n) = \mathcal{N}(\boldsymbol{\mu}; \Sigma)$, and let $J = \Sigma^{-1}$ be the information matrix. Then $J_{i,j} = 0$ if and only if $p \models (X_i \perp X_j \mid \mathcal{X} - \{X_i, X_j\})$.*

The proof is left as an exercise (exercise 7.3).

Example 7.2 *Consider the covariance matrix of example 7.1. Simple algebraic operations allow us to compute its inverse:*

$$J = \begin{pmatrix} 0.3125 & -0.125 & 0 \\ -0.125 & 0.5833 & 0.3333 \\ 0 & 0.3333 & 0.3333 \end{pmatrix}$$

As we can see, the entry in the matrix corresponding to X_1, X_3 is zero, reflecting the fact that they are conditionally independent given X_2. ■

Theorem 7.2 asserts the fact that the information matrix captures independencies between pairs of variables, conditioned on all of the remaining variables in the model. These are precisely the same independencies as the pairwise Markov independencies of definition 4.10. Thus, we can view the information matrix J for a Gaussian density p as precisely capturing the pairwise Markov independencies in a Markov network representing p. Because a Gaussian density is a positive distribution, we can now use theorem 4.5 to construct a Markov network that is a unique minimal I-map for p. As stated in this theorem, the construction simply introduces an edge between X_i and X_j whenever $(X_i \perp X_j \mid \mathcal{X} - \{X_i, X_j\})$ does not hold in p. But this latter condition holds precisely when $J_{i,j} \neq 0$. **Thus, we can view the information matrix as directly defining a minimal I-map Markov network for p, whereby nonzero entries correspond to edges in the network.**

7.2 Gaussian Bayesian Networks

We now show how we can define a continuous joint distribution using a Bayesian network. This representation is based on the *linear Gaussian model*, which we defined in definition 5.14. Although this model can be used as a CPD within any network, it turns out that continuous networks defined solely in terms of linear Gaussian CPDs are of particular interest:

Definition 7.1 Gaussian Bayesian network

We define a Gaussian Bayesian network *to be a Bayesian network all of whose variables are continuous, and where all of the CPDs are linear Gaussians.* ■

An important and surprising result is that linear Gaussian Bayesian networks are an alternative representation for the class of multivariate Gaussian distributions. This result has two parts. The first is that a linear Gaussian network always defines a joint multivariate Gaussian distribution.

Theorem 7.3 *Let Y be a linear Gaussian of its parents $X_1, \ldots, X_k$:*

$$p(Y \mid \boldsymbol{x}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}; \sigma^2\right).$$

Assume that $X_1, \ldots, X_k$ are jointly Gaussian with distribution $\mathcal{N}(\boldsymbol{\mu}; \Sigma)$. Then:

- *The distribution of Y is a normal distribution $p(Y) = \mathcal{N}\left(\mu_Y; \sigma_Y^2\right)$ where:*

$$\begin{aligned} \mu_Y &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu} \\ \sigma_Y^2 &= \sigma^2 + \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}. \end{aligned}$$

- *The joint distribution over $\{\boldsymbol{X}, Y\}$ is a normal distribution where:*

$$\mathbf{Cov}[X_i; Y] = \sum_{j=1}^{k} \beta_j \Sigma_{i,j}.$$

From this theorem, it follows easily by induction that if $\mathcal{B}$ is a linear Gaussian Bayesian network, then it defines a joint distribution that is jointly Gaussian.

Example 7.3

Consider the linear Gaussian network $X_1 \to X_2 \to X_3$, where

$$\begin{aligned} p(X_1) &= \mathcal{N}(1; 4) \\ p(X_2 \mid X_1) &= \mathcal{N}(0.5X_1 - 3.5; 4) \\ p(X_3 \mid X_2) &= \mathcal{N}(-X_2 + 1; 3). \end{aligned}$$

Using the equations in theorem 7.3, we can compute the joint Gaussian distribution $p(X_1, X_2, X_3)$. For the mean, we have that:

$$\begin{aligned} \mu_2 &= 0.5\mu_1 - 3.5 = 0.5 \cdot 1 - 3.5 = -3 \\ \mu_3 &= (-1)\mu_2 + 1 = (-1) \cdot (-3) + 1 = 4. \end{aligned}$$

The variance of X_2 and X_3 can be computed as:

$$\begin{aligned} \Sigma_{22} &= 4 + (1/2)^2 \cdot 4 = 5 \\ \Sigma_{33} &= 3 + (-1)^2 \cdot 5 = 8. \end{aligned}$$

We see that the variance of the variable is a sum of two terms: the variance arising from its own Gaussian noise parameter, and the variance of its parent variables weighted by the strength of the dependence. Finally, we can compute the covariances as follows:

$$\begin{aligned} \Sigma_{12} &= (1/2) \cdot 4 = 2 \\ \Sigma_{23} &= (-1) \cdot \Sigma_{22} = -5 \\ \Sigma_{13} &= (-1) \cdot \Sigma_{12} = -2. \end{aligned}$$

The third equation shows that, although X_3 does not depend directly on X_1, they have a nonzero covariance. Intuitively, this is clear: X_3 depends on X_2, which depends on X_1; hence, we expect X_1 and X_3 to be correlated, a fact that is reflected in their covariance. As we can see, the covariance between X_1 and X_3 is the covariance between X_1 and X_2, weighted by the strength of the dependence of X_3 on X_2.

In general, putting these results together, we can see that the mean and covariance matrix for $p(X_1, X_2, X_3)$ is precisely our covariance matrix of example 7.1. ■

The converse to this theorem also holds: the result of conditioning is a normal distribution where there is a linear dependency on the conditioning variables. The expressions for converting a multivariate Gaussian to a linear Gaussian network appear complex, but they are based on simple algebra. They can be derived by taking the linear equations specified in theorem 7.3, and reformulating them as defining the parameters β_i in terms of the means and covariance matrix entries.

Theorem 7.4

Let $\{\boldsymbol{X}, Y\}$ have a joint normal distribution defined in equation (7.3). Then the conditional density

$$p(Y \mid \boldsymbol{X}) = \mathcal{N}\left(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{X}; \sigma^2\right),$$

is such that:

$$\begin{aligned} \beta_0 &= \mu_Y - \Sigma_{Y\boldsymbol{X}} \Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1} \boldsymbol{\mu}_{\boldsymbol{X}} \\ \boldsymbol{\beta} &= \Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1} \Sigma_{Y\boldsymbol{X}} \\ \sigma^2 &= \Sigma_{YY} - \Sigma_{Y\boldsymbol{X}} \Sigma_{\boldsymbol{X}\boldsymbol{X}}^{-1} \Sigma_{\boldsymbol{X}Y}. \end{aligned}$$

This result allows us to take a joint Gaussian distribution and produce a Bayesian network, using an identical process to our construction of a minimal I-map in section 3.4.1.

Theorem 7.5

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$, and let p be a joint Gaussian distribution over $\mathcal{X}$. Given any ordering $X_1, \ldots, X_n$ over $\mathcal{X}$, we can construct a Bayesian network graph $\mathcal{G}$ and a Bayesian network $\mathcal{B}$ over $\mathcal{G}$ such that:

1. $\mathrm{Pa}_{X_i}^{\mathcal{G}} \subseteq \{X_1, \ldots, X_{i-1}\}$;
2. *the CPD of X_i in $\mathcal{B}$ is a linear Gaussian of its parents;*
3. *$\mathcal{G}$ is a minimal I-map for p.*

The proof is left as an exercise (exercise 7.4). As for the case of discrete networks, the minimal I-map is not unique: different choices of orderings over the variables will lead to different network structures. For example, the distribution in figure 7.1b can be represented either as the network where $X \rightarrow Y$ or as the network where $Y \rightarrow X$.

This equivalence between Gaussian distributions and linear Gaussian networks has important practical ramifications. On one hand, we can conclude that, for linear Gaussian networks, the joint distribution has a compact representation (one that is quadratic in the number of variables). Furthermore, the transformations from the network to the joint and back have a fairly simple and efficiently computable closed form. Thus, we can easily convert one representation to another, using whichever is more convenient for the current task. Conversely, **while the two representations are equivalent in their expressive power, there is not a one-to-one correspondence between their parameterizations. In particular, although in the worst case, the linear Gaussian representation and the Gaussian representation have the same number of parameters (exercise 7.6), there are cases where one representation can be significantly more compact than the other.**

Example 7.4

Consider a linear Gaussian network structured as a chain:

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n.$$

Assuming the network parameterization is not degenerate (that is, the network is a minimal I-map of its distribution), we have that each pair of variables X_i, X_j are correlated. In this case, as shown in theorem 7.1, the covariance matrix would be dense — none of the entries would be zero. Thus, the representation of the covariance matrix would require a quadratic number of parameters. In the information matrix, however, for all X_i, X_j that are not neighbors in the chain, we have that X_i and X_j are conditionally independent given the rest of the variables in the network; hence, by theorem 7.2, $J_{i,j} = 0$. Thus, the information matrix has most of the entries being zero; the only nonzero entries are on the tridiagonal (the entries i, j for $j = i - 1, i, i + 1$). ■

However, not all structure in a linear Gaussian network is represented in the information matrix.

Example 7.5

In a v-structure $X \rightarrow Z \leftarrow Y$, we have that X and Y are marginally independent, but not conditionally independent given Z. Thus, according to theorem 7.2, the X, Y entry in the information matrix would not be 0. Conversely, because the variables are marginally independent, the X, Y entry in the covariance entry would be zero.

Complicating the example somewhat, assume that X and Y also have a joint parent W; that is, the network is structured as a diamond. In this case, X and Y are still not independent given the remaining network variables Z, W, and hence the X, Y entry in the information matrix is nonzero. Conversely, they are also not marginally independent, and thus the X, Y entry in the covariance matrix is also nonzero. ■

These examples simply recapitulate, in the context of Gaussian networks, the fundamental difference in expressive power between Bayesian networks and Markov networks.

7.3 Gaussian Markov Random Fields

We now turn to the representation of multivariate Gaussian distributions via an undirected graphical model. We first show how a Gaussian distribution can be viewed as an MRF. This formulation is derived almost immediately from the information form of the Gaussian. Consider again equation (7.2). We can break up the expression in the exponent into two types of terms: those that involve single variables X_i and those that involve pairs of variables X_i, X_j. The terms that involve only the variable X_i are:

$$-\frac{1}{2} J_{i,i} x_i^2 + h_i x_i, \tag{7.4}$$

where we recall that the potential vector $\boldsymbol{h} = J\boldsymbol{\mu}$. The terms that involve the pair X_i, X_j are:

$$-\frac{1}{2}[J_{i,j} x_i x_j + J_{j,i} x_j x_i] = -J_{i,j} x_i x_j, \tag{7.5}$$

due to the symmetry of the information matrix. Thus, the information form immediately induces a pairwise Markov network, whose node potentials are derived from the potential vector and the

diagonal elements of the information matrix, and whose edge potentials are derived from the off-diagonal entries of the information matrix. We also note that, when $J_{i,j} = 0$, there is no edge between X_i and X_j in the model, corresponding directly to the independence assumption of the Markov network.

Gaussian MRF

Thus, any Gaussian distribution can be represented as a pairwise Markov network with quadratic node and edge potentials. This Markov network is generally called a *Gaussian Markov random field (GMRF)*. Conversely, consider any pairwise Markov network with quadratic node and edge potentials. Ignoring constant factors, which can be assimilated into the partition function, we can write the node and edge energy functions (log-potentials) as:

$$\begin{aligned} \epsilon_i(x_i) &= d_0^i + d_1^i x_i + d_2^i x_i^2 \\ \epsilon_{i,j}(x_i, x_j) &= a_{00}^{i,j} + a_{01}^{i,j} x_i + a_{10}^{i,j} x_j + a_{11}^{i,j} x_i x_j + a_{02}^{i,j} x_i^2 + a_{20}^{i,j} x_j^2, \end{aligned} \tag{7.6}$$

where we used the log-linear notation of section 4.4.1.2. By aggregating like terms, we can reformulate any such set of potentials in the log-quadratic form:

$$p'(\boldsymbol{x}) = \exp(-\frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x} + \boldsymbol{h}^T \boldsymbol{x}), \tag{7.7}$$

where we can assume without loss of generality that J is symmetric. This Markov network defines a valid Gaussian density if and only if J is a positive definite matrix. If so, then J is a legal information matrix, and we can take $\boldsymbol{h}$ to be a potential vector, resulting in a distribution in the form of equation (7.2).

However, unlike the case of Gaussian Bayesian networks, it is not the case that every set of quadratic node and edge potentials induces a legal Gaussian distribution. Indeed, the decomposition of equation (7.4) and equation (7.5) can be performed for any quadratic form, including one not corresponding to a positive definite matrix. For such matrices, the resulting function $\exp(\boldsymbol{x}^T A \boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x})$ will have an infinite integral, and cannot be normalized to produce a valid density. Unfortunately, **other than generating the entire information matrix and testing whether it is positive definite, there is no simple way to check whether the MRF is valid. In particular, there is no local test that can be applied to the network parameters that precisely characterizes valid Gaussian densities.** However, there *are* simple tests that are sufficient to induce a valid density. While these conditions are not necessary, they appear to cover many of the cases that occur in practice.

We first provide one very simple test that can be verified by direct examination of the information matrix.

Definition 7.2

diagonally dominant

A quadratic MRF parameterized by J is said to be diagonally dominant *if, for all i,*

$$\sum_{j \neq i} |J_{i,j}| < J_{i,i}.$$

■

For example, the information matrix in example 7.2 is diagonally dominant; for instance, for $i = 2$ we have:

$$|-0.125| + 0.3333 < 0.5833.$$

One can now show the following result:

Proposition 7.1 *Let $p'(\boldsymbol{x}) = \exp(-\frac{1}{2}\boldsymbol{x}^T J \boldsymbol{x} + \boldsymbol{h}^T \boldsymbol{x})$ be a quadratic pairwise MRF. If J is diagonally dominant, then p' defines a valid Gaussian MRF.*

The proof is straightforward algebra and is left as an exercise (exercise 7.8).

The following condition is less easily verified, since it cannot be tested by simple examination of the information matrix. Rather, it checks whether the distribution can be written as a quadratic pairwise MRF whose node and edge potentials satisfy certain conditions. Specifically, recall that a Gaussian MRF consists of a set of node potentials, which are log-quadratic forms in x_i, and a set of edge potentials, which are log-quadratic forms in x_i, x_j. We can state a condition in terms of the coefficients for the nonlinear components of this parameterization:

Definition 7.3 *A quadratic MRF parameterized as in equation (7.6) is said to be* pairwise normalizable *if:*

pairwise normalizable

- *for all i, $d_2^i > 0$;*
- *for all i, j, the 2×2 matrix*

$$\begin{pmatrix} a_{02}^{i,j} & a_{11}^{i,j}/2 \\ a_{11}^{i,j}/2 & a_{20}^{i,j} \end{pmatrix}$$

is positive semidefinite. ■

Intuitively, this definition states that each edge potential, considered in isolation, is normalizable (hence the name "pairwise-normalizable").

We can show the following result:

Proposition 7.2 *Let $p'(\boldsymbol{x})$ be a quadratic pairwise MRF, parameterized as in equation (7.6). If p' is pairwise normalizable, then it defines a valid Gaussian distribution.*

Once again, the proof follows from standard algebraic manipulations, and is left as an exercise (exercise 7.9).

We note that, like the preceding conditions, this condition is sufficient but not necessary:

Example 7.6 *Consider the following information matrix:*

$$\begin{pmatrix} 1 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 1 \end{pmatrix}$$

It is not difficult to show that this information matrix is positive definite, and hence defines a legal Gaussian distribution. However, it turns out that it is not possible to decompose this matrix into a set of three edge potentials, each of which is positive definite. ■

Unfortunately, evaluating whether pairwise normalizability holds for a given MRF is not always trivial, since it can be the case that one parameterization is not pairwise normalizable, yet a different parameterization that induces precisely the same density function is pairwise normalizable.

Example 7.7

Consider the information matrix of example 7.2, with a mean vector $\mathbf{0}$*. We can define this distribution using an MRF by simply choosing the node potential for* X_i *to be* $J_{i,i}x_i^2$ *and the edge potential for* X_i, X_j *to be* $2J_{i,j}x_ix_j$*. Clearly, the* X_1, X_2 *edge does* not *define a normalizable density over* X_1, X_2*, and hence this MRF is not pairwise normalizable. However, as we discussed in the context of discrete MRFs, the MRF parameterization is nonunique, and the same density can be induced using a continuum of different parameterizations. In this case, one alternative parameterization of the same density is to define all node potentials as* $\epsilon_i(x_i) = 0.05x_i^2$*, and the edge potentials to be* $\epsilon_{1,2}(x_1, x_2) = 0.2625x_1^2 + 0.0033x_2^2 - 0.25x_1x_2$*, and* $\epsilon_{2,3}(x_2, x_3) = 0.53x_2^2 + 0.2833x_3^2 + 0.6666x_2x_3$*. Straightforward arithmetic shows that this set of potentials induces the information matrix of example 7.2. Moreover, we can show that this formulation is pairwise normalizable: The three node potentials are all positive, and the two edge potentials are both positive definite. (This latter fact can be shown either directly or as a consequence of the fact that each of the edge potentials is diagonally dominant, and hence also positive definite.)* ■

This example illustrates that the pairwise normalizability condition is easily checked for a specific MRF parameterization. However, if our aim is to encode a particular Gaussian density as an MRF, we may have to actively search for a decomposition that satisfies the relevant constraints. If the information matrix is small enough to manipulate directly, this process is not difficult, but if the information matrix is large, finding an appropriate parameterization may incur a nontrivial computational cost.

7.4 Summary

This chapter focused on the representation and independence properties of Gaussian networks.

We showed an equivalence of expressive power between three representational classes: multivariate Gaussians, linear Gaussian Bayesian networks, and Gaussian MRFs. In particular, any distribution that can be represented in one of those forms can also be represented in another. We provided closed-form formulas that allow us convert between the multivariate Gaussian representation and the linear Gaussian Bayesian network. The conversion for Markov networks is simpler in some sense, inasmuch as there is a direct mapping between the entries in the information (inverse covariance) matrix of the Gaussian and the quadratic forms that parameterize the edge potentials in the Markov network. However, unlike the case of Bayesian networks, here we must take care, since not every quadratic parameterization of a pairwise Markov network induces a legal Gaussian distribution: The quadratic form that arises when we combine all the pairwise potentials may not have a finite integral, and therefore may not be normalizable. In general, there is no local way of determining whether a pairwise MRF with quadratic potentials is normalizable; however, we provided some easily checkable sufficient conditions that are often sufficient in practice.

The equivalence between the different representations is analogous to the equivalence of Bayesian networks, Markov networks, and discrete distributions: any discrete distribution can be encoded both as a Bayesian network and as a Markov network, and vice versa. However, as in the discrete case, this equivalence does *not* imply equivalence of expressive power with respect to independence assumptions. In particular, **the expressive power of the directed**

and undirected representations in terms of independence assumptions is exactly the same as in the discrete case: Directed models can encode the independencies associated with immoralities, whereas undirected models cannot; conversely, undirected models can encode a symmetric diamond, whereas directed models cannot. As we saw, the undirected models have a particularly elegant connection to the natural representation of the Gaussian distribution in terms of the information matrix; in particular, zeros in the information matrix for p correspond precisely to missing edges in the minimal I-map Markov network for p.

Finally, we note that the class of Gaussian distributions is highly restrictive, making strong assumptions that often do not hold in practice. Nevertheless, it is a very useful class, due to its compact representation and computational tractability (see section 14.2). Thus, in many cases, we may be willing to make the assumption that a distribution is Gaussian even when that is only a rough approximation. This approximation may happen a priori, in encoding a distribution as a Gaussian even when it is not. Or, in many cases, we perform the approximation as part of our inference process, representing intermediate results as a Gaussian, in order to keep the computation tractable. Indeed, as we will see, the Gaussian representation is ubiquitous in methods that perform inference in a broad range of continuous models.

7.5 Relevant Literature

The equivalence between the multivariate and linear Gaussian representations was first derived by Wermuth (1980), who also provided the one-to-one transformations between them. The introduction of linear Gaussian dependencies into a Bayesian network framework was first proposed by Shachter and Kenley (1989), in the context of influence diagrams.

Speed and Kiiveri (1986) were the first to make the connection between the structure of the information matrix and the independence assumptions in the distribution. Building on earlier results for discrete Markov networks, they also made the connection to the undirected graph as a representation. Lauritzen (1996, Chapter 5) and Malioutov et al. (2006) give a good overview of the properties of Gaussian MRFs.

7.6 Exercises

Exercise 7.1

Prove lemma 7.1. Note that you need to show both that the marginal distribution is a Gaussian, and that it is parameterized as $\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{Y}}; \Sigma_{\boldsymbol{YY}})$.

Exercise 7.2

a. Show that, for any joint density function $p(X, Y)$, if we have $(X \perp Y)$ in p, then $\boldsymbol{Cov}[X; Y] = 0$.

b. Show that, if $p(X, Y)$ is Gaussian, and $\boldsymbol{Cov}[X; Y] = 0$, then $(X \perp Y)$ holds in p.

c. Show a counterexample to 2 for non-Gaussian distributions. More precisely, show a construction of a joint density function $p(X, Y)$ such that $\boldsymbol{Cov}[X; Y] = 0$, while $(X \perp Y)$ does not hold in p.

Exercise 7.3

Prove theorem 7.2.

Exercise 7.4

Prove theorem 7.5.

Exercise 7.5

Consider a Kalman filter whose transition model is defined in terms of a pair of matrices A, Q, and whose observation model is defined in terms of a pair of matrices H, R, as specified in equation (6.3) and equation (6.4). Describe how we can extract a 2-TBN structure representing the conditional independencies in this process from these matrices. (Hint: Use theorem 7.2.)

Exercise 7.6

In this question, we compare the number of independent parameters in a multivariate Gaussian distribution and in a linear Gaussian Bayesian network.

a. Show that the number of independent parameters in Gaussian distribution over $X_1, \ldots, X_n$ is the same as the number of independent parameters in a fully connected linear Gaussian Bayesian network over $X_1, \ldots, X_n$.

b. In example 7.4, we showed that the number of parameters in a linear Gaussian network can be substantially smaller than in its multivariate Gaussian representation. Show that the converse phenomenon can also happen. In particular, show an example of a distribution where the multivariate Gaussian representation requires a linear number of nonzero entries in the covariance matrix, while a corresponding linear Gaussian network (one that is a minimal I-map) requires a quadratic number of nonzero parameters. (Hint: The minimal I-map does not have to be the optimal one.)

Exercise 7.7

conditional covariance

Let p be a joint Gaussian density over $\mathcal{X}$ with mean vector $\boldsymbol{\mu}$ and information matrix J. Let $X_i \in \mathcal{X}$, and $\boldsymbol{Z} \subset \mathcal{X} - \{X_i\}$. We define the *conditional covariance* of X_i, X_j given $\boldsymbol{Z}$ as:

$$\mathbf{Cov}_p[X_i; X_j \mid \boldsymbol{Z}] = \mathbf{E}_p[(X_i - \mu_i)(X_j - \mu_j) \mid \boldsymbol{Z}] = \mathbf{E}_{\boldsymbol{z} \sim p(\boldsymbol{Z})}\left[\mathbf{E}_{p(X_i, X_j \mid \boldsymbol{z})}[(x_i - \mu_i)(x_j - \mu_j)]\right].$$

partial correlation coefficient

The conditional variance of X_i is defined by setting $j = i$. We now define the *partial correlation coefficient*

$$\rho_{i,j} = \frac{\mathbf{Cov}_p[X_i; X_j \mid \mathcal{X} - \{X_i, X_j\}]}{\sqrt{\mathbf{Var}_p[X_i \mid \mathcal{X} - \{X_i, X_j\}]\mathbf{Var}_p[X_j \mid \mathcal{X} - \{X_i, X_j\}]}}.$$

Show that

$$\rho_{i,j} = -\frac{J_{i,j}}{\sqrt{J_{i,i}J_{j,j}}}.$$

Exercise 7.8

Prove proposition 7.1.

Exercise 7.9

Prove proposition 7.2.