5 Gaussian Graphical Models and Factor Graphs

Our development of graphical models thus far has been conceptually quite general, and we have seen some examples that involve random variables drawn from finite alphabets. However, in many applications, including some we mentioned in our introduction to the subject, the phenomena of interest are more naturally modeled using random variables over infinite alphabets, including continuous-valued variables.

While there are no major conceptual differences between graphical models for discrete- and continuous-valued distributions, there can be mathematical subtleties in their formal analysis.

In this installment of the notes, we begin to consider the continuous case more explicitly, to understand some of the issues involved and broaden our framework of thinking. Our focus will be on a particular class of continuous-valued random variables: multivariate *Gaussian* distributions. There are several reasons for this.

First, Gaussian distributions are natural models. Indeed, many phenomena are well-modeled using such distributions. The Gaussian is an example of a *stable* family of distributions, meaning that if we add two independent Gaussian random variables, the result is Gaussian. Moreover, if we add many independent random variables, the result is often well-approximated as Gaussian, a result formalized by the celebrated *Central Limit Theorem*. From these perspectives, the Gaussian family can be viewed as a kind of "attractor" to which phenomena converge under mixing.

From another perspective, Gaussian distributions are often good adversarial models in various settings, corresponding to worst-case models of uncertainty, so these models frequently lead to robust inference algorithms. This further makes them appealing for practical applications.

Finally, Gaussian distributions are mathematically very well-behaved objects, and avoid the kind of subtleties in analysis that complicate more general developments of continuous-valued random variables. In fact, as we'll see, the framework for analyzing inferences about Gaussian phenomena is ultimately based on the tools of relatively simple finite-dimensional *linear algebra* and this has important implications for inference with such models.

We begin by recalling some definitions associated to Gaussian random variables.

5.1 Gaussian Random Variables

The following is the familiar definition of a Gaussian (or normal) random variable.

Definition 1. *A random variable* x *is Gaussian if its probability density function (pdf)* $p_{\mathsf{x}}(\cdot)$ *can be expressed in the form*

$$p_{\mathsf{x}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2}, \quad x \in \mathbb{R},$$

for some parameters $\mu \in \mathbb{R}$ *and* $\sigma \in \mathbb{R}^+$.[1]

As notation, we use $x \sim \mathtt{N}(\mu, \sigma^2)$ to indicate that x is Gaussian distributed with the associated parameters, which we also sometimes express using the related notation $p_x(x) = \mathtt{N}(x; \mu, \sigma^2)$.

As additional terminology, $\mathtt{N}(0, 1)$ is referred to as the *standard* or *unit* Gaussian distribution.

5.2 Jointly Gaussian Random Variables

There are several equivalent ways to define a *jointly* Gaussian collection of random variables $\mathbf{x} = (x_1, \ldots, x_N)$, examples of which are as follows:

1. A collection of random variables $\mathbf{x}$ is jointly Gaussian if it can be expressed as a linear combination of i.i.d. scalar Gaussian variables, i.e., if there exists some matrix $\mathbf{A}$, constant vector $\mathbf{b}$ and random vector $\mathbf{u}$ with i.i.d. $\mathtt{N}(0, 1)$ entries such that $\mathbf{x} = \mathbf{A}\mathbf{u} + \mathbf{b}$.

2. A collection of random variables $\mathbf{x}$ is jointly Gaussian if every linear combinations of elements of $\mathbf{x}$ is a scalar Gaussian random variable, i.e., if $y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ is a Gaussian random variable for every choice of constant vector $\mathbf{a}$.

3. A collection of random variables $\mathbf{x}$ is jointly Gaussian if their joint distribution can be expressed in the form[2]

$$p_{x_1,\ldots,x_N}(x_1, \ldots, x_N) \triangleq p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Lambda}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Lambda}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \tag{1}$$

   where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ is the mean vector and $\mathbf{\Lambda} = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\right]$ is the covariance matrix.[3] This is referred to as the corresponding *multivariate* Gaussian (or normal) distribution, and (1) is referred to as the *covariance form* of the Gaussian distribution.

4. A collection of random variables $\mathbf{x}$ is jointly Gaussian if their joint distribution can be expressed in the form

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{Z} \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{J}\mathbf{x} + \mathbf{h}^{\mathrm{T}}\mathbf{x}\right\} \tag{2a}$$

[1] We include the limiting distribution corresponding to $\sigma \to 0$, i.e., x is trivially Gaussian if it is almost surely a constant: $\mathbb{P}(x = \mu) = 1$ for some μ.

[2] As a reminder, when $\mathbf{A}$ is a matrix, $|\mathbf{A}|$ denotes its determinant.

[3] Note that covariance matrices are symmetric matrices, a property we will frequently use. Actually, they are even more constrained: they are positive semidefinite matrices, although this will be less important to our development.

with

$$Z = \frac{(2\pi)^{N/2}}{|\mathbf{J}|^{1/2}} \exp\left\{\frac{1}{2}\mathbf{h}^{\mathrm{T}}\mathbf{J}^{-1}\mathbf{h}\right\}, \tag{2b}$$

where $\mathbf{h}$ is referred to as the *potential vector* and $\mathbf{J}$ is referred to as the *information (or precision) matrix*.[4] Eq. (2) is referred to as the *information form* of the Gaussian distribution.

Several remarks are worth making.

First, as convenient terminology, we say $\mathbf{x}$ is a *Gaussian random vector* when its constituent variables are jointly Gaussian.

Second, our development will emphasize Characterizations 3 and 4. Based on these characterizations, we (interchangeably) adopt the following covariance form notation $\mathbf{x} \sim \mathtt{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ and information form notation $\mathbf{x} \sim \mathtt{N}^{-1}(\mathbf{h}, \mathbf{J})$ for a Gaussian random vector, which makes their respective parameterizations explicit.

Third, obtaining form (2) from (1) is straightforward. In particular, it suffices to expand the quadratic form in (2), exploit the symmetry of $\boldsymbol{\Lambda}$ (and thus $\boldsymbol{\Lambda}^{-1}$), ignore constants of proportionality, and match terms, i.e.,

$$p_{\mathbf{x}}(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \propto \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\underbrace{\boldsymbol{\Lambda}^{-1}}_{\triangleq \mathbf{J}}\mathbf{x} + \underbrace{\mu^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}}_{\triangleq \mathbf{h}^{\mathrm{T}}}\mathbf{x}\right\}, \tag{3}$$

from which we see that the parameterizations are related according to

$$\mathbf{J} = \boldsymbol{\Lambda}^{-1} \qquad \text{and} \qquad \mathbf{h} = \mathbf{J}\boldsymbol{\mu}. \tag{4}$$

Fourth, note that the covariance form is inadequate for expressing the joint distribution in cases where the covariance matrix is singular, which occurs when one of the variables can be expressed as a deterministic linear combination of the others. In such cases, the information form can still be used though.[5]

Fifth, it is worth emphasizing that joint Gaussianity imposes strong constraints on the interactions among the variables. In particular, random variables that are individually (or marginally) Gaussian are *not*, in general, jointly Gaussian, as our characterizations each make clear.

Finally, a useful consequence of our definitions and terminology is that there are natural and consistent extensions. For example—and of particular interest in our development—a collection of random *vectors* $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is defined to be jointly Gaussian if the combined set of all their constituent elements taken together are

[4]Note that $\mathbf{J}$ is also a positive semidefinite symmetric matrix since it is the inverse of one. The symmetry in particular will be useful to our development.

[5]To use the covariance form in such cases one can reduce the collection of variables to a linearly independent set, for which the associated covariance is nonsingular.

jointly Gaussian random variables, i.e., if the "supervector"

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix}$$

is a Gaussian random vector.[6] This supervector form, and the associated block-oriented linear algebra, is frequently useful, as we will see.

5.3 Inference with Jointly Gaussian Variables

Before we can develop efficient inference algorithms for jointly Gaussian random variables, we first need to discuss the properties of the basic inference operations involving such variables. In particular, we now develop marginalization and conditioning among jointly Gaussian random variables.

For our development in this section we consider, without loss of generality, a collection of jointly Gaussian random variables partitioned into two groups, corresponding to jointly Gaussian vectors x_1 and x_2. With $\mathbf{x}$ denoting the supervector

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \tag{5}$$

their parameterization in covariance form is

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right] = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}, \tag{6}$$

while their parameterization in information form is

$$\mathbf{h} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\mu} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix}, \qquad \mathbf{J} = \boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix}. \tag{7}$$

5.3.1 Marginalization

In this section we develop the form and parameters of marginal distributions derived from multivariate Gaussians. We begin with the form.

Claim 1. *All marginalizations of the multivariate Gaussian distributions result in multivariate Gaussian distributions.*

Proof. Without loss of generality, we focus on $p_{\mathbf{x}_1}$. This result is an immediate consequence of Characterization 2. In particular, since $\mathbf{x}$ is Gaussian, $\mathbf{a}^{\mathrm{T}}\mathbf{x}$ is Gaussian for all choices of $\mathbf{a}$. In particular, it holds for all $\mathbf{a}$ of the form

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{0} \end{bmatrix},$$

[6] Note that the vectors involved need not each have the same number of elements.

with $\mathbf{a}_1$ having the dimension of $\mathbf{x}_1$, which in turn implies that $\mathbf{a}^{\mathrm{T}}\mathbf{x} = \mathbf{a}_1^{\mathrm{T}}\mathbf{x}_1$ is Gaussian for all choices of $\mathbf{a}_1$. Hence, $\mathbf{x}_1$ is Gaussian.

Alternatively, we could establish these results directly without resort to Characterization 2. For instance, to establish that $\mathbf{x}_1$ is Gaussian, observe that

$$\begin{aligned} p_{\mathbf{x}_1}(\mathbf{x}_1) &= \int_{\mathbf{x}_2} p_{\mathbf{x}_1,\mathbf{x}_2}(\mathbf{x}_1,\mathbf{x}_2) \\ &\propto \int_{\mathbf{x}_2} \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{J}\mathbf{x} + \mathbf{h}^{\mathrm{T}}\mathbf{x}\right\} \\ &= \exp\left\{-\frac{1}{2}\mathbf{x}_1^{\mathrm{T}}\mathbf{J}_{11}\mathbf{x}_1 + \mathbf{h}_1^{\mathrm{T}}\mathbf{x}_1\right\}\phi(\mathbf{x}_1) \end{aligned} \tag{8}$$

where

$$\phi(\mathbf{x}_1) \triangleq \int_{\mathbf{x}_2} \exp\left\{-\frac{1}{2}\mathbf{x}_2^{\mathrm{T}}\mathbf{J}_{22}\mathbf{x}_2 + (\underbrace{\mathbf{h}_2 - \mathbf{x}_1\mathbf{J}_{12}}_{\triangleq\tilde{\mathbf{h}}_2(\mathbf{x}_1)})^{\mathrm{T}}\mathbf{x}_2\right\} \tag{9}$$

Now note that the $\phi(\mathbf{x}_1)$ in (9) is the normalization constant for a Gaussian distribution over $\mathbf{x}_2$ in information form, with $\mathbf{x}_1$ viewed as constant, so using (2b) we deduce that

$$\phi(\mathbf{x}_1) \propto \exp\left\{-\frac{1}{2}\tilde{\mathbf{h}}_2(\mathbf{x}_1)^{\mathrm{T}}\mathbf{J}_{22}\tilde{\mathbf{h}}_2(\mathbf{x}_1)\right\} \propto \exp\left\{-\frac{1}{2}\mathbf{x}_1^{\mathrm{T}}\mathbf{J}'\mathbf{x}_1 + \mathbf{h}'^{\mathrm{T}}\mathbf{x}_1\right\}. \tag{10}$$

for some $\mathbf{J}'$ and $\mathbf{h}'$ that depend on $\tilde{\mathbf{h}}_2$ and $\mathbf{J}_{22}$. Hence, (8) takes the form of a Gaussian distribution in information form. □

Since marginalization preserves Gaussianity, to compute a given marginal we simply need to compute its associated parameters, in either covariance or information form.

To obtain the parameters of the marginal distribution in covariance form is immediate: we simply read off the relevant entries of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. In particular, since

$$\boldsymbol{\mu}_1 = \mathbb{E}\left[\mathbf{x}_1\right] \qquad \text{and} \qquad \boldsymbol{\Lambda}_{11} = \mathbb{E}\left[(\mathbf{x}_1 - \boldsymbol{\mu}_1)(\mathbf{x}_1 - \boldsymbol{\mu}_1)^{\mathrm{T}}\right],$$

it follows immediately that $\mathbf{x}_1 \sim \mathtt{N}(\boldsymbol{\mu}_1, \boldsymbol{\Lambda}_{11})$.

Expressing the parameters of the marginal distribution in information form as a function of $\mathbf{h}$ and $\mathbf{J}$ is more work. Indeed,

$$\mathbf{x}_1 \sim \mathtt{N}^{-1}(\mathbf{h}', \mathbf{J}'), \tag{11a}$$

where

$$\mathbf{h}' = \mathbf{h}_1 - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{h}_2 \tag{11b}$$

and

$$\mathbf{J}' = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}, \tag{11c}$$

and where the expression for $\mathbf{J}'$ is referred to as a *Schur complement* of the block $\mathbf{J}_{22}$ in the matrix $\mathbf{J}$.[7]

To obtain (11), it suffices to partially invert $\mathbf{J}$. In particular, note that to transform $\mathbf{J}$ into a blockwise lower-triangular matrix we can use one step of (row-oriented) blockwise Gaussian elimination:

$$\begin{bmatrix} \mathbf{I} & -\mathbf{J}_{12}\mathbf{J}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{J}' & \mathbf{0} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix}, \tag{12}$$

with $\mathbf{J}' \triangleq \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}$.

To obtain (11c), right-multiplying both sides of (12) by $\mathbf{\Lambda}$ and using that $\mathbf{J}\mathbf{\Lambda} = \mathbf{I}$ gives

$$\begin{bmatrix} \mathbf{I} & -\mathbf{J}_{12}\mathbf{J}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{J}' & \mathbf{0} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{J}'\mathbf{\Lambda}_{11} & \mathbf{J}'\mathbf{\Lambda}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{13}$$

from which we see that matching upper left entries yields

$$\mathbf{J}' = \mathbf{\Lambda}_{11}^{-1}, \tag{14}$$

as required.

Next, to find obtain (11b) requires evaluating $\mathbf{h}' = \mathbf{J}'\boldsymbol{\mu}_1$. We first write $\mathbf{J}\boldsymbol{\mu} = \mathbf{h}$ in its expanded from

$$\begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix}, \tag{15}$$

then apply the Gaussian elimination step transform given in (12) to both sides of (15) to obtain

$$\begin{bmatrix} \mathbf{J}' & \mathbf{0} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{J}_{12}\mathbf{J}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1 - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{h}_2 \\ \mathbf{h}_2 \end{bmatrix}. \tag{16}$$

Hence, since the left-hand side is

$$\begin{bmatrix} \mathbf{J}'\boldsymbol{\mu}_1 \\ \mathbf{J}_{12}\boldsymbol{\mu}_1 + \mathbf{J}_{22}\boldsymbol{\mu}_2 \end{bmatrix},$$

we obtain, after equating to the right-hand side of (16),

$$\mathbf{h}' \triangleq \mathbf{J}'\boldsymbol{\mu}_1 = \mathbf{h}_1 - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{h}_2,$$

as required.

[7]Sometimes, it is convenient to use the notation $\mathbf{J}/\mathbf{J}_{22}$ for this Schur complement. Analogously, $\mathbf{J}_{22} - \mathbf{J}_{21}\mathbf{J}_{11}^{-1}\mathbf{J}_{12}$ is the Schur complement of block $\mathbf{J}_{11}$ in $\mathbf{J}$, and is denoted $\mathbf{J}/\mathbf{J}_{11}$.

5.3.2 Conditioning

In this section we develop the form and parameters of conditional distributions derived from multivariate Gaussians. We begin with the form.

Claim 2. *The distribution for any subset of a collection of jointly Gaussian random variables conditioned on any other subset of the same collection is multivariate Gaussian.*

Proof. Without loss of generality, we let $\mathbf{x}_1$ denote the first subset, and $\mathbf{x}_2$ denote the second subset. (If there are additional variables in the collection, we can first marginalize them out, which we know results in $\mathbf{x}$ as defined in (5) still being Gaussian.) To see that $\mathbf{x}_1$ is Gaussian conditioned on $\mathbf{x}_2 = \mathbf{x}_2$, note that with $\mathbf{x}_2$ viewed as a constant we have

$$\begin{aligned} p_{\mathbf{x}_1|\mathbf{x}_2}(\mathbf{x}_1|\mathbf{x}_2) &\propto p_{\mathbf{x}_1,\mathbf{x}_2}(\mathbf{x}_1,\mathbf{x}_2) \\ &\propto \exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{x}_1^{\mathrm{T}} & \mathbf{x}_2^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{J}_{11} & \mathbf{J}_{12}\\ \mathbf{J}_{21} & \mathbf{J}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix} + \begin{bmatrix}\mathbf{h}_1^{\mathrm{T}} & \mathbf{h}_2^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}\right\} \qquad (17) \\ &\propto \exp\left\{-\frac{1}{2}\left(\mathbf{x}_1^{\mathrm{T}}\mathbf{J}_{11}\mathbf{x}_1 + 2\mathbf{x}_2^{\mathrm{T}}\mathbf{J}_{21}\mathbf{x}_1\right) + \mathbf{h}_1^{\mathrm{T}}\mathbf{x}_1\right\} \\ &= \exp\left\{-\frac{1}{2}\mathbf{x}_1^{\mathrm{T}}\mathbf{J}_{11}\mathbf{x}_1 + (\underbrace{\mathbf{h}_1 - \mathbf{J}_{12}\mathbf{x}_2}_{\triangleq \mathbf{h}_1'})^{\mathrm{T}}\mathbf{x}_1\right\}. \qquad (18) \end{aligned}$$

Since (18) is the information form of a Gaussian random vector, we conclude that the conditional distribution is Gaussian. □

Since the conditional distribution is Gaussian, to fully specify it, we need only its covariance or information form parameters, which we now develop.

In this case, the information form parameters are easiest to obtain. In particular, via (18) in our proof above, we can read off its potential vector and information matrix; specifically, we obtain

$$\mathbf{x}_1|\mathbf{x}_2 \sim \mathtt{N}^{-1}(\mathbf{h}_1', \mathbf{J}_{11}), \qquad (19)$$

where

$$\mathbf{h}_1' = \mathbf{h}_1 - \mathbf{J}_{12}\mathbf{x}_2. \qquad (20)$$

To obtain the covariance form parameters as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ is more work. Indeed, as we show in a moment,

$$\mathbf{x}_1|\mathbf{x}_2 \sim \mathtt{N}(\boldsymbol{\mu}', \boldsymbol{\Lambda}'), \qquad (21a)$$

where

$$\boldsymbol{\mu}' = \boldsymbol{\mu}_1 + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \qquad (21b)$$

and

$$\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}, \tag{21c}$$

where we now recognize the expression for $\boldsymbol{\Lambda}'$ is a Schur complement of the matrix $\boldsymbol{\Lambda}$.

To obtain (21), it suffices to partially invert $\boldsymbol{\Lambda}$ in this case. In particular, note that to transform $\boldsymbol{\Lambda}$ into a blockwise lower-triangular matrix we can use one step of (row-oriented) blockwise Gaussian elimination:

$$\begin{bmatrix} \mathbf{I} & -\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}' & \mathbf{0} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}, \tag{22}$$

with $\boldsymbol{\Lambda}' \triangleq \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}$.

To obtain (21c), right-multiplying both sides of (22) by $\mathbf{J}$ and using that $\mathbf{J}\boldsymbol{\Lambda} = \mathbf{I}$ gives

$$\begin{bmatrix} \mathbf{I} & -\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}' & \mathbf{0} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}'\mathbf{J}_{11} & \boldsymbol{\Lambda}'\mathbf{J}_{12} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \tag{23}$$

from which we see that matching upper left entries yields

$$\boldsymbol{\Lambda}' = \mathbf{J}_{11}^{-1}, \tag{24}$$

as required.

Next, to obtain (21b) requires evaluating $\boldsymbol{\mu}' = \boldsymbol{\Lambda}'\mathbf{h}_1'$. We first write $\boldsymbol{\Lambda}\mathbf{h} = \boldsymbol{\mu}$ in its expanded from

$$\begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \tag{25}$$

then apply the Gaussian elimination step transform given in (22) to both sides of (25) to obtain

$$\begin{bmatrix} \boldsymbol{\Lambda}' & \mathbf{0} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_2 \end{bmatrix}. \tag{26}$$

Hence, since the left-hand side is

$$\begin{bmatrix} \boldsymbol{\Lambda}'\mathbf{h}_1 \\ \boldsymbol{\Lambda}_{12}\mathbf{h}_1 + \boldsymbol{\Lambda}_{22}\mathbf{h}_2 \end{bmatrix},$$

we obtain, after equating to the right-hand side of (26),

$$\boldsymbol{\Lambda}'\mathbf{h}_1 = \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\mu}_2. \tag{27}$$

But matching the upper-right entries in (23) we also see that

$$\boldsymbol{\Lambda}'\mathbf{J}_{12} = -\boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}. \tag{28}$$

Hence,

$$\boldsymbol{\mu}' = \boldsymbol{\Lambda}'\mathbf{h}_1' = \boldsymbol{\Lambda}'(\mathbf{h}_1 - \mathbf{J}_{12}\mathbf{x}_2) = \boldsymbol{\Lambda}'\mathbf{h}_1 - \boldsymbol{\Lambda}'\mathbf{J}_{12}\mathbf{x}_2 = \boldsymbol{\mu}_1 + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \tag{29}$$

as required, where to obtain the first equality we have used (20), and where to obtain the last equality we have used (27) and (28).

As an aside, we have an additional interpretation of (21). Although beyond our current scope, estimation theory establishes that the estimate $\hat{\mathbf{x}}_1(\mathbf{x}_2)$ of $\mathbf{x}_1$ based on observing $\mathbf{x}_2 = \mathbf{x}_2$ that minimizes the mean-square estimation error $\mathbb{E}\left[\|\hat{\mathbf{x}}_1(\mathbf{x}_2) - \mathbf{x}_1\|^2\right]$ is the conditional mean: $\hat{\mathbf{x}}_1(\mathbf{x}_2) = \mathbb{E}\left[\mathbf{x}_1 | \mathbf{x}_2 = \mathbf{x}_2\right]$. From this perspective, we see that in the jointly Gaussian case, $\boldsymbol{\mu}'$ in (21b) is this estimate, which we note turns out to be a *linear*[8] function of $\mathbf{x}_2$. Moreover, $\boldsymbol{\Lambda}'$ in (21c) turns out to be the covariance of the resulting estimation error, which, perhaps surprisingly, does not depend on $\mathbf{x}_2$.

5.3.3 Conditional Marginals

As mentioned at the outset of the previous section, when there are other variables involved, say $\mathbf{x}_3$, we know the conditional marginal $p_{\mathbf{x}_1|\mathbf{x}_2}$ is still Gaussian, and the preceding analysis applies. However, it is important to keep in mind in such cases that the information form parameters are more complicated to obtain. This is because starting from the trivariate distribution for $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ we have to first marginalize to obtain the bivariate distribution for $\mathbf{x}_1, \mathbf{x}_2$, and the resulting 2×2 information matrix bears a complicated relationship to the original 3×3 information matrix, consistent with our earlier discussion of marginalization in information form. Hence, even though the subsequent conditioning is simple in information form, the overall conditional marginal computation in information form is cumbersome due to marginalization.

It is also worth noting that using covariance form doesn't help. Indeed, with this form, the marginalization step becomes simple, but the subsequent conditioning becomes more complicated. In this sense, the complexity (such as involving the computation of Schur complements) is unavoidable in the absence of special structure in the Gaussian model. We examine such structure next.

5.4 Independence Structure in Gaussian distributions

Having discussed the behavior of basic inference operations involving jointly Gaussian random variables, let us now turn to understanding how structure in such distributions expresses itself, which we ultimately exploit to implement inference efficiently.

For our development, we consider, without loss of generality, a collection of jointly Gaussian random variables paritioned into three groups, corresponding to jointly

[8] To be more precise, we should say "affine" to include the constant offsets, but we will use the term "linear" generically.

Gaussian vectors x_1, x_2, and x_3. With x denoting the supervector

$$\mathsf{x} = \begin{bmatrix} \mathsf{x}_1 \\ \mathsf{x}_2 \\ \mathsf{x}_3 \end{bmatrix}, \tag{30}$$

their parameterization in covariance form is

$$\boldsymbol{\mu} = \mathbb{E}[\mathsf{x}] = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \mathbb{E}\left[(\mathsf{x}-\boldsymbol{\mu})(\mathsf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right] = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} & \boldsymbol{\Lambda}_{33} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} & \boldsymbol{\Lambda}_{23} \\ \boldsymbol{\Lambda}_{31} & \boldsymbol{\Lambda}_{32} & \boldsymbol{\Lambda}_{33} \end{bmatrix}, \tag{31}$$

while their parameterization in information form is

$$\mathbf{h} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\mu} = \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \end{bmatrix}, \qquad \mathbf{J} = \boldsymbol{\Lambda}^{-1} = \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} & \mathbf{J}_{33} \\ \mathbf{J}_{21} & \mathbf{J}_{22} & \mathbf{J}_{23} \\ \mathbf{J}_{31} & \mathbf{J}_{32} & \mathbf{J}_{33} \end{bmatrix}. \tag{32}$$

For our development, we continue to focus on a triple of jointly Gaussian random vectors $\mathsf{x}_1, \mathsf{x}_2, \mathsf{x}_3$ whose parameterization is give in (30)–(32)

5.4.1 Marginal Independence

Our main result is the following.

Theorem 1. *We have the marginal independence* $\mathsf{x}_1 \perp\!\!\!\perp \mathsf{x}_2$ *if and only if*

$$\boldsymbol{\Lambda}_{12} = \mathbf{0},$$

i.e., if and only if x_1 *and* x_2 *are uncorrelated.*

This theorem establishes that the marginal independence structure among a collection of jointly Gaussian random variables is expressed in terms of the sparsity of their covariance matrix.

Proof. To verify the "only if" part of the theorem, we simply note that

$$\boldsymbol{\Lambda}_{12} = \mathbb{E}\left[(\mathsf{x}_1-\mu_1)(\mathsf{x}_2-\mu_2)^{\mathrm{T}}\right] = \mathbb{E}\left[(\mathsf{x}_1-\mu_1)\right]\mathbb{E}\left[(\mathsf{x}_2-\mu_2)^{\mathrm{T}}\right] = \mathbf{0},$$

where to obtain the middle equality we have used the independence of x_1 and x_2.

To verify the "if" part of the theorem, we note that from Claim 1 it follows that $\mathsf{x}_1, \mathsf{x}_2$ are jointly Gaussian. Moreover, from the ensuing discussion their mean vector is a subvector of $\boldsymbol{\mu}$ in (31), and their covariance matrix is a submatrix of $\boldsymbol{\Lambda}$ in (31). Specifically, in terms of the parameters defined in (31), we have

$$\mathsf{x}' \triangleq \begin{bmatrix} \mathsf{x}_1 \\ \mathsf{x}_2 \end{bmatrix}, \qquad \boldsymbol{\mu}' = \mathbb{E}[\mathsf{x}'] = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \mathbb{E}\left[(\mathsf{x}'-\boldsymbol{\mu}')(\mathsf{x}'-\boldsymbol{\mu}')^{\mathrm{T}}\right] = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}. \tag{33}$$

Now if $\mathbf{\Lambda}_{12} = \mathbf{0}$ then

$$\begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{\Lambda}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{22}^{-1} \end{bmatrix},$$

so $p_{\mathbf{x}'}(\mathbf{x}')$ factors into the product of a function solely of $\mathbf{x}_1$ and one solely of $\mathbf{x}_2$:

$$\begin{aligned} p_{\mathbf{x}'}(\mathbf{x}') &\propto \exp\left\{-\frac{1}{2}\begin{bmatrix}(\mathbf{x}_1-\boldsymbol{\mu}_1)^{\mathrm{T}} & (\mathbf{x}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{22}^{-1}\end{bmatrix}\begin{bmatrix}(\mathbf{x}_1-\boldsymbol{\mu}_1) \\ (\mathbf{x}_2-\boldsymbol{\mu}_2)\end{bmatrix}\right\} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^{\mathrm{T}}\mathbf{\Lambda}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\right\}\exp\left\{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^{\mathrm{T}}\mathbf{\Lambda}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right\}, \end{aligned}$$

so $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent. □

5.4.2 Conditional Independence

Our main result is the following.

Theorem 2. *We have the conditional independence* $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_2 \mid \mathbf{x}_3$ *if and only if*

$$\mathbf{J}_{12} = \mathbf{0}.$$

This theorem establishes that the conditional independence structure among a collection of jointly Gaussian random variables is expressed in terms of the sparsity of their information matrix.

Proof. To begin, note that with $\mathbf{x}_3$ viewed as a constant (since we are conditioning on it),

$$p_{\mathbf{x}_1,\mathbf{x}_2|\mathbf{x}_3}(\mathbf{x}_1,\mathbf{x}_2|\mathbf{x}_3) \propto p_{\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) \propto \exp\left\{-\frac{1}{2}\sum_{i,j}\mathbf{x}_i^{\mathrm{T}}\mathbf{J}_{ij}\mathbf{x}_j - \sum_i \mathbf{h}_i^{\mathrm{T}}\mathbf{x}_i\right\}. \quad (34)$$

Now $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent conditioned on $\mathbf{x}_3 = \mathbf{x}_3$ if and only if (34) factors into the product of a function solely of $\mathbf{x}_1$ and one solely of $\mathbf{x}_2$. But this happens if and only if there is no single term in the summation in (34) that involves both $\mathbf{x}_1$ and $\mathbf{x}_2$. Now the only such terms are $-(1/2)\mathbf{x}_1^{\mathrm{T}}\mathbf{J}_{12}\mathbf{x}_2$ and $-(1/2)\mathbf{x}_2^{\mathrm{T}}\mathbf{J}_{21}\mathbf{x}_1$, which are identical since $\mathbf{J}_{12} = \mathbf{J}_{21}$, and vanish if and only if $\mathbf{J}_{12} = \mathbf{0}$. □

5.5 Graphical Representation of Gaussian Distributions

The results of Sections 5.3 and 5.4 provide everything we need to know to build probabilistic graphical models for collections of jointly Gaussian random variables.

Before proceeding, we make an important observation about representation of Gaussian distributions. In the case of random variables over finite-alphabets, we can

conceptually and practically represent any distributions of interest using probability tables, as we have discussed. Moreover, the amount of storage required for the table grows with the alphabet size. In the case of continuous-valued random variables, we cannot represent the corresponding distributions as tables.

However, as one can see from the preceding development, for the specific case of jointly Gaussian random variables, the associated distributions *can* be represented using tables. In particular, since the family of Gaussian distributions is "closed" under marginalization and conditioning, in representing any distribution of interest we need only keep track of its parameters, either in covariance or information form. In other words, like discrete distributions, Gaussian distributions have *finite-dimensional* representations. Moreover, the amount of storage required now grows with the *dimension* of the variables involved in the distribution, not their cardinality. Obviously, this has implications for implementation.

5.5.1 Directed Graphical Models

Directed graphical models for Gaussian distributions are straightforward to develop. In particular, as we have discussed, directed graphical models represent factorizations of a joint distribution into the product of conditional distributions, i.e.,

$$p_{\mathbf{x}_1,\ldots,\mathbf{x}_N}(\mathbf{x}_1,\ldots,\mathbf{x}_N) = \prod_{i=1}^{N} p_{\mathbf{x}_i|\mathbf{x}_{\pi_i}}(\mathbf{x}_i|\mathbf{x}_{\pi_i}).$$

In the Gaussian case, all the constituent conditional distributions are Gaussian, so each is characterized by its local covariance or information form parameters. As we have discussed, a key feature of the conditional distributions in the Gaussian case is that the mean or potential vector in the representation is *linear* in the parent variables (and the covariance or information matrix has no dependence on the parent variables). This has important implications for implementation, since the associated mean or potential tables need only encode the parameters of this linear mapping.

When constructing such directed graphical models, we can start from a complete directed graph corresponding to a particular topological ordering and prune away edges according to marginal and conditional independencies that exist in the distribution. As we have discussed, these independencies are expressed through the sparsity structure in the covariance and information matrices, respectively.

5.5.2 Undirected Graphical Models

Undirected graphical models for Gaussian distributions are similarly straightforward to develop. In particular, as we have also discussed, undirected graphical models represent conditional independencies among variables in a joint distribution. In particular, the absence of an edge between nodes i and j expresses that the distribution

satisfies the conditional independency

$$\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j \mid \mathbf{x}_{\mathcal{V}\setminus\{i,j\}}. \tag{35}$$

But from Theorem 2 in Section 5.4.2, we saw that the conditional independency (35) holds if and only if the information matrix for $\mathbf{x}_1, \ldots, \mathbf{x}_N$ satisfies $\mathbf{J}_{ij} = \mathbf{0}$. Thus, this means that undirected graphs for Gaussian distributions can be built from their information form parameterizations with particular ease.

5.6 Factor graphs

Given that we have the definition of jointly Gaussian random variables at hand, let us introduce the third type of graphical models that we study in this class.

As we have seen, directed and undirected graphical models can be used to capture different types of factorization structure in a joint distribution. And as we have discussed, such factorization structure is what allows for more compact descriptions of these joint distributions, and, ultimately, more efficient inference.

In the case of directed graphical models, the constituent factors take the form of conditional probability distributions based on the parent-child structure in the graph. By contrast, in the case of undirected graphical models, the constituent factors take the form of potential functions over cliques of nodes in the graph. Ultimately, these two types of graphs can represent some forms of factorization structure more effectively than others.

An alternative to these types of models is to develop a graphical model that is specifically tailored to fully capturing *arbitrary* factorization structure. This is the objective of *factor graphs*, which we now develop. And we note that while factor graphs provide greater flexibility in capturing factorization structure in the representation of distributions, when used where the other graphical models would have otherwise sufficed, they can lead to higher complexity inference. As such, they augment rather than replace the other two types of graphical models.

5.6.1 Definition and Properties

A factor graph is different from our earlier graphical models in that it consists of two types of nodes: variable nodes, and factor nodes. Precisely, a *factor graph* $\mathcal{G}$ is a bipartite graph with one partition corresponding to variable nodes $\mathcal{V}$ and other partition corresponding to factor nodes $\mathcal{F}$ with (undirected) edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{F}$, i.e. $\mathcal{G} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$. To distinguish them, in our depictions we denote variables nodes using circles, and factor nodes using squares.

Associated with each variable node is one of the variables over which the family of joint distribution is defined. Associated with each factor node is a factor function corresponding to a factor in the distribution. Moreover, the arguments of the factor function associated a factor node are precisely the variable nodes that are connected

to this factor node. Evidently, if there are N variable nodes, there is never a need for more than 2^N factor nodes, and in practice, we are often interested in modeling distributions with far fewer factors than this.

Let $\mathcal{V}_j$ denote the variable nodes to which factor node j has an edge. Then the family of joint distributions corresponding to a factor graph is given by

$$p_{\mathsf{x}_\mathcal{V}}(x_\mathcal{V}) = \frac{1}{Z} \prod_{j \in \mathcal{F}} f_j(x_{\mathcal{V}_j}),$$

where the factors $f_j(\cdot)$ are nonnegative functions but otherwise unconstrained; the constant Z serves to normalize the distribution.

5.6.2 Example: factor graphs of a simple Gaussian distribution

To gain further insights, let's consider a simple instance of jointly Gaussian distribution. In particular, we will consider the family of distributions of the form

$$p_{\mathsf{x},\mathsf{y},\mathsf{z}}(x, y, z) = f_1(x, y)\, f_2(y, z)\, f_3(x, z), \tag{36}$$

as seen in the next example. Families of the form (36), and their generalizations, arise quite naturally.

Example 1. Consider a triple of random variables x, y, and z with joint distribution

$$p_{\mathsf{x},\mathsf{y},\mathsf{z}}(x, y, z) = \frac{1}{2(\pi)^{3/2}} \exp\left\{ -\frac{3}{4}(x^2 + y^2 + z^2) + \frac{1}{2}(xy + xz + yz) \right\},$$

over the alphabet $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} = \mathbb{R}^3$, which we see can be written in the factored form (36) with

$$\begin{aligned} f_1(x, y) &= \exp\left\{ -\frac{3}{4}x^2 + \frac{1}{2}xy \right\} \\ f_2(y, z) &= \exp\left\{ -\frac{3}{4}y^2 + \frac{1}{2}yz \right\} \\ f_3(x, z) &= \exp\left\{ -\frac{3}{4}z^2 + \frac{1}{2}xz \right\}. \end{aligned}$$

This is an example of a particular multivariate Gaussian distribution. Note that it is straightforward to verify by integration and examination of factors that there no unconditional independencies among these variables. Also, a related calculation reveals that there are no (nontrivial) conditional independencies either.

From the definition of factor graphs, the joint distribution can be efficiently expressed using the factor graph of Fig. 1a. If we try to express this distribution using an undirected graph, we are forced to use the fully connected graph depicted in Fig. 1b.

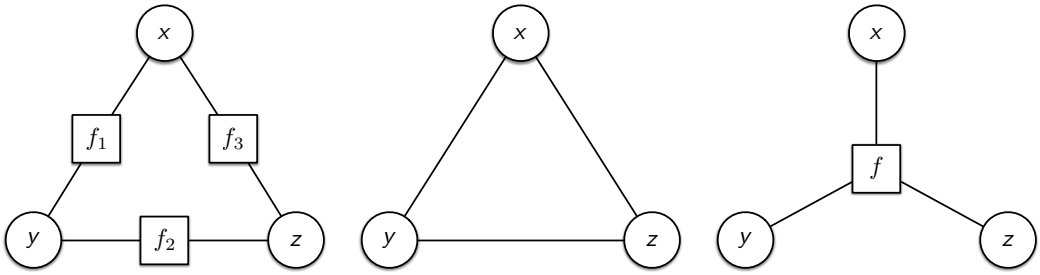

Figure 1: (a) Factor graph representing distributions of the form (36). (b) Undirected graph representing distributions of the form (36). (c) Factor graph corresponding to undirected graph of Fig. 1b.

To emphasize the distinction, it is worth observing that the factor graph corresponding to the (full) family of distributions represented by Fig. 1b is Fig. 1c, which is quite different from that in Fig. 1a. Thus, it is an example where the factor graph is capable of capturing the factorization structure in the joint distribution, but the undirected graph is not.

As seen in the example, it is worth emphasizing that unlike undirected graphical models, factor graphs are not constrained to have factors only for maximal cliques, so we have more flexibility in capturing factorization structure in joint distributions. This typically results in factor graphs corresponding to more compact representations of joint distributions. Evidently, the complexity of description (number of parameters) of members of the family of joint distributions captured by a factor graph is

$$\sum_{j\in\mathcal{F}} |\mathcal{X}|^{|\mathcal{V}_j|} \leq |\mathcal{F}|\,|\mathcal{X}|^D, \qquad D = \max_{j\in\mathcal{F}} |\mathcal{V}_j|, \tag{37}$$

where we comment that $|\mathcal{F}|$ can be exponential in $N = |\mathcal{V}|$ in some cases, as explored in the homework.

It is worth emphasizing that, like our other graphical model types, factor graphs are universal, in that the can represent any distribution. Indeed, a factor graph consisting of a single factor connected to all N variable nodes corresponds to the family of all possible distributions.

The preceding discussion does not mean, however, that our previously developed directed and undirected graphical models are no longer needed. Indeed, when the factorization of a distribution matches the forms for which these graphs are specifically tailored, this often leads to the most efficient inference, as we will see later in the subject.

5.6.3 Factor Graph Representation of Gaussian Distributions

The *pairwise* factorization of jointly Gaussian distributions observed in Example 1 can be extended to all Gaussian distributions, as we develop now. Gaussian graphical models are an instance of *pairwise* models, i.e., models characterized by their pairwise potentials. To see this, note that the distribution of a collection of variables

$$\mathbf{x} = \big[\mathbf{x}_1, \ldots, \mathbf{x}_N\big]$$

can be expressed in information form via

$$p_{\mathbf{x}}(\mathbf{x}) \propto \exp\left\{-\sum_{i,j} \mathbf{x}_i^{\mathrm{T}} \mathbf{J}_{ij} \mathbf{x}_j + \sum_i \mathbf{h}_i^{\mathrm{T}} \mathbf{x}_i\right\} \tag{38}$$

$$= \prod_{i,j} \underbrace{\exp\left\{-\mathbf{x}_i^{\mathrm{T}} \mathbf{J}_{ij} \mathbf{x}_j\right\}}_{\triangleq f_{ij}(\mathbf{x}_i, \mathbf{x}_j)} \prod_i \underbrace{\exp\left\{\mathbf{h}^{\mathrm{T}} \mathbf{x}_i\right\}}_{\triangleq f_i(\mathbf{x}_i)}. \tag{39}$$

Hence, in the Gaussian case, factors never involve more than two nodes. By contrast, when building undirected graphical models for Gaussian distributions, the structure in the information matrix can be such that there are cliques of size greater than two in the resulting graph. In such cases, the undirected graph alone expresses a less efficient factorization of the distribution. As a result—and as with other pairwise models—the structure in Gaussian graphical models is sometimes expressed using factor graphs, as we have already seen in Example 1.

5.6.4 More Factor Graph Examples (optional)

Let us consider additional illustrative examples of factor graphs that arise in different application domains. First, let's consider a familiar puzzle game.

Example 2. The game of Sudoku can be viewed as a problem of inference on a factor graph. With the (i, j)th position in the 9×9 grid, $i, j \in \{1, \ldots, 9\}$ we associate a variable $\mathsf{x}_{i,j} \in \mathcal{X} = \{1, \ldots, 9\}$. Let

$$g(u_1, \ldots, u_9) \triangleq \prod_{i \in \{1,\ldots,9\}} \prod_{\{j \in \{1,\ldots,9\}\colon\, j \neq i\}} \mathbb{1}_{u_i \neq u_j},$$

which we note is 1 if and only all arguments take distinct values. Then the grid row constraints (that the numbers along each row must be distinct) can be expressed as

$$f_i^{\mathrm{r}}(x_{i,\{1,\ldots,9\}}) = g(x_{i,\{1,\ldots,9\}}) = 1, \qquad i = 1, \ldots, 9,$$

the grid column constraints (that the numbers down each column must be distinct) can be expressed as

$$f_j^{\mathrm{c}}(x_{\{1,\ldots,9\},j}) = g(x_{\{1,\ldots,9\},j}) = 1, \qquad j = 1, \ldots, 9,$$

and the 3×3 subgrid constraints (that the numbers within each of the 9 subgrids must be distinct) can be expressed as

$$f^{\mathrm{s}}_{k,l}(x_{\{1+(k-1),2+(k-1),3+(k-1)\},\{1+(l-1),2+(l-1),3+(l-1)\}}) \\ = g(x_{\{1+3(k-1),2+3(k-1),3+3(k-1)\},\{1+3(l-1),2+3(l-1),3+3(l-1)\}}) = 1, \\ (k,l) \in \{1,2,3\} \times \{1,2,3\}.$$

With this notation, the distribution over $\{x_{i,j},\ i,j \in \{1,\ldots,9\}\}$ such that all valid configurations are equally likely is

$$p(\{x_{i,j},\ i,j \in \{1,\ldots,9\}) \propto \prod_{i=1}^{9} f^{\mathrm{r}}(x_{i,\{1,\ldots,9\}}) \prod_{j=1}^{9} f^{\mathrm{c}}_{j}(x_{\{1,\ldots,9\},j}) \\ \cdot \prod_{k,l=1}^{9} f^{\mathrm{s}}_{k,l}\big(x_{\{1+(k-1),2+(k-1),3+(k-1)\},\{1+(l-1),2+(l-1),3+(l-1)\}}\big),$$

which can be expressed as a factor graph with 81 variables nodes and 27 factor nodes.

The goal of the game is, when some of the variables are specified, to determine the remaining ones. This can be interpreted as a posterior marginal computation, where the conditioning is on the specified variables. When the given variables are such that solution is unique, as the game is usually configured, the posterior marginals are singular distributions.

Next, let's explore the factor graph corresponding to the simple error-correcting code introduced earlier in the notes.

Example 3. Consider the so-called $(7, 4)$ Hamming code, which can be represented by a factor graph. In this code, there are 4 information bits, represented by x_1, x_2, x_3, x_4, each of which is independent and equally likely to be 0 or 1. There are 3 parity[9] bits, represented by x_5, x_6, x_7, which are generated from the information bits as follows

$$x_5 = x_1 \oplus x_2 \oplus x_4$$
$$x_6 = x_1 \oplus x_3 \oplus x_4$$
$$x_7 = x_2 \oplus x_3 \oplus x_4,$$

where $\oplus$ denotes the binary addition (i.e., exclusive-OR) operator. The (7-bit) codeword is the concatenation of the information bits with the parity bits.

Since all 16 sequences of 4 information bits are equally likely, so are the corresponding 16 codewords, and thus

$$p_{x_1,\ldots,x_7}(x_1,\ldots,x_7) = \frac{1}{16}\, f_{\mathrm{a}}(x_1,x_2,x_4,x_5)\, f_{\mathrm{b}}(x_1,x_3,x_4,x_6)\, f_{\mathrm{c}}(x_2,x_3,x_4,x_7), \quad (40a)$$

[9]Recall that the parity of a collection of bits indicates whether the number of 1 bits is even or odd.

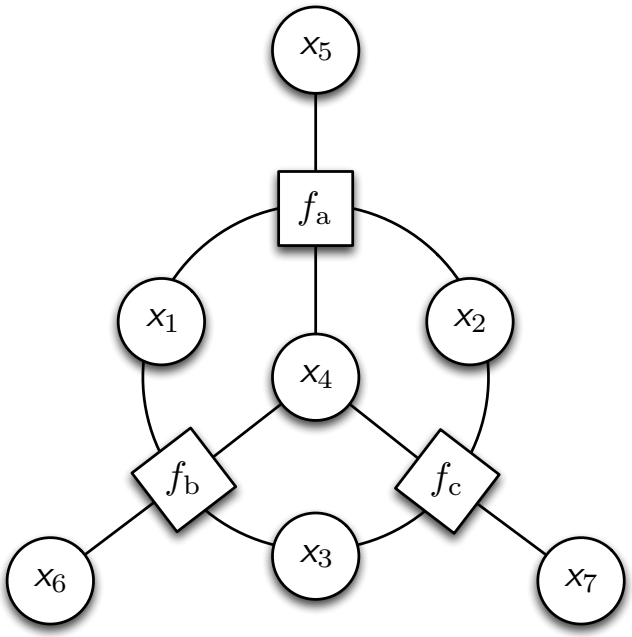

Figure 2: Factor graph representing the $(7, 4)$ Hamming code.

where

$$f_{\mathrm{a}}(x_1, x_2, x_4, x_5) = \mathbb{1}_{x_1 \oplus x_2 \oplus x_4 \oplus x_5 = 0} \tag{40b}$$

$$f_{\mathrm{b}}(x_1, x_3, x_4, x_6) = \mathbb{1}_{x_1 \oplus x_3 \oplus x_4 \oplus x_6 = 0} \tag{40c}$$

$$f_{\mathrm{c}}(x_2, x_3, x_4, x_7) = \mathbb{1}_{x_2 \oplus x_3 \oplus x_4 \oplus x_7 = 0}. \tag{40d}$$

The factor graph representation of the Hamming codeword distribution (40) is depicted in Fig. 2.

For comparison, suppose we were to represent (40) using an undirected graphical model. Then we would need to treat each of the factors f_{a}, f_{b}, and f_{c} as clique potentials. The result is depicted in Fig. 3. Note that this requires introducing edges to reflect this clique structure. However, the effect of this is that Fig. 3 ultimately describes a richer family of distributions than the factor graph of Fig. 2. For example, note that the nodes corresponding to x_1, x_2, x_3, x_4 form a clique, so that a factor $\psi(x_1, x_2, x_3, x_4)$ will also appear in the factorization of the distribution associated with this undirected graph. In this sense, the factor graph is a better representation for the Hamming codeword distribution—it only includes the necessary factors, and thus gives a more compact representation.

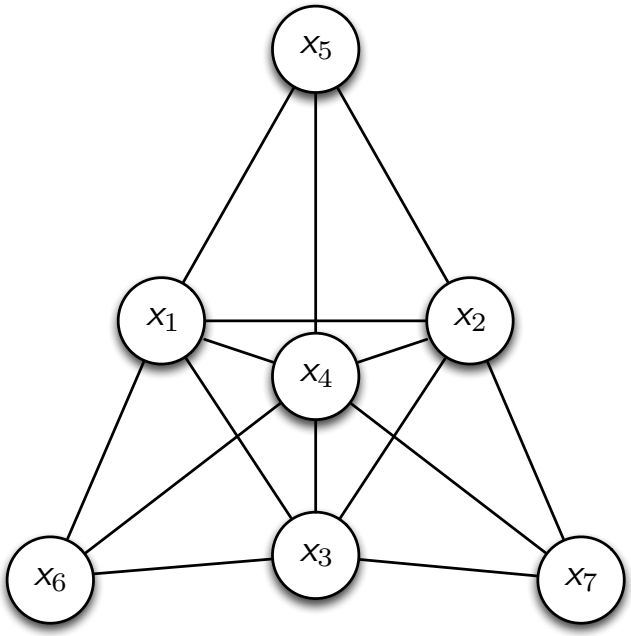

Figure 3: Undirected graphical model representing the $(7, 4)$ Hamming code.