6 Gaussian Graphical Models

Today we describe how collections of jointly Gaussian random variables can be represented as directed and undirected graphical models. Our knowledge of these graphical models will all carry over to the Gaussian case with the added benefit that Gaussian random variables will allow us to exploit a variety of linear algebra tools.

Why focus on Gaussians rather than continuous distributions in general? The choice of having a special case for Gaussians is warranted by the many nice properties Gaussian random variables possess. For example, the Gaussian distribution is an example of a *stable* family, meaning that if we add any two independent Gaussians, we get another Gaussian. The Gaussian distribution is the only continuous, stable family with finite variance. Moreover, the Central Limit Theorem suggests that the family is an "attractor" since summing many i.i.d. random variables that need not be Gaussian results in a random variable that converges in distribution to a Gaussian. In fact, under mild conditions, we need not require the random variables being summed to be identically distributed either.

6.1 Multivariate (Jointly) Gaussian Random Variables

There are many equivalent ways to define multivariate Gaussians, also called Gaussian random vectors. Here are a few characterizations for random vector $\mathbf{x}$ being multivariate Gaussian:

(i) Linear combination of i.i.d. scalar Gaussian variables: There exists some matrix $\mathbf{A}$, constant vector $\mathbf{b}$ and random vector $\mathbf{u}$ of i.i.d. $\mathtt{N}(0,1)$ entries such that $\mathbf{x} = \mathbf{A}\mathbf{u} + \mathbf{b}$.

(ii) All linear combinations of elements of $\mathbf{x}$ are scalar Gaussian random variables: $y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ is Gaussian for all $\mathbf{a}$.

(iii) Covariance form: The probability density function of $\mathbf{x}$ can be written as

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

denoted as $\mathbf{x} \sim \mathtt{N}(\boldsymbol{\mu}, \mathbf{\Lambda})$ with mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ and covariance matrix $\mathbf{\Lambda} = \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\right]$.

(iv) Information form: The probability density function of $\mathbf{x}$ can be written as

$$p_{\mathbf{x}}(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{J}\mathbf{x} + \mathbf{h}^{\mathrm{T}}\mathbf{x}\right\},$$

denoted as $\mathsf{x} \sim \mathtt{N}^{-1}(\mathbf{h}, \mathbf{J})$ with potential $\mathbf{h}$ and information (or precision) matrix $\mathbf{J}$. Note that $\mathbf{J} = \mathbf{\Lambda}^{-1}$ and $\mathbf{h} = \mathbf{J}\boldsymbol{\mu}$.

We will focus on the last two characterizations, while exploiting the first two as key properties.

6.2 Operations on Gaussian random vectors

For the covariance and information forms, we consider how marginalization and conditioning operations are done. Let

$$\mathsf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \sim \mathtt{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{21} & \mathbf{\Lambda}_{22} \end{bmatrix} \right) = \mathtt{N}^{-1}\left(\begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} \right).$$

Marginalization is easy when we have x represented in covariance form: Due to characterization (ii) from earlier, marginals of x are Gaussian. Computing marginals just involves reading off entries from $\boldsymbol{\mu}$ and $\mathbf{\Lambda}$, e.g.

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

In contrast, computing marginals using the information form is more complicated:

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

where $\mathbf{h}' = \mathbf{h}_1 - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{h}_2$ and $\mathbf{J}' = \mathbf{J}_{11} - \mathbf{J}_{12}\mathbf{J}_{22}^{-1}\mathbf{J}_{21}$. The expression for $\mathbf{J}'$ is called the *Schur complement*.

Conditioning is easy when we have x represented in information form: We use the fact that conditionals of a Gaussian random vector are Gaussian. Setting conditioning variables constant in the joint distribution and reading off the quadratic form of the remaining variables, it becomes apparent that conditioning involves reading off entries from $\mathbf{J}$, e.g., when conditioning on $\mathbf{x}_2$,

$$\begin{aligned}
p_{\mathbf{x}_1|\mathbf{x}_2}(\mathbf{x}_1|\mathbf{x}_2) &\propto p_{\mathbf{x}_1,\mathbf{x}_2}(\mathbf{x}_1, \mathbf{x}_2) \\
&\propto \exp\left\{ -\frac{1}{2} \begin{pmatrix} \mathbf{x}_1^{\mathrm{T}} & \mathbf{x}_2^{\mathrm{T}} \end{pmatrix} \begin{bmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{bmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} + \begin{pmatrix} \mathbf{h}_1^{\mathrm{T}} & \mathbf{h}_2^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \right\} \\
&= \exp\left\{ -\frac{1}{2} \left(\mathbf{x}_1^{\mathrm{T}} \mathbf{J}_{11} \mathbf{x}_1 + 2\mathbf{x}_2^{\mathrm{T}} \mathbf{J}_{21} \mathbf{x}_1 + \mathbf{x}_2^{\mathrm{T}} \mathbf{J}_{22} \mathbf{x}_2 \right) + \mathbf{h}_1^{\mathrm{T}} \mathbf{x}_1 + \mathbf{h}_2^{\mathrm{T}} \mathbf{x}_2 \right\} \\
&= \exp\left\{ -\frac{1}{2} \mathbf{x}_1^{\mathrm{T}} \mathbf{J}_{11} \mathbf{x}_1 + (\mathbf{h}_1^{\mathrm{T}} - \mathbf{x}_2^{\mathrm{T}} \mathbf{J}_{21}) \mathbf{x}_1 + \mathbf{h}_2^{\mathrm{T}} \mathbf{x}_2 - \frac{1}{2} \mathbf{x}_2^{\mathrm{T}} \mathbf{J}_{22} \mathbf{x}_2 \right\} \\
&= \exp\left\{ -\frac{1}{2} \mathbf{x}_1^{\mathrm{T}} \mathbf{J}_{11} \mathbf{x}_1 + (\mathbf{h}_1 - \mathbf{J}_{12} \mathbf{x}_2)^{\mathrm{T}} \mathbf{x}_1 + \mathbf{h}_2^{\mathrm{T}} \mathbf{x}_2 - \frac{1}{2} \mathbf{x}_2^{\mathrm{T}} \mathbf{J}_{22} \mathbf{x}_2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2} \mathbf{x}_1^{\mathrm{T}} \mathbf{J}_{11} \mathbf{x}_1 + (\mathbf{h}_1 - \mathbf{J}_{12} \mathbf{x}_2)^{\mathrm{T}} \mathbf{x}_1 \right\},
\end{aligned}$$

where the last step uses the fact that $\mathbf{x}_2$, which we are conditioning on, is treated as a constant. In particular, we see that

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

where $\mathbf{h}_1' = \mathbf{h}_1 - \mathbf{J}_{12}\mathbf{x}_2$. While we can read off entries of $\mathbf{J}$ to obtain the information matrix for $\mathbf{x}_1|\mathbf{x}_2$, namely $\mathbf{J}_{11}$, the potential vector needs to be updated. Note that conditioning using the covariance form is more complicated, involving a Schur complement:

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

where $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ and $\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}_{11} - \boldsymbol{\Lambda}_{12}\boldsymbol{\Lambda}_{22}^{-1}\boldsymbol{\Lambda}_{21}$.

We can interpret the conditional distribution as follows. Note that $\boldsymbol{\mu}' = \mathbb{E}[\mathbf{x}_1|\mathbf{x}_2]$, also known as the Bayes least-squares estimate of $\mathbf{x}_1$ from $\mathbf{x}_2$, is linear in $\mathbf{x}_2$, a special property of Gaussians. Moreover,

$$\boldsymbol{\mu}' = \underset{\substack{\hat{\mathbf{x}}_1(\mathbf{x}_2) \text{ s.t.} \\ \hat{\mathbf{x}}_1(\mathbf{x}_2)=\mathbf{A}\mathbf{x}_2+\mathbf{b}}}{\arg\min} \quad \mathbb{E}\left[\|\mathbf{x}_1 - \hat{\mathbf{x}}_1(\mathbf{x}_2)\|^2\right],$$

where $\boldsymbol{\Lambda}'$ is the resulting mean-square error for estimater $\boldsymbol{\mu}'$.

We see that both the covariance and information forms are useful depending on whether we are marginalizing or conditioning. Converting between the two requires matrix inversion, e.g., solving linear equations. This involves Gaussian elimination and use of the Schur complement, which we will say a little more about at the end of today's lecture.

6.3 Gaussian graphical models

To represent a Gaussian random vector as a graphical model, we will need to know conditional independencies. From $\boldsymbol{\Lambda}$ and $\mathbf{J}$, we can read off the following independencies:

Theorem 1. *For* $\mathbf{x} \sim \mathtt{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$, $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j$ *if and only if* $\boldsymbol{\Lambda}_{ij} = \mathbf{0}$.

Theorem 2. *For* $\mathbf{x} \sim \mathtt{N}^{-1}(\mathbf{h}, \mathbf{J})$, $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j|\mathbf{x}_{rest}$ *if and only if* $\mathbf{J}_{ij} = \mathbf{0}$.

The information matrix $\mathbf{J}$ is particularly useful since it describes pairwise Markov conditional independencies and encodes a minimal undirected I-map for $\mathbf{x}$. To obtain the undirected Gaussian graphical model from $\mathbf{J}$, add an edge between $\mathbf{x}_i$ and $\mathbf{x}_j$ whenever $\mathbf{J}_{ij} \neq \mathbf{0}$. To obtain a Gaussian directed graphical model, choose an ordering of the $\mathbf{x}_i$'s and apply the chain rule:

$$p_{\mathbf{x}_1,\ldots,\mathbf{x}_n} = p_{\mathbf{x}_1} p_{\mathbf{x}_2|\mathbf{x}_1} p_{\mathbf{x}_3|\mathbf{x}_2,\mathbf{x}_1} \cdots p_{\mathbf{x}_n|\mathbf{x}_{n-1},\ldots,\mathbf{x}_1}.$$

Note that each factor on the right-hand side is Gaussian, with mean linear in its parents, due to the Bayes least-square estimate being linear as discussed previously.

6.4 Gaussian Markov process

An example of a directed Gaussian graphical model is the Gauss-Markov process, shown in Figure 1. We can express the process in *innovation form*:

$$\mathbf{x}_{i+1} = \mathbf{A}\mathbf{x}_i + \mathbf{B}\mathbf{v}_i, \tag{1}$$

where $\mathbf{x}_1 \sim \mathtt{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$, $\mathbf{v}_i \sim$ i.i.d. $\mathtt{N}(\mathbf{0}, \boldsymbol{\Lambda}_\mathbf{v})$ are independent of $\mathbf{x}_1$, and $\mathbf{B}\mathbf{v}_i$ is called the *innovation*. This is a linear dynamical system because the evolution model is linear.

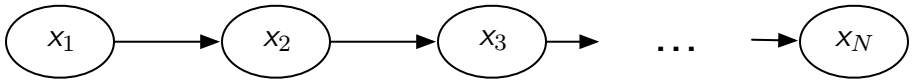

Figure 1: Gauss-Markov process.

Consider the case where we do not observe the $\mathbf{x}_i$'s directly, i.e., $\mathbf{x}_i$'s are hidden, but we observe $\mathbf{y}_i$ related to each $\mathbf{x}_i$ through a Gaussian conditional probability distribution:

$$\mathbf{y}_i = \mathbf{C}\mathbf{x}_i + \mathbf{w}_i, \tag{2}$$

where $\mathbf{w}_i \sim$ i.i.d. $\mathtt{N}(\mathbf{0}, \boldsymbol{\Lambda}_\mathbf{w})$ independent of $\mathbf{v}_j$'s and $\mathbf{x}_1$. The resulting graphical model is shown in Figure 2 Collectively, equations (1) and (2) are referred to as *standard state space form*.

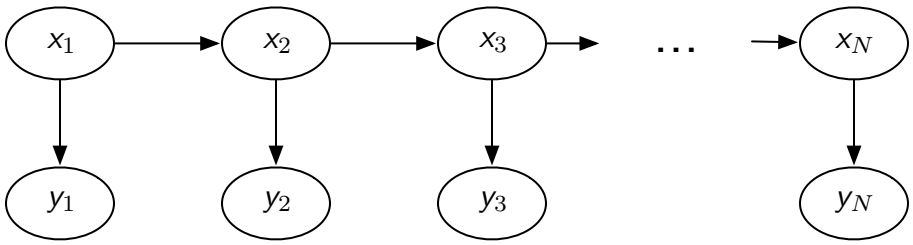

Figure 2: Hidden Gauss-Markov process.

Generally, Gaussian inference involves exploiting linear algebraic structure.

6.5 Matrix inversion

Lastly, we develop the matrix inversion lemma. Let

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

where $\mathbf{E}$ and $\mathbf{H}$ are invertible. We want to invert $\mathbf{M}$. First, we block diagonalize via pre- and post-multiplying:

$$\underbrace{\begin{bmatrix}\mathbf{I} & -\mathbf{F}\mathbf{H}^{-1}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}}_{\mathbf{X}}\underbrace{\begin{bmatrix}\mathbf{E} & \mathbf{F}\\ \mathbf{G} & \mathbf{H}\end{bmatrix}}_{\mathbf{M}}\underbrace{\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{H}^{-1}\mathbf{G} & \mathbf{I}\end{bmatrix}}_{\mathbf{Z}} = \underbrace{\begin{bmatrix}\overbrace{\mathbf{E}-\mathbf{F}\mathbf{H}^{-1}\mathbf{G}}^{\mathbf{M}/\mathbf{H}} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}}_{\mathbf{W}},$$

where, $\mathbf{M}/\mathbf{H}$ is the Schur complement. Noting that $\mathbf{W}^{-1} = \mathbf{Z}^{-1}\mathbf{M}^{-1}\mathbf{X}^{-1}$, then $\mathbf{M}^{-1}$ is given by

$$\begin{aligned}\mathbf{M}^{-1} &= \mathbf{Z}\mathbf{W}^{-1}\mathbf{X}\\ &= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{H}^{-1}\mathbf{G} & \mathbf{I}\end{bmatrix}\left(\begin{bmatrix}(\mathbf{M}/\mathbf{H}) & \mathbf{0}\\ \mathbf{0} & \mathbf{H}\end{bmatrix}\right)^{-1}\begin{bmatrix}\mathbf{I} & -\mathbf{F}\mathbf{H}^{-1}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{H}^{-1}\mathbf{G} & \mathbf{I}\end{bmatrix}\begin{bmatrix}(\mathbf{M}/\mathbf{H})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{I} & -\mathbf{F}\mathbf{H}^{-1}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}.\end{aligned}$$

Taking the determinant of both sides of the above equation yields

$$\begin{aligned}|\mathbf{M}|^{-1} &= |\mathbf{M}^{-1}|\\ &= \left|\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{H}^{-1}\mathbf{G} & \mathbf{I}\end{bmatrix}\right|\left|\begin{bmatrix}(\mathbf{M}/\mathbf{H})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}^{-1}\end{bmatrix}\right|\left|\begin{bmatrix}\mathbf{I} & -\mathbf{F}\mathbf{H}^{-1}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\right|\\ &= \left|\begin{bmatrix}(\mathbf{M}/\mathbf{H})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{H}^{-1}\end{bmatrix}\right|\\ &= |(\mathbf{M}/\mathbf{H})^{-1}||\mathbf{H}|^{-1}\\ &= |(\mathbf{M}/\mathbf{H})|^{-1}|\mathbf{H}|^{-1},\end{aligned} \tag{3}$$

where we use the fact that

$$\left|\begin{bmatrix}\mathbf{A} & 0\\ \mathbf{C} & \mathbf{D}\end{bmatrix}\right| = \left|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ 0 & \mathbf{D}\end{bmatrix}\right| = |\mathbf{A}||\mathbf{D}|$$

whenever $\mathbf{A}$ and $\mathbf{D}$ are square matrices. Rearranging terms in equation (3) gives $|\mathbf{M}| = |\mathbf{M}/\mathbf{H}||\mathbf{H}|$, hence the notation for the Schur complement.

We could alternatively decompose $\mathbf{M}$ in terms of $\mathbf{E}$ and $\mathbf{M}/\mathbf{E} = \mathbf{H} - \mathbf{G}\mathbf{E}^{-1}\mathbf{F}$ to get an expression for $\mathbf{M}^{-1}$, which looks similar to the above equation. Equating the two and rearranging terms gives the *matrix inversion lemma*:

$$(\underbrace{\mathbf{E}-\mathbf{F}\mathbf{H}^{-1}\mathbf{G}}_{\mathbf{M}/\mathbf{H}})^{-1} = \mathbf{E}^{-1} + \mathbf{E}^{-1}\mathbf{F}(\underbrace{\mathbf{H}-\mathbf{G}\mathbf{E}^{-1}\mathbf{F}}_{\mathbf{M}/\mathbf{E}})^{-1}\mathbf{G}\mathbf{E}^{-1}.$$

This will be useful later when we develop Gaussian inference algorithms.

6.438 Algorithms for Inference

Fall 2014