Probabilistic Graphical Models

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2021)
http://cs182.sist.shanghaitech.edu.cn

Outline

Outline

Introduction

- Probabilities play a central role in machine learning. So far, we have formulated and solved complicated probabilistic models purely by algebraic manipulation.
- Probabilistic graphical models may be seen as diagrammatic representations of probability distributions.
  - They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.
  - Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph.
  - Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations (i.e., general-purpose graph algorithms), in which underlying mathematical expressions are carried along implicitly.

Graphical Representations

- A graph comprises nodes (a.k.a. vertices) connected by links (a.k.a. edges or arcs).
- In a probabilistic graphical model, each node represents a random variable or group of random variables and each link expresses a probabilistic relationship between variables.
- A probabilistic graphical model captures the way in which the joint distribution over all the random variables can be decomposed into a product of factors each depending only on a subset of the random variables.
- Two major types of probabilistic graphical models:
  - Directed graphical models (a.k.a. Bayesian networks)
  - Undirected graphical models (a.k.a. Markov random fields).
- Directed graphs are useful for expressing causal relationships between random variables, whereas undirected graphs are better suited to expressing soft constraints between random variables.
- Only Bayesian networks will be covered here.

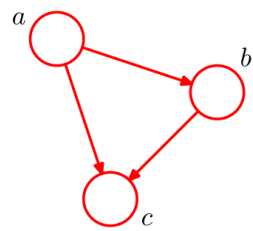
Outline

An Illustrative Example: Fully Connected

- Decomposition of joint distribution into factors:

$$p(a, b, c) = p(c \mid a, b)p(a, b) = p(c \mid a, b)p(b \mid a)p(a)$$

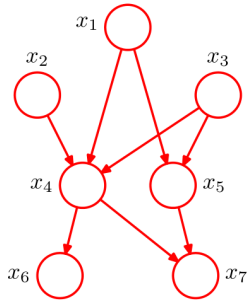
- We introduce a node for each of the random variables a, b, and c and associate each node with the corresponding conditional distribution.
- Node a is a parent of node b; node b is a child of node a.
- This graph is fully connected because there is a link between every pair of nodes.



- This is not the only possible decomposition and graphical representation for $p(a, b, c)$. E.g., another decomposition:

$$p(a, b, c) = p(a \mid b, c)p(c \mid b)p(b)$$

An Illustrative Example: Not Fully Connected



- This is not a fully connected graph because, for instance, there is no link from x_1 to x_2 or from x_3 to x_7.
- Decomposition of joint distribution into factors:

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4 \mid x_1, x_2, x_3) \cdot p(x_5 \mid x_1, x_3)p(x_6 \mid x_4)p(x_7 \mid x_4, x_5)$$

General Directed Acyclic Graphs

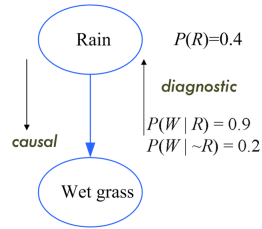
- For a directed acyclic graph (DAG) with K nodes, the factorization of the joint distribution over the K variables is given by:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k \mid \mathrm{pa}_k)$$

  where $\mathbf{x} = \{x_1, \ldots, x_K\}$ and pa_k denotes the set of parents of x_k.
- The joint distribution defined by a DAG is given by the product, over all of the nodes of the graph, of a conditional distribution for each node conditioned on the variables corresponding to the parents of that node in the graph.
- This key equation expresses the factorization properties of the joint distribution for a directed graphical model.
- In a DAG, there always exists an ordering of the nodes such that no links go from any node to any lower-numbered node, but this ordering is not necessarily unique.

Causal Graph and Diagnostic Inference

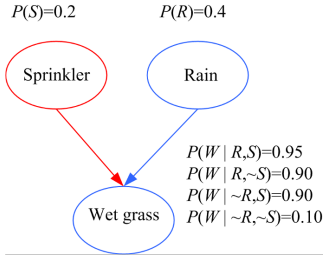
- The graph models that rain causes the grass to get wet.
- Causal graph: rain is the cause of wet grass.
- Diagnostic inference: knowing that the grass is wet, what is the probability that rain is the cause?



- Bayes' rule:

$$
\begin{aligned}
P(R \mid W) &= \frac{P(W \mid R)P(R)}{P(W)} \\
&= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)} \\
&= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = \frac{0.36}{0.48} = 0.75 > P(R) = 0.4
\end{aligned}
$$

Two Causes: Causal Inference



- Causal (or predictive) inference: if the sprinkler is on, what is the probability that the grass is wet?

$$
\begin{aligned}
P(W \mid S) = \sum_R P(W, R \mid S) &= P(W \mid R, S)P(R \mid S) + P(W \mid \sim R, S)P(\sim R \mid S) \\
&= P(W \mid R, S)P(R) + P(W \mid \sim R, S)\, P(\sim R) \\
&= 0.95 \times 0.4 + 0.9 \times 0.6 \\
&= 0.92
\end{aligned}
$$

Two Causes: Diagnostic Inference

- Diagnostic inference: if the grass is wet, what is the probability that the sprinkler is on?

$$P(S \mid W) = \frac{P(W \mid S)P(S)}{P(W)}$$

where (calculated by marginalizing over the joint)

$$\begin{aligned} P(W) =& \sum_{R,S} P(W, R, S) \\ =& P(W \mid R, S)P(R, S) + P(W \mid \sim R, S)P(\sim R, S) \\ &+ P(W \mid R, \sim S)P(R, \sim S) + P(W \mid \sim R, \sim S)P(\sim R, \sim S) \\ =& P(W \mid R, S)P(R)P(S) + P(W \mid \sim R, S)P(\sim R)P(S) \\ &+ P(W \mid R, \sim S)P(R)P(\sim S) + P(W \mid \sim R, \sim S)P(\sim R)P(\sim S) \\ =& 0.52 \end{aligned}$$

So

$$P(S \mid W) = \frac{0.92 \times 0.2}{0.52} = 0.35 > P(S) = 0.2$$

Two Causes: Explaining Away

- Let us assume that it rained, then (Bayes' rule)

$$P(S \mid R, W) = \frac{P(W \mid R, S)P(S \mid R)}{P(W \mid R)} = \frac{P(W \mid R, S)P(S)}{P(W \mid R)} = 0.21$$

- Explaining away:

$$0.21 = P(S \mid R, W) < P(S \mid W) = 0.35$$

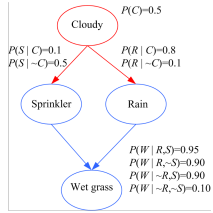
  Knowing that it has rained decreases the probability that the sprinkler is on.

- Knowing that the grass is wet, rain and sprinkler become dependent:

$$P(S \mid R, W) \neq P(S \mid W)$$

Dependent Causes

- Bayesian networks explicitly encode independencies and allow breaking down inference into calculation over small groups of variables propagated from evidence nodes to query nodes.



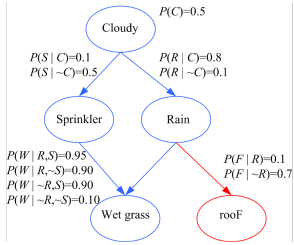
- Causal inference:

$$
\begin{aligned}
P(W \mid C) &= \sum_{R,S} P(W, R, S \mid C) \\
&= P(W \mid R, S, C)P(R, S \mid C) + P(W \mid \sim R, S, C)P(\sim R, S \mid C) \\
&\quad + P(W \mid R, \sim S, C)P(R, \sim S \mid C) + P(W \mid \sim R, \sim S, C)P(\sim R, \sim S \mid C) \\
&= P(W \mid R, S)P(R \mid C)P(S \mid C) + P(W \mid \sim R, S)P(\sim R \mid C)P(S \mid C) \\
&\quad + P(W \mid R, \sim S)P(R \mid C)P(\sim S \mid C) + P(W \mid \sim R, \sim S)P(\sim R \mid C)P(\sim S \mid C)
\end{aligned}
$$

- Independence: W and C are independent given R and S; R and S are independent given C.
- We can also calculate $P(C \mid W)$ for a diagnostic inference.

Local Structures

- The graphical representation is visual and helps understanding.
- The network represents conditional independence statements. The joint distribution can be broken down into local structures, which eases, analysis, computation, and inference:



$$P(C, S, R, W, F) = P(C)P(S \mid C)P(R \mid C) \cdot P(W \mid S, R)P(F \mid R)$$

- In the previous example, the variables are binary. In general,

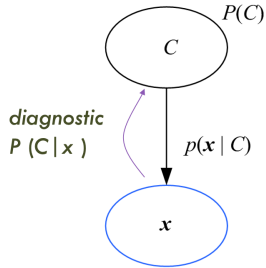
$$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid \text{pa}_i)$$

where x_i is either continuous or discrete with ≥ 2 possible values.

Outline

Bayesian Networks for Classification

- Graphical models are frequently used to visualize generative models for representing the process that we believe has created the data.
- For classification, the corresponding graphical model:



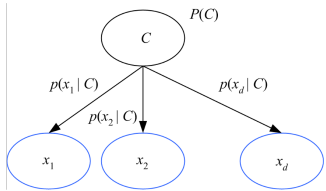
- Bayes' rule inverts the edge:

$$P(C \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C)P(C)}{P(\mathbf{x})}$$

- Classification as diagnostic inference under a Bayesian network formulation.

Naive Bayes' Classier

- As a special case for its computational advantages, the naive Bayes' classifier ignores possible dependencies, namely, correlations, among the input variables and reduces a multivariate problem to a group of univariate problems.



- Given C, it assumes that the input variables x_j are independent:

$$p(\mathbf{x} \mid C) = \prod_{j=1}^{d} p(x_j \mid C)$$

Outline

An Example

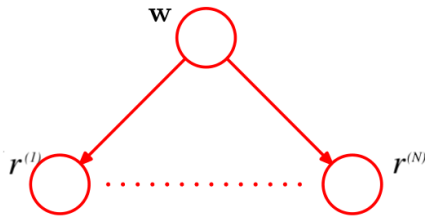
- Training data for (polynomial) regression problem:
  - Input data: $\mathbf{x} = (x^{(1)}, \ldots, x^{(N)})^T$
  - Observed data (dependent variables): $\mathbf{r} = (r^{(1)}, \ldots, r^{(N)})^T$
- Prediction problem: predict the value of t for a new test point x.
- Directed graphical model for this regression problem:
  - Random variables:
    - Vector of coefficients/weights: $\mathbf{w}$
    - Observed data: $\mathbf{r}$
  - Parameters:
    - Input data: $\mathbf{x}$
    - Noise variance: σ^2
    - Precision (i.e., inverse variance) hyperparameter of Gaussian prior over $\mathbf{w}$: α

Joint Distribution as Directed Graphical Model

- The N data points are assumed to be i.i.d.
- Joint distribution over random variables:

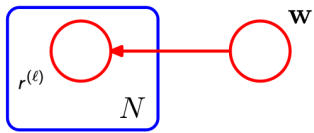
$$p(\mathbf{r}, \mathbf{w}) = \prod_{\ell=1}^{N} p(r^{(\ell)} \mid \mathbf{w}) p(\mathbf{w})$$

- This joint distribution can be represented by a graphical model.



A More Compact Graphical Representation

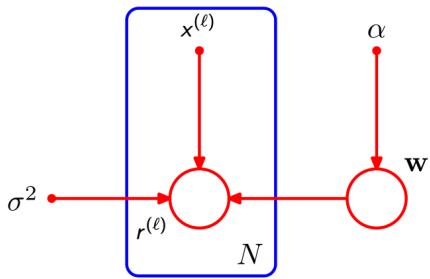
- To represent multiple nodes more compactly, we can draw a single representative node $r^{(\ell)}$ and surround it with a box, called plate, labeled with a number N indicating the number of nodes of this kind.



Explicit Representation of Model Parameters

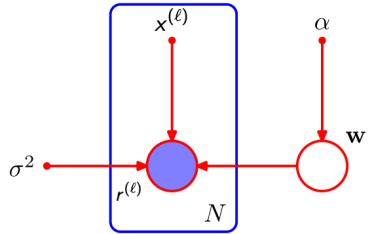
- Joint distribution over random variables with model parameters shown explicitly (deterministic parameters denoted by small solid nodes):

$$p(\mathbf{r}, \mathbf{w} \mid \mathbf{x}, \alpha, \sigma^2) = \prod_{\ell=1}^{N} p(r^{(\ell)} \mid x^{(\ell)}, \mathbf{w}, \sigma^2) p(\mathbf{w} \mid \alpha)$$



Variables and Posterior Distribution

- The observed variables are those random variables set to their observed values, typically shown as shaded nodes.
- The other random variables are referred to as latent variables or hidden variables.



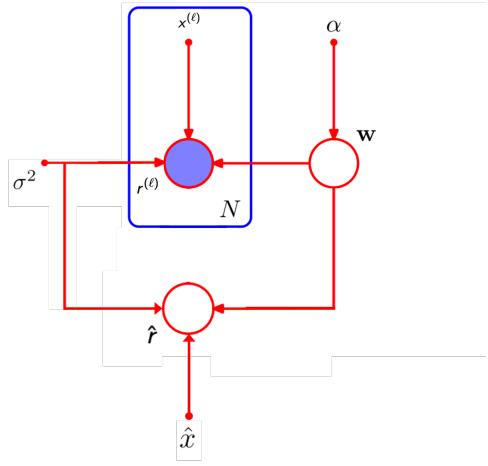
- We can apply Bayes' theorem to evaluate the posterior distribution over the polynomial coefficients $\mathbf{w}$:

$$p(\mathbf{w} \mid \mathbf{r}) = \frac{p(\mathbf{r}, \mathbf{w})}{p(\mathbf{r})} = \frac{\prod_{\ell=1}^{N} p(r^{(\ell)} \mid \mathbf{w}) p(\mathbf{w})}{p(\mathbf{r})}$$

where again we have omitted the deterministic parameters.

Prediction Based on Regression Model

- Given a new input $\hat{x}$, we want to find the corresponding probability distribution for $\hat{r}$ conditioned on the observed data.
- The graphical model that describes this problem is



Graphical Model for Prediction

- Joint distribution over all random variables in this model, conditioned on the deterministic parameters:

$$\begin{aligned} p(\hat{r}, \mathbf{r}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) &= p(\hat{r} \mid \hat{x}, \mathbf{w}, \sigma^2) p(\mathbf{r}, \mathbf{w} \mid \mathbf{x}, \alpha, \sigma^2) \\ &= p(\hat{r} \mid \hat{x}, \mathbf{w}, \sigma^2) \left[\prod_{\ell=1}^{N} p(r^{(\ell)} \mid x^{(\ell)}, \mathbf{w}, \sigma^2) \right] p(\mathbf{w} \mid \alpha) \end{aligned}$$

- Predictive distribution for $\hat{r}$:

$$\begin{aligned} p(\hat{r} \mid \hat{x}, \mathbf{x}, \mathbf{r}, \alpha, \sigma^2) &= \int p(\hat{r}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \mathbf{r}, \alpha, \sigma^2) d\mathbf{w} \\ &= \int p(\hat{r} \mid \hat{x}, \mathbf{w}, \sigma^2) p(\mathbf{w} \mid \mathbf{x}, \mathbf{r}, \alpha, \sigma^2) d\mathbf{w} \\ &= \int p(\hat{r} \mid \hat{x}, \mathbf{w}, \sigma^2) \frac{p(\mathbf{r}, \mathbf{w} \mid \mathbf{x}, \alpha, \sigma^2)}{p(\mathbf{r} \mid \mathbf{x}, \sigma^2)} d\mathbf{w} \propto \int p(\hat{r}, \mathbf{r}, \mathbf{w} \mid \hat{x}, \mathbf{x}, \alpha, \sigma^2) d\mathbf{w} \end{aligned}$$

where the random variables in $\mathbf{r}$ are set to the observed values in the training set.

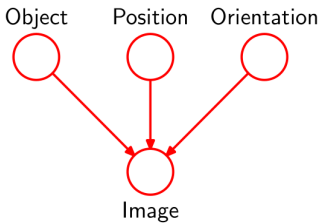
Outline

Ancestral Sampling

- Problem formulation:
  - Given a joint distribution $p(x_1, \ldots, x_K)$ over K variables that factorizes according to a DAG, assuming that the variables have been ordered such that each node has a higher number than any of its parents.
  - The goal is to draw a sample $\hat{x}_1, \ldots, \hat{x}_K$ from the joint distribution.
- Sampling procedure:
  - Start with x_1 and draw a sample from $p(x_1)$, which we call $\hat{x}_1$.
  - For each of the remaining nodes (from $\hat{x}_2$ to $\hat{x}_K$ in order), draw a sample from $p(x_n \mid \mathrm{pa}_n)$ in which the parent variables have been set to their sampled values.
  - In case we want to obtain a sample from some marginal distribution corresponding to a subset of the variables, we simply take the sampled values for the required nodes and ignore the rest.

Non-generative vs. Generative Models

- Non-generative models:
  - Graphical models that cannot be used to generate data.
  - Example: the polynomial regression model discussed above (because there is no probability distribution associated with the input variable x). We could make it generative by introducing a suitable prior distribution $p(x)$.
- Generative models:
  - Graphical models that capture the causal process by which the observed data was generated.
  - Example: the classification model discussed above; model representing the process by which images of objects are created.



Outline

Discrete Variables

- Probability distribution of a single discrete variable $\mathbf{x}$ having K possible states (using the 1-of-K representation) is given by:

$$p(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- Since the parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$ are subject to the constraint

$$\sum_{k=1}^{K} \mu_k = 1$$

only $K - 1$ values for μ_k need to be specified in order to dene the distribution.

Joint Distribution

- Joint distribution of two discrete variables $\mathbf{x}_1$ and $\mathbf{x}_2$, each of which has K states:

$$p(\mathbf{x}_1, \mathbf{x}_2 \mid \boldsymbol{\mu}) = \prod_{k=1}^{K} \prod_{\ell=1}^{K} \mu_{k\ell}^{x_{1k} x_{2\ell}}$$

  where x_{1k} denotes the kth component of $\mathbf{x}_1$ (and similarly for $x_{2\ell}$) and $\mu_{k\ell}$ denotes the probability of observing both $x_{1k} = 1$ and $x_{2\ell} = 1$.

- Since the parameters are subject to the constraint

$$\sum_{k=1}^{K} \sum_{\ell=1}^{K} \mu_{k\ell} = 1$$

  the joint distribution is governed by only $K^2 - 1$ parameters.

- For M discrete variables, the joint distribution is governed by $K^M - 1$ parameters – which grows exponentially with M.

Removing Links I

- The joint distribution $p(\mathbf{x}_1, \mathbf{x}_2)$ can be factorized as $p(\mathbf{x}_2 \mid \mathbf{x}_1)p(\mathbf{x}_1)$, which corresponds to the following graphical model:

- $p(\mathbf{x}_1)$ is governed by $(K-1)$ parameters and $p(\mathbf{x}_2 \mid \mathbf{x}_1)$ is governed by $(K-1)$ parameters for each of the K possible values of $\mathbf{x}_1$. So the total number of parameters that need to be specified is $(K-1) + K(K-1) = K^2 - 1$.
- The total number of parameters should be $2(K-1)$ if $\mathbf{x}_1$ and $\mathbf{x}_2$ were independent corresponding to the following graphical model:

- For M variables, the total number of parameters would be $M(K-1)$, which grows linearly with M. Thus the number of parameters can be reduced by removing links.

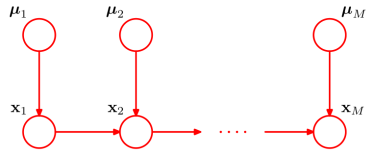
Removing Links II

- If we have M discrete variables with the joint distribution modeled using a fully connected directed graph with one variable corresponding to each node, we have $K^M - 1$ parameters. If there are no links in the graph, the total number of parameters is $M(K-1)$.
- A graphical model having intermediate levels of connectivity is in the form of a chain of M discrete variables:

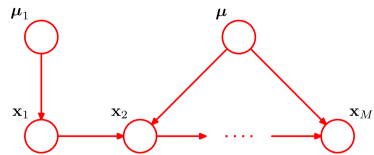
- The marginal distribution $p(\mathbf{x}_1)$ requires $K-1$ parameters, whereas each of the $M-1$ conditional distributions requires $K(K-1)$ parameters. This gives a total parameter count of $K-1+(M-1)K(K-1)$, which is quadratic in K and which grows linearly (rather than exponentially) with the length M of the chain.

Parameter Sharing

- Besides removing some links, another way of reducing the number of independent parameters is through parameter sharing (a.k.a. tying of parameters).
- An extension of the chain model by including priors over the parameters governing the discrete distributions:

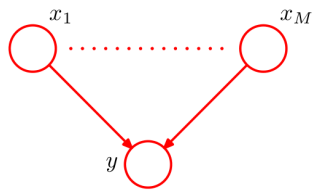


- A simplified model with a single set of parameters $\boldsymbol{\mu}$ shared among all the conditional distributions $p(\mathbf{x}_i \mid \mathbf{x}_{i-1})$:



Parameterized Distributions I

- Yet another way to control the exponential growth in the number of parameters is to use parameterized models for the conditional distributions instead of complete tables of conditional probability values.
- Example:
  - A graph of binary variables comprising M parents $x_1, \ldots, x_M$ and a single child y:



  - Each of the parent variables x_i is governed by a single parameter μ_i representing the probability $p(x_i = 1)$, giving M parameters in total for the parent nodes.

Parameterized Distributions II

- Example (cont'd):
  - The conditional distribution $p(y \mid x_1, \ldots, x_M)$ would require 2^M parameters to represent the probability $p(y = 1)$ for each of the 2^M possible settings of the parent variables.
  - A more parsimonious form for the conditional distribution based on a logistic sigmoid function on a linear combination of the parent variables:

$$p(y = 1 \mid x_1, \ldots, x_M) = \sigma\left(w_0 + \sum_{i=1}^{M} w_i x_i\right) = \sigma(\mathbf{w}^T \mathbf{x})$$

    where $\mathbf{x} = (x_0, x_1, \ldots, x_M)^T$ is an $(M+1)$-dimensional vector of parent states augmented with an additional variable x_0 whose value is clamped to 1, and $\mathbf{w} = (w_0, w_1, \ldots, w_M)^T$ is a vector of $M+1$ parameters.
  - The number of parameters now grows linearly with M, though the conditional distribution is of a more restricted form.

Outline

Introduction

- We express a multivariate Gaussian distribution as a directed graph corresponding to a linear-Gaussian model over the component variables.
- This allows us to impose interesting structure on the distribution, with the general Gaussian and the diagonal covariance Gaussian representing opposite extremes.
- Consider a DAG over D variables in which node i represents a random variable x_i governed by a Gaussian distribution:

$$p(x_i \mid \mathrm{pa}_i) = \mathcal{N}(x_i \mid \mu_i, v_i) \tag{1}$$

where the mean

$$\mu_i = \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i$$

is a linear combination of the states of its parent nodes and v_i is the variance of the conditional distribution.

Joint Distribution I

- Log of the joint distribution of $\mathbf{x} = (x_1, \ldots, x_D)^T$:

$$\begin{aligned}\log p(\mathbf{x}) &= \sum_{i=1}^{D} \log p(x_i \mid \mathrm{pa}_i) \\ &= -D \log v_i - \sum_{i=1}^{D} \frac{1}{2v_i} \left(x_i - \sum_{j \in \mathrm{pa}_i} w_{ij} x_j - b_i \right)^2 + const\end{aligned}$$

  where *const* denotes the terms independent of $\mathbf{x}$.

- Since $\log p(\mathbf{x})$ is a quadratic function of the components of $\mathbf{x}$, $p(\mathbf{x})$ is a multivariate Gaussian distribution.

Joint Distribution II

- Since each variable x_i has (conditional on the states of its parents) a Gaussian distribution of the form (1), we can express x_i as:

$$x_i = \sum_{j \in \mathrm{pa}_i} w_{ij} x_j + b_i + \sqrt{v_i}\epsilon_i \tag{2}$$

where ϵ_i is a zero-mean, unit-variance Gaussian random variable satisfying

$$\mathbb{E}[\epsilon_i] = 0$$
$$\mathbb{E}[\epsilon_i \epsilon_j] = I_{ij}$$

with I_{ij} denoting the (i, j)th element of the identity matrix.

Recursive Computation of Mean and Covariance

- Taking the expectation of (2), we have

$$\mathbb{E}[x_i] = \sum_{j \in \mathrm{pa}_i} w_{ij} \mathbb{E}[x_j] + b_i$$

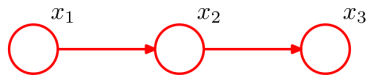
- We can find the components of the mean $\mathbb{E}[\mathbf{x}] = (\mathbb{E}[x_i], \ldots, \mathbb{E}[x_D])^T$ by starting at the lowest-numbered node and working recursively through the graph.
- Similarly, the (i, j)th element $(i \leq j)$ of the covariance matrix for $p(\mathbf{x})$ can be evaluated recursively based on the following recursion relation starting from the lowest numbered node:

$$\begin{aligned}
\mathrm{Cov}[x_i, x_j] =& \mathbb{E}\left[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])\right] = \mathbb{E}\Big[(x_i - \mathbb{E}[x_i])\{\sum_{k \in \mathrm{pa}_j} w_{jk}(x_k - \mathbb{E}[x_k]) + \sqrt{v_j}\epsilon_j\}\Big] \\
=& \mathbb{E}\Big[\sum_{k \in \mathrm{pa}_j} w_{jk}(x_i - \mathbb{E}[x_i])(x_k - \mathbb{E}[x_k])\Big] + \mathbb{E}\Big[(x_i - \mathbb{E}[x_i])\sqrt{v_j}\epsilon_j\Big] \\
=& \sum_{k \in \mathrm{pa}_j} w_{jk} \mathrm{Cov}[x_i, x_k] + I_{ij}\sqrt{v_i}\sqrt{v_j}
\end{aligned}$$

Two Extreme Cases

- Case 1: a graph with no links
  - The parameters w_{ij} are zero.
  - Mean: $[b_1, \ldots, b_D]^T$
  - Covariance matrix: $\text{diag}([v_1, \ldots, v_D]^T)$
  - $p(\mathbf{x})$ has $2D$ parameters and represents a set of D independent univariate Gaussian distributions.
- Case 2: a fully connected graph
  - Each node has all lower-numbered nodes as parents, i.e., the matrix w_{ij} has $i-1$ entries on the ith row and hence is a lower triangular matrix with no entries on the leading diagonal, giving a total of $D(D-1)/2$ parameters.
  - The total number of independent parameters $\{w_{ij}\}$ and $\{v_i\}$ in the covariance matrix is $D(D+1)/2$, corresponding to a general symmetric covariance matrix.
- Graphs having some intermediate level of complexity correspond to joint Gaussian distributions with partially constrained covariance matrices.

An Illustrative Example



- A DAG over three Gaussian variables.
- Mean and covariance matrix of joint distribution:

$$\boldsymbol{\mu} = (b_1, b_2 + w_{21}b_1, b_3 + w_{32}b_2 + w_{32}w_{21}b_1)^T$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} v_1 & w_{21}v_1 & w_{32}w_{21}v_1 \\ w_{21}v_1 & v_2 + w_{21}^2 v_1 & w_{32}(v_2 + w_{21}^2 v_1) \\ w_{32}w_{21}v_1 & w_{32}(v_2 + w_{21}^2 v_1) & v_3 + w_{32}^2(v_2 + w_{21}^2 v_1) \end{pmatrix}$$

Extension to Multivariate Case

- In general, each node of the DAG represents a multivariate Gaussian variable.
- Conditional distribution:

$$p(\mathbf{x}_i \mid \mathrm{pa}_i) = \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

  where

$$\boldsymbol{\mu}_i = \sum_{j \in \mathrm{pa}_i} \mathbf{W}_{ij}\mathbf{x}_j + \mathbf{b}_i$$

- Since $\mathbf{x}_i$ and $\mathbf{x}_j$ may have different dimensionalities, in general $\mathbf{W}_{ij}$ is not a square matrix.
- Similar to the univariate case, it is easy to show that the joint distribution over all variables is Gaussian.