Lecture 2

David A. Levin

University of Oregon

January 2020

Bayesian Estimation

- In this set-up, we assume the parameter(s) θ has a **prior** distribution $\pi(\theta)$.
- Given θ, the random variables $(X_1, \ldots, X_n)$ have a distribution $f(x_1 \ldots, x_n \mid \theta)$.
- If the distributions of X_i's are conditionally independent, then

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

- Inference about θ is made using the **posterior** distribution

$$f(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{\int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\,d\theta}$$

Bayesian Estimation

- In this set-up, we assume the parameter(s) θ has a **prior** distribution $\pi(\theta)$.
- Given θ, the random variables $(X_1, \ldots, X_n)$ have a distribution $f(x_1 \ldots, x_n \mid \theta)$.
- If the distributions of X_i's are conditionally independent, then

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)\,.$$

- Inference about θ is made using the **posterior** distribution

$$f(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{\int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta}$$

Bayesian Estimation

- In this set-up, we assume the parameter(s) θ has a **prior** distribution $\pi(\theta)$.
- Given θ, the random variables $(X_1, \ldots, X_n)$ have a distribution $f(x_1 \ldots, x_n \mid \theta)$.
- If the distributions of X_i's are conditionally independent, then

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) .$$

- Inference about θ is made using the **posterior** distribution

$$f(\theta \mid x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \mid \theta)\pi(\theta)}{\int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta}$$

Bayesian Estimation

- In this set-up, we assume the parameter(s) θ has a **prior** distribution $\pi(\theta)$.
- Given θ, the random variables $(X_1, \ldots, X_n)$ have a distribution $f(x_1 \ldots, x_n \,|\, \theta)$.
- If the distributions of X_i's are conditionally independent, then

$$f(x_1, \ldots, x_n \,|\, \theta) = \prod_{i=1}^{n} f(x_i \,|\, \theta) \,.$$

- Inference about θ is made using the **posterior** distribution

$$f(\theta \,|\, x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n \,|\, \theta)\pi(\theta)}{\int f(x_1, \ldots, x_n \,|\, \theta)\pi(\theta)\, d\theta}$$

- The issue is often computing the normalizing constant in the posterior:

$$\int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta$$

- If θ is high-dimensional, this especially can be difficult. Modern Bayesian statistics uses many methods including Markov Chain Monte Carlo to evaluate this constant.
- Often the prior is chosen so that the posterior is easy to determine; if the posterior has the same parametric form as the prior, the distribution is called **conjugate**

- The issue is often computing the normalizing constant in the posterior:
$$\int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta$$
- If θ is high-dimensional, this especially can be difficult. Modern Bayesian statistics uses many methods including Markov Chain Monte Carlo to evaluate this constant.
- Often the prior is chosen so that the posterior is easy to determine; if the posterior has the same parametric form as the prior, the distribution is called **conjugate**

- The issue is often computing the normalizing constant in the posterior:
$$\int f(x_1, \ldots, x_n \mid \theta)\pi(\theta)\, d\theta$$
- If θ is high-dimensional, this especially can be difficult. Modern Bayesian statistics uses many methods including Markov Chain Monte Carlo to evaluate this constant.
- Often the prior is chosen so that the posterior is easy to determine; if the posterior has the same parametric form as the prior, the distribution is called **conjugate**

Normal Example

- Suppose that
$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\mu-\nu)^2\right)$$
That is $\theta \sim N(\nu, \tau^2)$.
- Suppose that $X_1, \ldots, X_n$ are i.i.d.
$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right),$$
that is, X_i are $N(\mu, \sigma^2)$.
- Here ν, τ^2, σ^2 are assumed known. They are called **hyperparameters**.
- Calculating the posterior seems messy but it all works out because the prior is conjugate.

Normal Example

- Suppose that

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\mu - \nu)^2\right)$$

  That is $\theta \sim N(\nu, \tau^2)$.

- Suppose that $X_1, \ldots, X_n$ are i.i.d.

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right),$$

  that is, X_i are $N(\mu, \sigma^2)$.

- Here ν, τ^2, σ^2 are assumed known. They are called **hyperparameters**.
- Calculating the posterior seems messy but it all works out because the prior is conjugate.

Normal Example

- Suppose that

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\mu - \nu)^2\right)$$

  That is $\theta \sim N(\nu, \tau^2)$.

- Suppose that $X_1, \ldots, X_n$ are i.i.d.

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right),$$

  that is, X_i are $N(\mu, \sigma^2)$.

- Here ν, τ^2, σ^2 are assumed known. They are called **hyperparameters**.
- Calculating the posterior seems messy but it all works out because the prior is conjugate.

Normal Example

- Suppose that

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}(\mu-\nu)^2\right)$$

  That is $\theta \sim N(\nu, \tau^2)$.
- Suppose that $X_1, \ldots, X_n$ are i.i.d.

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right),$$

  that is, X_i are $N(\mu, \sigma^2)$.
- Here ν, τ^2, σ^2 are assumed known. They are called **hyperparameters**.
- Calculating the posterior seems messy but it all works out because the prior is conjugate.

$$
\begin{aligned}
f(\mu \mid x_1,\ldots,x_n) &\propto f(x_1,\ldots,x_n \mid \mu)\pi(\mu;\nu,\tau) \\
&= c(\sigma,\tau)\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{2\tau^2}(\mu-\nu)^2\right)\right) \\
&= c(\sigma,\tau,\boldsymbol{x})\exp\left(\mu\left(\frac{S_n}{2\sigma^2}+\frac{\nu}{2\tau^2}\right)-\mu^2\left(\frac{n}{2\sigma^2}+\frac{1}{2\tau^2}\right)\right) \\
&= c(\sigma,\tau,\boldsymbol{x})\exp\left[-\frac{1}{2}\,\frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}\left(\mu^2-\mu\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\left(\frac{\bar{x}n\tau^2+\nu\sigma^2}{\sigma^2\tau^2}\right)\right)\right] \\
&= c(\sigma,\tau,\boldsymbol{x})\exp\left[-\frac{1}{2v(\sigma,\tau)}\left(\mu-\frac{n\bar{x}\tau^2+\nu\sigma^2}{n\tau^2+\sigma^2}\right)^2\right]
\end{aligned}
$$

All the exponent that does not depend on μ is thrown into the multiplicative constant $c(\sigma,\tau,\boldsymbol{x})$.
But the only distribution this can be is Normal, with variance $v(\sigma,\tau)=\sigma^2\tau^2/(n\tau^2+\sigma^2)$, and with mean

$$
\bar{x}\frac{\tau^2}{\tau^2+\sigma^2/n}+\nu\frac{\sigma^2/n}{\tau^2+\sigma^2/n}.
$$

- If the goal is minimize

$$\mathbb{E}[(\mu - T)^2 \mid \boldsymbol{x}]$$

  among statistics T depending on $\boldsymbol{x}$, then the minimizer is $T = \mathbb{E}[\mu \mid \boldsymbol{x}]$.
- In this case, the Bayes estimator is

$$\bar{x}\frac{\tau^2}{\tau^2 + \sigma^2/n} + \nu\frac{\sigma^2/n}{\tau^2 + \sigma^2/n}$$

- This is a convex combination of the data-only estimator $\bar{X}$ and the prior mean ν. The weight of $\bar{X}$ tend to 1 as $n \to \infty$.
- Other inferences are possible, e.g. *credible intervals*, so we can find a and b so that

$$\mathbb{P}(a < \mu < b \mid \boldsymbol{x}) = 0.95\,.$$

- If the goal is minimize

$$\mathbb{E}[(\mu - T)^2 \mid \boldsymbol{x}]$$

  among statistics T depending on $\boldsymbol{x}$, then the minimizer is $T = \mathbb{E}[\mu \mid \boldsymbol{x}]$.
- In this case, the Bayes estimator is

$$\bar{x}\frac{\tau^2}{\tau^2 + \sigma^2/n} + \nu\frac{\sigma^2/n}{\tau^2 + \sigma^2/n}$$

- This is a convex combination of the data-only estimator $\bar{X}$ and the prior mean ν. The weight of $\bar{X}$ tend to 1 as $n \to \infty$.
- Other inferences are possible, e.g. *credible intervals*, so we can find a and b so that

$$\mathbb{P}(a < \mu < b \mid \boldsymbol{x}) = 0.95\,.$$

- If the goal is minimize

$$\mathbb{E}[(\mu - T)^2 \mid \boldsymbol{x}]$$

among statistics T depending on $\boldsymbol{x}$, then the minimizer is $T = \mathbb{E}[\mu \mid \boldsymbol{x}]$.

- In this case, the Bayes estimator is

$$\bar{x}\frac{\tau^2}{\tau^2 + \sigma^2/n} + \nu\frac{\sigma^2/n}{\tau^2 + \sigma^2/n}$$

- This is a convex combination of the data-only estimator $\bar{X}$ and the prior mean ν. The weight of $\bar{X}$ tend to 1 as $n \to \infty$.
- Other inferences are possible, e.g. *credible intervals*, so we can find a and b so that

$$\mathbb{P}(a < \mu < b \mid \boldsymbol{x}) = 0.95\,.$$

- If the goal is minimize

$$\mathbb{E}[(\mu - T)^2 \mid \boldsymbol{x}]$$

  among statistics T depending on $\boldsymbol{x}$, then the minimizer is $T = \mathbb{E}[\mu \mid \boldsymbol{x}]$.

- In this case, the Bayes estimator is

$$\bar{x}\frac{\tau^2}{\tau^2 + \sigma^2/n} + \nu\frac{\sigma^2/n}{\tau^2 + \sigma^2/n}$$

- This is a convex combination of the data-only estimator $\bar{X}$ and the prior mean ν. The weight of $\bar{X}$ tend to 1 as $n \to \infty$.
- Other inferences are possible, e.g. *credible intervals*, so we can find a and b so that

$$\mathbb{P}(a < \mu < b \mid \boldsymbol{x}) = 0.95\,.$$

Graphical Models

- A **graphical model** specifies a factorization which a joint density must obey.
- Example: If $X_1, \ldots, X_n$ are independent, then the joint density completely factors into marginals:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) . \tag{1}$$

- This is too strong a constraint to require of general applicable probability models.
- A graphical model specifies a class of joint densities which obey a factorization which still affords useful simplification but preserves some generality.

Graphical Models

- A **graphical model** specifies a factorization which a joint density must obey.
- Example: If $X_1, \ldots, X_n$ are independent, then the joint density completely factors into marginals:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)\,. \tag{1}$$

- This is too strong a constraint to require of general applicable probability models.
- A graphical model specifies a class of joint densities which obey a factorization which still affords useful simplification but preserves some generality.

Graphical Models

- A **graphical model** specifies a factorization which a joint density must obey.
- Example: If $X_1, \ldots, X_n$ are independent, then the joint density completely factors into marginals:

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)\,. \tag{1}$$

- This is too strong a constraint to require of general applicable probability models.
- A graphical model specifies a class of joint densities which obey a factorization which still affords useful simplification but preserves some generality.

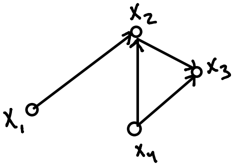
Graphical Models

- A **graphical model** specifies a factorization which a joint density must obey.
- Example: If $X_1, \dots, X_n$ are independent, then the joint density completely factors into marginals:

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)\,. \tag{1}$$

- This is too strong a constraint to require of general applicable probability models.
- A graphical model specifies a class of joint densities which obey a factorization which still affords useful simplification but preserves some generality.

- nodes represent variables
- directed arrows represent dependencies
- Any variable with an arrow pointing to x_i is called a **parent** of x_i; we denote by pa_i all the parents of x_i.



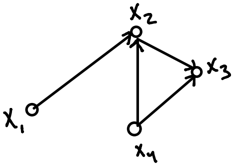
- A joint distribution respects the graphical model encoded by a directed graph if

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \mathrm{pa}_i) \,.$$

- In the example graph, a joint distribution obeys the graphical model if

$$p(x_1, x_2, x_3, x_4) = p(x_1)\, p(x_2 \mid x_1, x_4)\, p(x_3 \mid x_2, x_4)\, p(x_4) \,.$$

- nodes represent variables
- directed arrows represent dependencies
- Any variable with an arrow pointing to x_i is called a **parent** of x_i; we denote by pa_i all the parents of x_i.



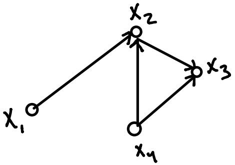
- A joint distribution respects the graphical model encoded by a directed graph if

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \mathsf{pa}_i) \,.$$

- In the example graph, a joint distribution obeys the graphical model if

$$p(x_1, x_2, x_3, x_4) = p(x_1)\, p(x_2 \mid x_1, x_4)\, p(x_3 \mid x_2, x_4)\, p(x_4) \,.$$

- nodes represent variables
- directed arrows represent dependencies
- Any variable with an arrow pointing to x_i is called a **parent** of x_i; we denote by pa_i all the parents of x_i.



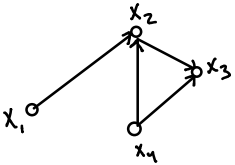
- A joint distribution respects the graphical model encoded by a directed graph if

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \mathsf{pa}_i)\,.$$

- In the example graph, a joint distribution obeys the graphical model if

$$p(x_1, x_2, x_3, x_4) = p(x_1)\, p(x_2 \mid x_1, x_4)\, p(x_3 \mid x_2, x_4)\, p(x_4)\,.$$

- nodes represent variables
- directed arrows represent dependencies
- Any variable with an arrow pointing to x_i is called a **parent** of x_i; we denote by pa_i all the parents of x_i.



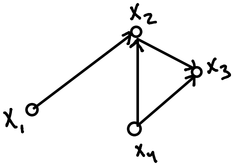
- A joint distribution respects the graphical model encoded by a directed graph if

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \mathsf{pa}_i) \,.$$

- In the example graph, a joint distribution obeys the graphical model if

$$p(x_1, x_2, x_3, x_4) = p(x_1)\, p(x_2 \mid x_1, x_4)\, p(x_3 \mid x_2, x_4)\, p(x_4) \,.$$

- nodes represent variables
- directed arrows represent dependencies
- Any variable with an arrow pointing to x_i is called a **parent** of x_i; we denote by pa_i all the parents of x_i.



- A joint distribution respects the graphical model encoded by a directed graph if

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \mathsf{pa}_i)\,.$$

- In the example graph, a joint distribution obeys the graphical model if

$$p(x_1, x_2, x_3, x_4) = p(x_1)\,p(x_2 \mid x_1, x_4)\,p(x_3 \mid x_2, x_4)\,p(x_4)\,.$$

- The graphical model encodes *conditional independence* statements.
- A set of variables X and Y are conditionally independent given Z if

$$\mathbb{P}(X \in A, Y \in B \mid Z) = \mathbb{P}(X \in A \mid Z)\mathbb{P}(X \in B \mid Z) .$$

- In terms of pdfs/pmfs,

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z) .$$

- Or equivalently,

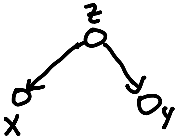
$$p(x \mid y, z) = p(x \mid z) .$$

- The graphical model encodes *conditional independence* statements.
- A set of variables X and Y are conditionally independent given Z if
$$\mathbb{P}(X \in A, Y \in B \mid Z) = \mathbb{P}(X \in A \mid Z)\mathbb{P}(X \in B \mid Z)\,.$$
- In terms of pdfs/pmfs,
$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z)\,.$$
- Or equivalently,
$$p(x \mid y, z) = p(x \mid z)\,.$$

- The graphical model encodes *conditional independence* statements.
- A set of variables X and Y are conditionally independent given Z if
$$\mathbb{P}(X \in A, Y \in B \mid Z) = \mathbb{P}(X \in A \mid Z)\mathbb{P}(X \in B \mid Z)\,.$$
- In terms of pdfs/pmfs,
$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z)\,.$$
- Or equivalently,
$$p(x \mid y, z) = p(x \mid z)\,.$$

- The graphical model encodes *conditional independence* statements.
- A set of variables X and Y are conditionally independent given Z if
$$\mathbb{P}(X \in A, Y \in B \mid Z) = \mathbb{P}(X \in A \mid Z)\mathbb{P}(X \in B \mid Z)\,.$$
- In terms of pdfs/pmfs,
$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z)\,.$$
- Or equivalently,
$$p(x \mid y, z) = p(x \mid z)\,.$$

Example 1

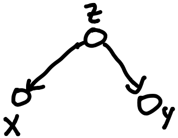


-
- Note that

$$p(x, y \mid z) = \frac{p(x, y, z)}{\sum_{x,y} p(x, y, z)} = \frac{p(x \mid z)\, p(y \mid z)\, p(z)}{\sum_x p(x \mid z) \sum_y p(y \mid z)\, p(z)} = p(x \mid z)\, p(y \mid z)$$

- Any time X and Y are separated by a **tail-to-tail** vertex Z, they are conditionally independent.
- Note that X and Y are not independent unconditionally, in general.
- Exercise: Show by example that X and Y are not necessarily independent.

Example 1



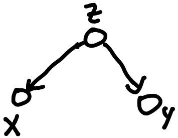
-

- Note that

$$p(x, y \mid z) = \frac{p(x, y, z)}{\sum_{x,y} p(x, y, z)} = \frac{p(x \mid z)\, p(y \mid z)\, p(z)}{\sum_x p(x \mid z) \sum_y p(y \mid z)\, p(z)} = p(x \mid z)\, p(y \mid z)$$

- Any time X and Y are separated by a **tail-to-tail** vertex Z, they are conditionally independent.
- Note that X and Y are not independent unconditionally, in general.
- Exercise: Show by example that X and Y are not necessarily independent.

Example 1



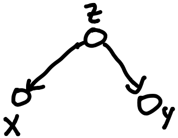
-

- Note that

$$p(x, y \mid z) = \frac{p(x, y, z)}{\sum_{x,y} p(x, y, z)} = \frac{p(x \mid z)\, p(y \mid z)\, p(z)}{\sum_x p(x \mid z) \sum_y p(y \mid z)\, p(z)} = p(x \mid z)\, p(y \mid z)$$

- Any time X and Y are separated by a **tail-to-tail** vertex Z, they are conditionally independent.
- Note that X and Y are not independent unconditionally, in general.
- Exercise: Show by example that X and Y are not necessarily independent.

Example 1

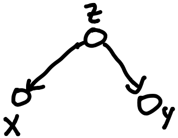


- Note that

$$p(x, y \mid z) = \frac{p(x, y, z)}{\sum_{x,y} p(x, y, z)} = \frac{p(x \mid z)\, p(y \mid z)\, p(z)}{\sum_x p(x \mid z) \sum_y p(y \mid z)\, p(z)} = p(x \mid z)\, p(y \mid z)$$

- Any time X and Y are separated by a **tail-to-tail** vertex Z, they are conditionally independent.
- Note that X and Y are not independent unconditionally, in general.
- Exercise: Show by example that X and Y are not necessarily independent.

Example 1

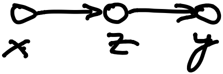


-
- Note that

$$p(x, y \mid z) = \frac{p(x, y, z)}{\sum_{x,y} p(x, y, z)} = \frac{p(x \mid z)\, p(y \mid z)\, p(z)}{\sum_x p(x \mid z) \sum_y p(y \mid z)\, p(z)} = p(x \mid z)\, p(y \mid z)$$

- Any time X and Y are separated by a **tail-to-tail** vertex Z, they are conditionally independent.
- Note that X and Y are not independent unconditionally, in general.
- Exercise: Show by example that X and Y are not necessarily independent.

Example 2

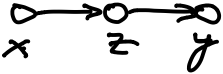


-
-

$$p(x, y \mid z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)\, p(z \mid x)\, p(y \mid z)}{p(z)}$$
$$= \frac{p(x \mid z)\, p(z)\, p(y \mid z)}{p(z)} = p(x \mid z)\, p(y \mid z)$$

- Thus, if x and y are connected by a **head-to-tail** vertex z, then they are conditionally independent.

Example 2

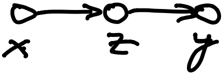


-
-

$$p(x, y \mid z) = \frac{p(x, y, z)}{p(z)} = \frac{p(x)\, p(z \mid x)\, p(y \mid z)}{p(z)}$$
$$= \frac{p(x \mid z)\, p(z)\, p(y \mid z)}{p(z)} = p(x \mid z)\, p(y \mid z)$$

- Thus, if x and y are connected by a **head-to-tail** vertex z, then they are conditionally independent.

Example 2



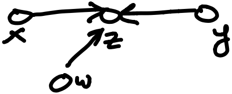
-
-

$$
\begin{aligned}
p(x, y \mid z) = \frac{p(x, y, z)}{p(z)} &= \frac{p(x)\, p(z \mid x)\, p(y \mid z)}{p(z)} \\
&= \frac{p(x \mid z)\, p(z)\, p(y \mid z)}{p(z)} = p(x \mid z)\, p(y \mid z)
\end{aligned}
$$

- Thus, if x and y are connected by a **head-to-tail** vertex z, then they are conditionally independent.

-
- $$p(x, y \mid z) = \frac{p(y \mid z)\, p(x \mid z)\, p(z)}{p(z)} = p(y \mid z)\, p(x \mid z)$$

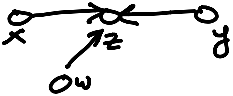
  so X and Y are independent given $\emptyset$.
- Exercise: Give an example to show that X and Y are not conditionally independent given Z.
- Even more is true:



  X and Y are not independent given W (in general).

-
- $$p(x,y \mid z) = \frac{p(y \mid z)\, p(x \mid z)\, p(z)}{p(z)} = p(y \mid z)\, p(x \mid z)$$

  so X and Y are independent given $\emptyset$.
- Exercise: Give an example to show that X and Y are not conditionally independent given Z.
- Even more is true:

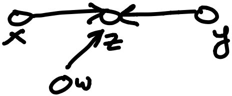


X and Y are not independent given W (in general).

-
-

$$p(x,y\mid z) = \frac{p(y\mid z)\,p(x\mid z)\,p(z)}{p(z)} = p(y\mid z)\,p(x\mid z)$$

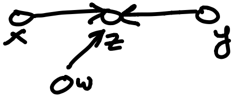
  so X and Y are independent given $\emptyset$.
- Exercise: Give an example to show that X and Y are not conditionally independent given Z.
- Even more is true:



X and Y are not independent given W (in general).

-
- $$p(x,y\mid z) = \frac{p(y\mid z)\,p(x\mid z)\,p(z)}{p(z)} = p(y\mid z)\,p(x\mid z)$$

  so X and Y are independent given $\emptyset$.
- Exercise: Give an example to show that X and Y are not conditionally independent given Z.
- Even more is true:



X and Y are not independent given W (in general).

- Let A, B, C be sets of vertices. The set C **blocks** A to B if every path from a vertex $a \in A$ to a vertex $b \in B$ contains either
  - a vertex $c \in C$ so that c is head-to-tail or tail-to-tail, or
  - a vertex v which neither belongs to C or is a decendent of any vertex in C.
-

Theorem 1 (*d*-separation).

If C blocks A and B in the di-graph, then the variables A and the variables B are conditionally independent given C.

- Let A, B, C be sets of vertices. The set C **blocks** A to B if every path from a vertex $a \in A$ to a vertex $b \in B$ contains either
    - a vertex $c \in C$ so that c is head-to-tail or tail-to-tail, or
    - a vertex v which neither belongs to C or is a decendent of any vertex in C.
-

Theorem 1 (*d*-separation).

If C blocks A and B in the di-graph, then the variables A and the variables B are conditionally independent given C.

- Let A, B, C be sets of vertices. The set C **blocks** A to B if every path from a vertex $a \in A$ to a vertex $b \in B$ contains either
  - a vertex $c \in C$ so that c is head-to-tail or tail-to-tail, or
  - a vertex v which neither belongs to C or is a decendent of any vertex in C.
-

Theorem 1 (d-separation).

If C blocks A and B in the di-graph, then the variables A and the variables B are conditionally independent given C.

- Let A, B, C be sets of vertices. The set C **blocks** A to B if every path from a vertex $a \in A$ to a vertex $b \in B$ contains either
    - a vertex $c \in C$ so that c is head-to-tail or tail-to-tail, or
    - a vertex v which neither belongs to C or is a decendent of any vertex in C.
-

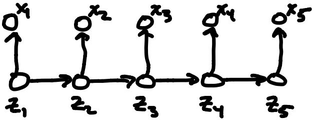
Theorem 1 (d-separation).

If C blocks A and B in the di-graph, then the variables A and the variables B are conditionally independent given C.

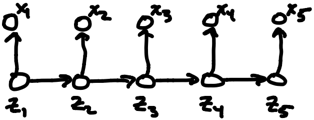
- Let A, B, C be sets of vertices. The set C **blocks** A to B if every path from a vertex $a \in A$ to a vertex $b \in B$ contains either
  - a vertex $c \in C$ so that c is head-to-tail or tail-to-tail, or
  - a vertex v which neither belongs to C or is a decendent of any vertex in C.
-

Theorem 1 (d-separation).

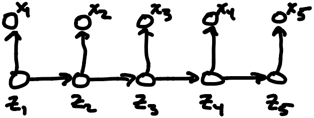
If C blocks A and B in the di-graph, then the variables A and the variables B are conditionally independent given C.



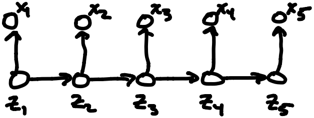
- The above is a **Hidden Markov Model**.
- We can use d-separation to establish useful conditional independence relations:
  - $$(X_1, \ldots, X_k) \perp (X_{k+1}, \ldots, X_n) \mid Z_k$$
  - $$X_1, \ldots, X_{k-1} \perp X_k \mid Z_k$$



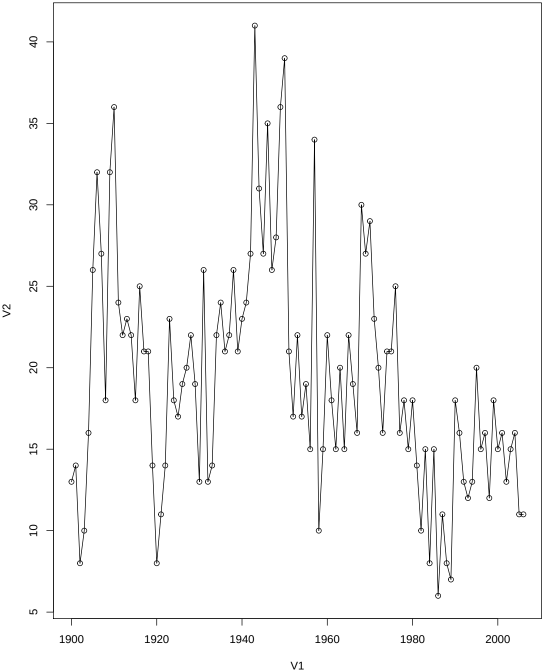
- The above is a **Hidden Markov Model**.
- We can use *d*-separation to establish useful conditional independence relations:
  - $$(X_1, \ldots, X_k) \perp (X_{k+1}, \ldots, X_n) \mid Z_k$$
  - $$X_1, \ldots, X_{k-1} \perp X_k \mid Z_k$$



- The above is a **Hidden Markov Model**.
- We can use d-separation to establish useful conditional independence relations:
  - $$(X_1, \dots, X_k) \perp (X_{k+1}, \dots, X_n) \mid Z_k$$
  - $$X_1, \dots, X_{k-1} \perp X_k \mid Z_k$$

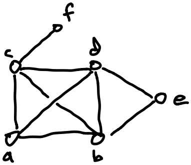


- The above is a **Hidden Markov Model**.
- We can use *d*-separation to establish useful conditional independence relations:
  - $$(X_1, \ldots, X_k) \perp (X_{k+1}, \ldots, X_n) \mid Z_k$$
  - $$X_1, \ldots, X_{k-1} \perp X_k \mid Z_k$$


V2
V1
5
10
15
20
25
30
35
40
1900
1920
1940
1960
1980
2000

- Can we do something similar, but for undirected graphs? That is, can we use undirected graphs to specify classes of joint probability distributions, where variables are identified with vertices?
- A **Markov random field** satisfies a simpler version of the d-separation theorem: A and B are blocked by C if every path from A to B includes a vertex in C.
- We will first specify a structure for joint distributions whose components are identified with vertices of the graph.
- We will state a theorem which shows the equivalence of Markov random fields with the distributions having the specified product structure.

- A **clique** in a graph is a set of vertices that are maximally connected.
- A **maximal** clique in a graph cannot be enlarged by adding another vertex to obtain another clique.
- Below $\{a, b, c\}$ is a clique, but not maximal, because $\{a, b, c, d\}$ is a clique containing it. $\{a, b, c, d\}$ is maximal, because including e or f does not create a clique.



- To each maximal clique S in a graph, associate a **potential**, defined as $e^{-H_S(x_S)}$.
- Define a probability by

$$p(\boldsymbol{x}) = Z^{-1} \prod_S e^{H_S(x_S)} = Z^{-1} e^{-\sum_S H_S(x_S)}.$$

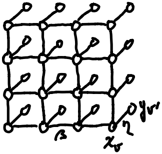
Theorem 2 (Hammersley-Clifford).

A joint distribution is a Markov random field if and only if it has the product form

$$p(\boldsymbol{x}) = Z^{-1} \prod_S e^{-H_S(x_S)}$$

- The product is over maximal cliques S;
- $x_S = (x_{s_1}, \ldots, x_{s_r})$ where $S = \{s_1, \ldots, s_r\}$;
- $Z = \sum_{\boldsymbol{x}} \prod_S e^{-H_S(x_S)}$ is a normalizing constant; computing Z can be quite difficult or expensive!

Ising Model



$$H_{\{v,w\}}(x_v, x_w) = -\beta x_v x_w$$

If $x_v = x_w$, $H(x_v, x_w) = -\beta$
$x_v \neq x_w$, $H(x_v, x_w) = \beta$

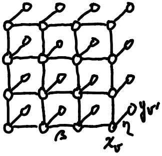
"Agreement is low energy"

$$H(x_v, y_{v'}) = -\eta x_v y_{v'}$$

- Each variable is ± 1;
- β is parameter (inverse temperature) representing interaction strength between variables in "base graph";
- Y_v is a "noisy" observation of X_v; the parameter η determines the "flip" probability.
- In image analysis applications, the Y's are observed, and the X's are unobserved.
- Total energy is

$$H(\boldsymbol{x}, \boldsymbol{y}) = -\beta \sum_{v \sim w} x_v x_w + h \sum_v x_v - \eta \sum_v x_v y_{v'}$$

Ising Model



$H_{\{v,w\}}(x_v, x_w) = -\beta x_v x_w$

If $x_v = x_w$, $H(x_v, x_w) = -\beta$
$x_v \neq x_w$, $H(x_v, x_w) = \beta$

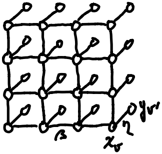
"Agreement is low energy"

$H(x_v, y_{v'}) = -\eta x_v y_{v'}$

- Each variable is ± 1;
- β is parameter (inverse temperature) representing interaction strength between variables in "base graph";
- Y_v is a "noisy" observation of X_v; the parameter η determines the "flip" probability.
- In image analysis applications, the Y's are observed, and the X's are unobserved.
- Total energy is

$$H(\boldsymbol{x}, \boldsymbol{y}) = -\beta \sum_{v \sim w} x_v x_w + h \sum_v x_v - \eta \sum_v x_v y_{v'}$$

Ising Model



$H_{\{v,w\}}(x_v, x_w) = -\beta x_v x_w$

If $x_v = x_w$, $H(x_v, x_w) = -\beta$
$x_v \neq x_w$, $H(x_v, x_w) = \beta$

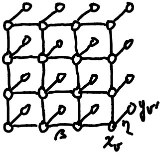
"Agreement is low energy"

$H(x_v, y_{v'}) = -\eta x_v y_{v'}$

- Each variable is ± 1;
- β is parameter (inverse temperature) representing interaction strength between variables in "base graph";
- Y_v is a "noisy" observation of X_v; the parameter η determines the "flip" probability.
- In image analysis applications, the Y's are observed, and the X's are unobserved.
- Total energy is

$$H(\boldsymbol{x}, \boldsymbol{y}) = -\beta \sum_{v \sim w} x_v x_w + h \sum_v x_v - \eta \sum_v x_v y_{v'}$$

Ising Model



$$H_{\{v,w\}}(x_v, x_w) = -\beta x_v x_w$$

If $x_v = x_w$, $H(x_v, x_w) = -\beta$
$x_v \neq x_w$, $H(x_v, x_w) = \beta$

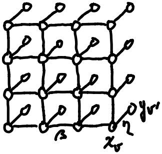
"Agreement is low energy"

$$H(x_v, y_{v'}) = -\eta x_v y_{v'}$$

- Each variable is ± 1;
- β is parameter (inverse temperature) representing interaction strength between variables in "base graph";
- Y_v is a "noisy" observation of X_v; the parameter η determines the "flip" probability.
- In image analysis applications, the Y's are observed, and the X's are unobserved.
- Total energy is

$$H(\mathbf{x}, \mathbf{y}) = -\beta \sum_{v \sim w} x_v x_w + h \sum_v x_v - \eta \sum_v x_v y_{v'}$$

Ising Model



$H_{\{v,w\}}(x_v, x_w) = -\beta x_v x_w$

If $x_v = x_w$, $H(x_v, x_w) = -\beta$
$x_v \neq x_w$, $H(x_v, x_w) = \beta$

"Agreement is low energy"

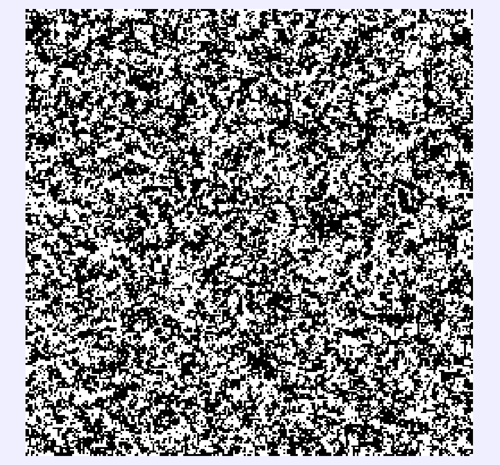
$H(x_v, y_{v'}) = -\eta x_v y_{v'}$

- Each variable is ± 1;
- β is parameter (inverse temperature) representing interaction strength between variables in "base graph";
- Y_v is a "noisy" observation of X_v; the parameter η determines the "flip" probability.
- In image analysis applications, the Y's are observed, and the X's are unobserved.
- Total energy is

$$H(\boldsymbol{x}, \boldsymbol{y}) = -\beta \sum_{v \sim w} x_v x_w + h \sum_v x_v - \eta \sum_v x_v y_{v'}$$

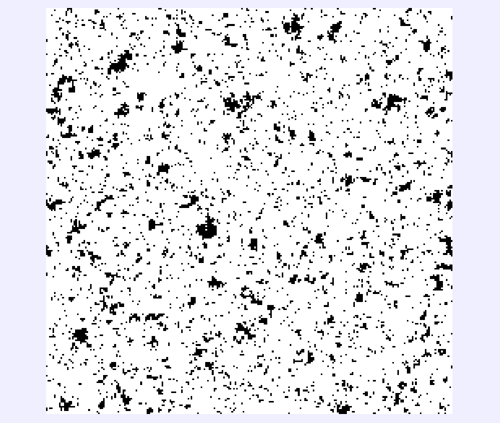
Three regimes

High temperature ($\beta < \beta_c$):



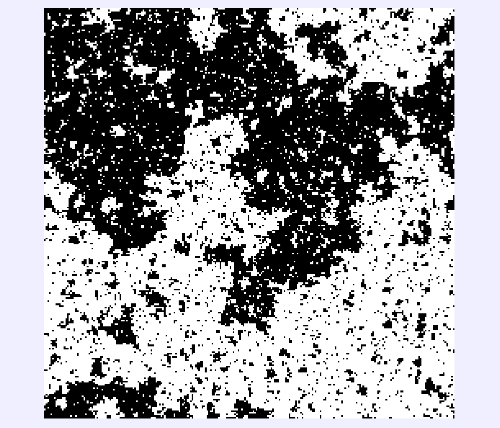
Three regimes

low temperature ($\beta > \beta_c$),



Three regimes

critical temperature ($\beta = \beta_c$),



A simple inference method

- Problem: Given observations y, can we infer x?
- We take a Maximum Likelihood approach; regard the x as parameters, and pick the value x which makes the data y most likely. That is, maximize $p(\boldsymbol{x}, \boldsymbol{y})$ over all $\boldsymbol{x}$, holding the observed $\boldsymbol{y}$ fixed.
- This is a difficult task! Search space is large!
- Iterated Conditional Modes: in some order, move through all nodes. At each node, flip the sign to obtain a higher probability configuration.

A simple inference method

- Problem: Given observations y, can we infer x?
- We take a Maximum Likelihood approach; regard the x as parameters, and pick the value x which makes the data y most likely. That is, maximize $p(\boldsymbol{x}, \boldsymbol{y})$ over all $\boldsymbol{x}$, holding the observed $\boldsymbol{y}$ fixed.
- This is a difficult task! Search space is large!
- Iterated Conditional Modes: in some order, move through all nodes. At each node, flip the sign to obtain a higher probability configuration.

A simple inference method

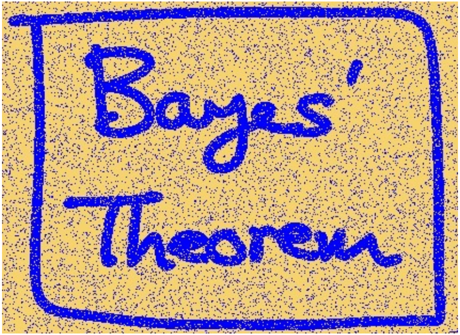
- Problem: Given observations y, can we infer x?
- We take a Maximum Likelihood approach; regard the x as parameters, and pick the value x which makes the data y most likely. That is, maximize $p(\boldsymbol{x}, \boldsymbol{y})$ over all $\boldsymbol{x}$, holding the observed $\boldsymbol{y}$ fixed.
- This is a difficult task! Search space is large!
- Iterated Conditional Modes: in some order, move through all nodes. At each node, flip the sign to obtain a higher probability configuration.

A simple inference method

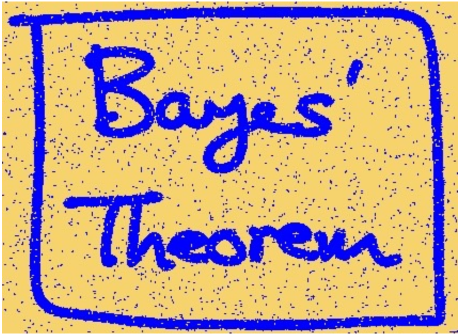
- Problem: Given observations y, can we infer x?
- We take a Maximum Likelihood approach; regard the x as parameters, and pick the value x which makes the data y most likely. That is, maximize $p(\boldsymbol{x}, \boldsymbol{y})$ over all $\boldsymbol{x}$, holding the observed $\boldsymbol{y}$ fixed.
- This is a difficult task! Search space is large!
- Iterated Conditional Modes: in some order, move through all nodes. At each node, flip the sign to obtain a higher probability configuration.

Example from C. Bishop, *Machine Learning and Pattern Recognition* :
Original image

Example from C. Bishop, *Machine Learning and Pattern Recognition* :
Corrupted image



Example from C. Bishop, *Machine Learning and Pattern Recognition* :
Restored image



Message passing and the sum-product algorithm

- In many algorithms we will need an effective way to compute marginals in a graphical model:

$$p(x_k) = \sum_{x_j : j \neq k} p(x_1, \ldots, x_n) \,.$$

- Note if each variable has say K possible values then this requires a sum over K^{n-1} terms; this grows exponentially as n grows, so does not scale.
- In models which have a tree-structure, the sum-product algorithm scales linearly in the size of the problem.

Message passing and the sum-product algorithm

- In many algorithms we will need an effective way to compute marginals in a graphical model:

$$p(x_k) = \sum_{x_j : j \neq k} p(x_1, \ldots, x_n)\,.$$

- Note if each variable has say K possible values then this requires a sum over K^{n-1} terms; this grows exponentially as n grows, so does not scale.
- In models which have a tree-structure, the sum-product algorithm scales linearly in the size of the problem.

Message passing and the sum-product algorithm

- In many algorithms we will need an effective way to compute marginals in a graphical model:

$$p(x_k) = \sum_{x_j : j \neq k} p(x_1, \ldots, x_n) .$$

- Note if each variable has say K possible values then this requires a sum over K^{n-1} terms; this grows exponentially as n grows, so does not scale.
- In models which have a tree-structure, the sum-product algorithm scales linearly in the size of the problem.

factor graphs

1. Factor graphs unite graphical models and Markov random fields together.
2. A factor graph is bipartite, and each node is either a function, or a variable.
3. The joint distribution is a product over the functions appearing in the graph.

factor graphs

1. Factor graphs unite graphical models and Markov random fields together.
2. A factor graph is bipartite, and each node is either a function, or a variable.
3. The joint distribution is a product over the functions appearing in the graph.

factor graphs

1. Factor graphs unite graphical models and Markov random fields together.
2. A factor graph is bipartite, and each node is either a function, or a variable.
3. The joint distribution is a product over the functions appearing in the graph.