Chapter 18
Learning: The Bayesian Approach

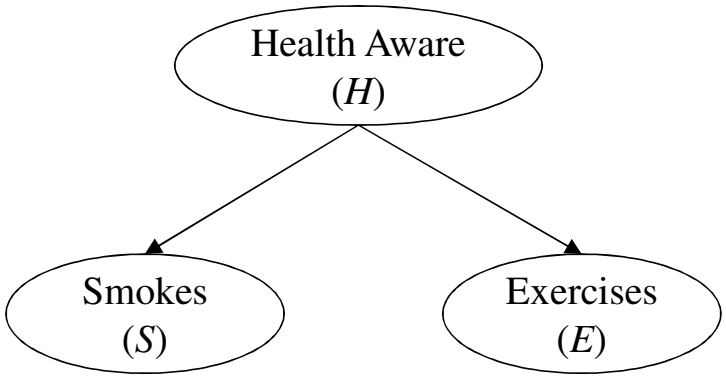
Adnan Darwiche[1]

[1]Lecture slides for *Modeling and Reasoning with Bayesian Networks,* Adnan Darwiche, Cambridge University Press, 2009.

Outline

We will discuss in this chapter a particular approach to learning Bayesian networks from data, known as the Bayesian approach, which is marked by its ability to integrate prior knowledge into the learning process, and to reduce learning to a problem of inference.

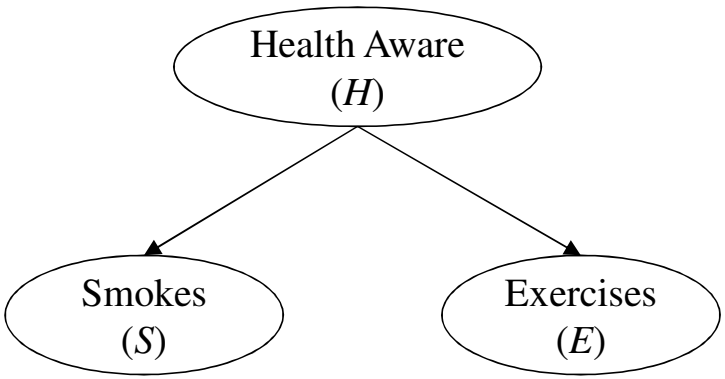
Introduction



Case	H	S	E
1	F	F	T
2	T	F	T
3	T	F	T
4	F	F	F
5	F	T	F

A network structure with a complete data set.

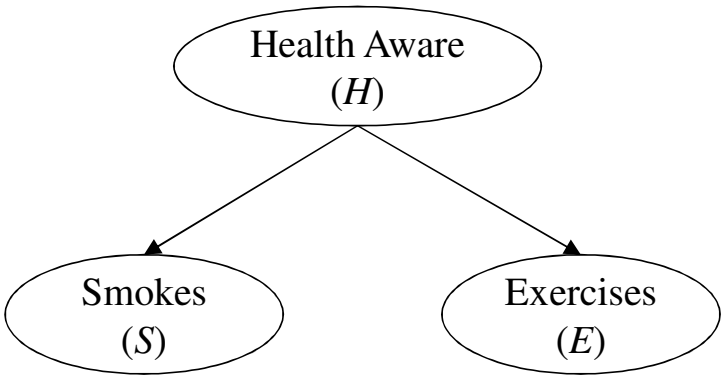
Introduction



Five parameter sets

$$\theta_H = (\theta_h, \theta_{\bar{h}})$$
$$\theta_{S|h} = (\theta_{s|h}, \theta_{\bar{s}|h})$$
$$\theta_{S|\bar{h}} = (\theta_{s|\bar{h}}, \theta_{\bar{s}|\bar{h}})$$
$$\theta_{E|h} = (\theta_{e|h}, \theta_{\bar{e}|h})$$
$$\theta_{E|\bar{h}} = (\theta_{e|\bar{h}}, \theta_{\bar{e}|\bar{h}})$$

Introduction

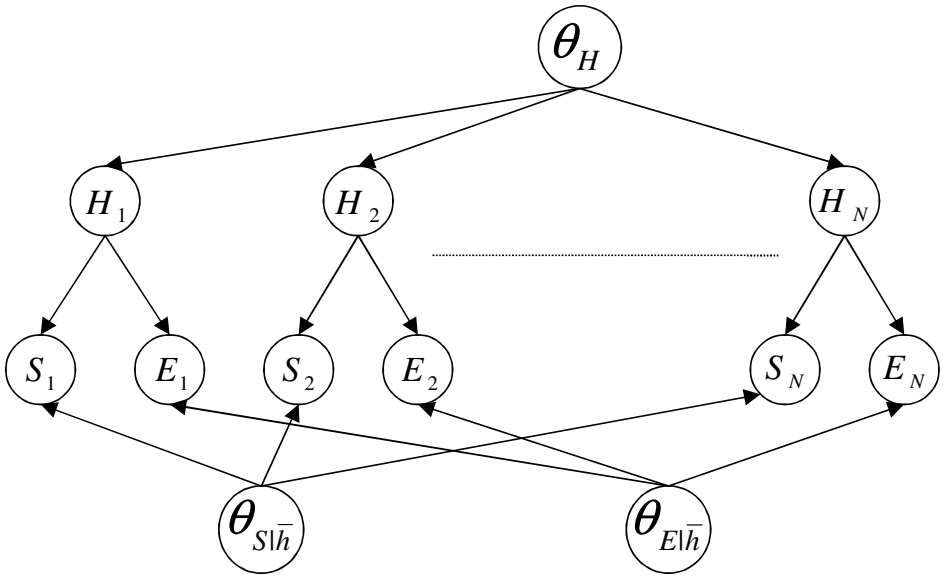


Prior knowledge

$$
\begin{aligned}
\theta_{S|h} &= (.1, .9) \\
\theta_{E|h} &= (.8, .2) \\
\theta_{H} &\in \{(.75, .25), (.90, .10)\} \\
\theta_{S|\bar{h}} &\in \{(.25, .75), (.50, .50)\} \\
\theta_{E|\bar{h}} &\in \{(.50, .50), (.75, .25)\}
\end{aligned}
$$

where each of the two values are considered equally likely.

Introduction



Meta network: variables θ_H, $\theta_{S|\bar{h}}$, $\theta_{E|\bar{h}}$ represent the possible values of unknown network parameters, where the CPTs of these variables encode our prior knowledge about these parameters. Moreover, variables H_i, S_i and E_i represent the values that variables H, S and E take in case i of the data set, allowing one to assert the data set as evidence on the given network.

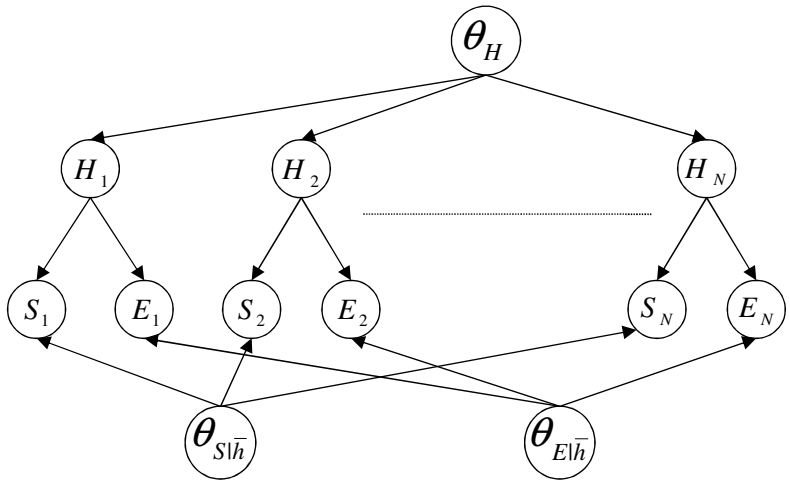
Introduction

By explicitly encoding prior knowledge about network parameters, and by treating data as evidence, the Bayesian approach can now reduce the process of learning to a process of computing posterior distributions:

$$\mathbb{P}(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}|\mathcal{D}),$$

where $\mathbb{P}$ is the distribution induced by the meta network, and $\mathcal{D}$ is the evidence entailed by the data set.

Introduction



One can identify parameter estimates that have the highest probability:

$$\underset{\theta_H, \theta_{S|\bar{h}}\theta_{E|\bar{h}}}{\text{argmax}} \; \mathbb{P}(\theta_H, \theta_{S|\bar{h}}\theta_{E|\bar{h}}|\mathcal{D})$$

These are known as MAP estimates, for maximum a posteriori estimates.

Introduction

The Bayesian approach does not commit to a single value of network parameters θ as it can work with a distribution over the possible values of these parameters, $\mathbb{P}(\theta|\mathcal{D})$.

The Bayesian approach can compute the expected value of a given query with respect to the distribution over network parameters. For example, the expected probability of observing a person that both smokes and exercises can be computed as follows:

$$\sum_{\theta} \Pr_{\theta}(s, e)\mathbb{P}(\theta|\mathcal{D}),$$

where $\Pr_{\theta}(.)$ is the distribution induced by the base network and parametrization θ.

Outline

- Define the notion of a meta network formally.
- Describe a particular class of meta networks that is commonly assumed in Bayesian learning.
- Parameter estimation while assuming that each parameter has a finite number of possible values.
- Parameter estimation for the continuous case.
- Learning network structure.

Parameter Sets

Let X be a variable with values $x_1, \ldots, x_k$, and let $\mathbf{U}$ be its parents

A parameter set for variable X and parent instantiation $\mathbf{u}$, denoted by $\theta_{X|\mathbf{u}}$, is the set of network parameters $(\theta_{x_1|\mathbf{u}}, \ldots, \theta_{x_k|\mathbf{u}})$. A parameter set that admits a finite number of values is said to be discrete, otherwise it is said to be continuous.

The parameter set $\theta_{S|\bar{h}}$ admits the following two values:

$$\theta_{S|\bar{h}} \in \{(.25, .75), (.50, .50)\}$$

This parameter set is therefore discrete, and each of its values corresponds to an assignment of probabilities to the set of co-varying parameters $(\theta_{s|\bar{h}}, \theta_{\bar{s}|\bar{h}})$. Hence, if $\theta_{S|\bar{h}} = (.25, .75)$, then $\theta_{s|\bar{h}} = .25$ and $\theta_{\bar{s}|\bar{h}} = .75$.

Parameter Sets

To further spell out our notational conventions for parameter sets, consider the following expression:

$$\sum_{\theta_{S|\bar{h}}} \theta_{s|\bar{h}} \theta_{\bar{s}|\bar{h}}$$

That is, we are summing over all possible values of the parameter set $\theta_{S|\bar{h}}$, and then multiplying the values of parameters corresponding to each element of the summand. The above expression will therefore evaluate to:

$$(.25)(.75) + (.50)(.50)$$

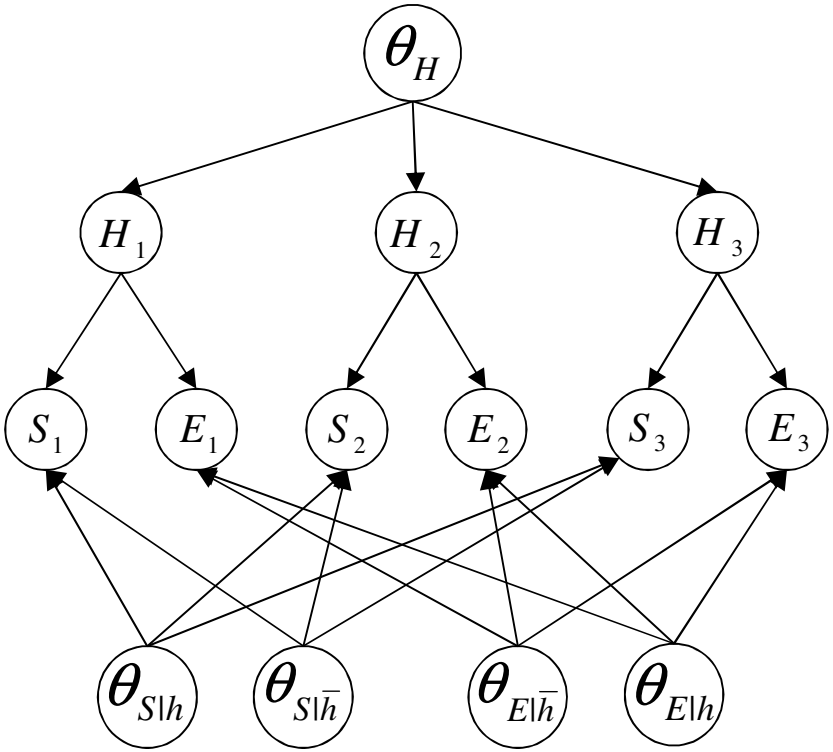
We will write an number of expressions later that resemble the form given above.

Meta Networks

Let G be a network structure

A meta network of size N for structure G is constructed using N instances of structure G, with variable X in G appearing as X_i in the ith instance of G. Moreover, for every variable X in G and its parent instantiation $\mathbf{u}$, the meta network contains the parameter set $\theta_{X|\mathbf{u}}$ and corresponding edges $\theta_{X|\mathbf{u}} \rightarrow X_1, \ldots, \theta_{X|\mathbf{u}} \rightarrow X_N$

Meta Networks



A meta network for the structure $S \leftarrow H \rightarrow E$

Meta Networks

We will distinguish between the base network, which is a classical Bayesian network, and the meta network.

We will also use θ to denote the set of all parameters for the base network, and call it a parametrization. Equivalently, θ will represent the collection of parameter sets in the meta network.

The distribution induced by a base network and parametrization θ will be denoted by $\Pr_\theta(.)$ and called a base distribution.

The distribution induced by a meta network will be denoted by $\mathbb{P}(.)$ and called a meta distribution.

Prior Knowledge

Prior knowledge on network parameters is encoded in the meta network using the CPTs of parameter sets.

For example, we have assumed in that the two values of parameter set $\theta_{S|\bar{h}}$ are equally likely. Hence, the CPT of this parameter set is as follows:

$\theta_{S\|\bar{h}} = (\theta_{s\|\bar{h}}, \theta_{\bar{s}\|\bar{h}})$	$\mathbb{P}(\theta_{S\|\bar{h}})$
$(.25, .75)$	50%
$(.50, .50)$	50%

These CPTs are then given as input to the learning process and lead to a major distinction with the ML approach to learning.

Prior Knowledge

The CPTs of other variables in a meta network (i.e., ones that do not correspond to parameter sets) are determined by the intended semantics of such networks.

Consider a variable X in the base network having parents $\mathbf{U}$, and let $X_1, \ldots, X_n$ be the instances of X and $\mathbf{U}_1, \ldots, \mathbf{U}_n$ be the instances of $\mathbf{U}$ in the meta network. All instances of X will have the same CPT in the meta network:

$$\mathbb{P}(X_i|\mathbf{u}_i, \theta_{X|\mathbf{u}^1}, \ldots, \theta_{X|\mathbf{u}^m}) = \theta_{X|\mathbf{u}^j}, \text{ where } \mathbf{u}^j = \mathbf{u}_i$$

Example:

$$\begin{aligned} \mathbb{P}(S_i|H_i = h, \theta_{S|h}, \theta_{S|\bar{h}}) &= \theta_{S|h} \\ \mathbb{P}(S_i|H_i = \bar{h}, \theta_{S|h}, \theta_{S|\bar{h}}) &= \theta_{S|\bar{h}} \end{aligned}$$

Data as Evidence

Case	H	S	E
1	h	$\bar{s}$	e
2	h	$\bar{s}$	$\bar{e}$
3	$\bar{h}$	s	$\bar{e}$

The data set can be viewed as the following variable instantiation:

$$\mathcal{D} = (H_1 = h) \wedge (S_1 = \bar{s}) \wedge (E_1 = e) \wedge \ldots \wedge (H_3 = \bar{h}) \wedge (S_3 = s) \wedge (E_3 = \bar{e})$$

One can assert this data set as evidence on the meta network, and then compute the corresponding posterior distribution on network parameters.

Data as Evidence

We initially have the following distribution on parameter sets:

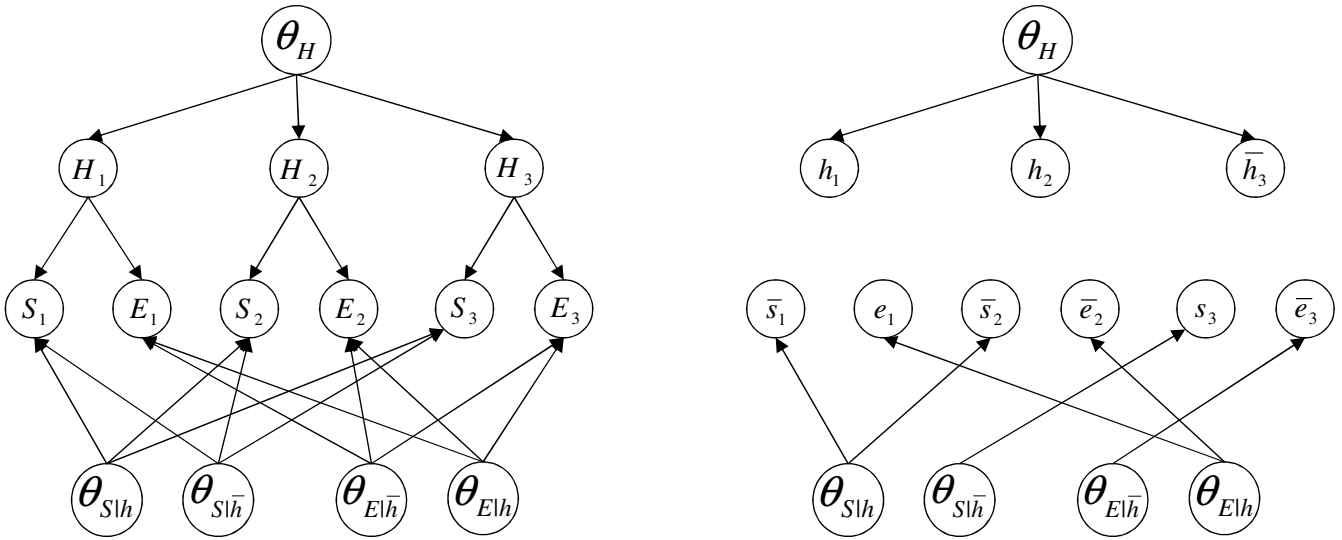
$$\mathbb{P}(\theta_H, \theta_{S|h}, \theta_{S|\bar{h}}, \theta_{E|h}, \theta_{E|\bar{h}}) = \mathbb{P}(\theta_H)\mathbb{P}(\theta_{S|h})\mathbb{P}(\theta_{S|\bar{h}})\mathbb{P}(\theta_{E|h})\mathbb{P}(\theta_{E|\bar{h}})$$

Note how the prior distribution could be decomposed in this case, which is possible for any meta network given by (since parameter sets are root nodes and are therefore d-separated).

This decomposition holds for the posterior distribution as well, given that the data set is complete:

$$\begin{aligned}
&\mathbb{P}(\theta_H, \theta_{S|h}, \theta_{S|\bar{h}}, \theta_{E|h}, \theta_{E|\bar{h}}|\mathcal{D}) \\
&= \mathbb{P}(\theta_H|\mathcal{D})\mathbb{P}(\theta_{S|h}|\mathcal{D})\mathbb{P}(\theta_{S|\bar{h}}|\mathcal{D})\mathbb{P}(\theta_{E|h}|\mathcal{D})\mathbb{P}(\theta_{E|\bar{h}}|\mathcal{D})
\end{aligned}$$

Data as Evidence



(a) meta network (b) pruned meta network

Pruning edges of a meta network based on a complete data set. Removed edges are either outgoing from observed variables or representing superfluous dependencies.

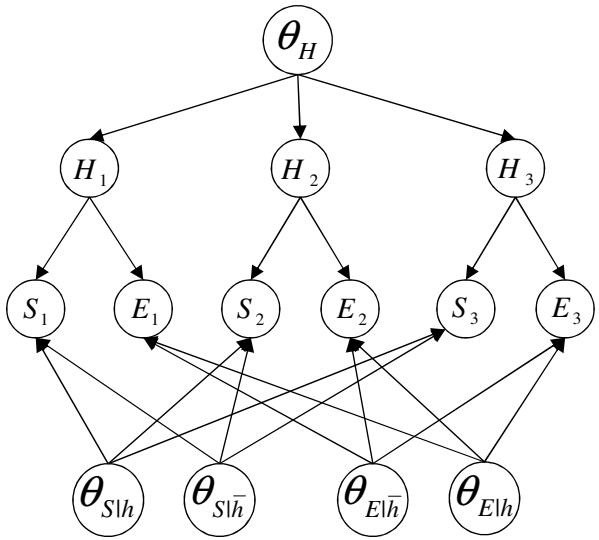
Parameter Independence

Consider a meta network and let Σ_1 and Σ_2 each contain a collection of parameter sets, $\Sigma_1 \cap \Sigma_2 = \emptyset$

The following conditions, known as parameter independence, are then guaranteed to hold:

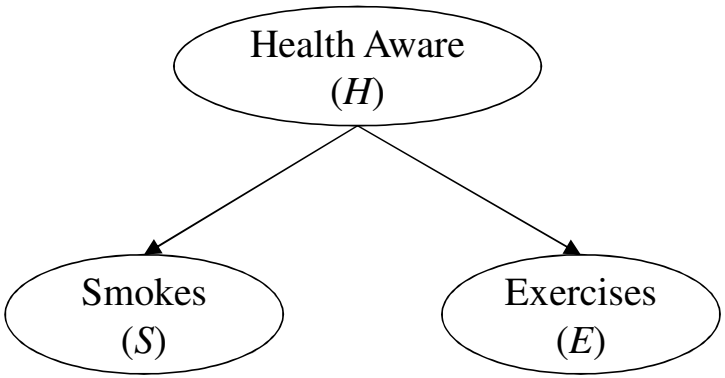
- Σ_1 and Σ_2 are independent, $\mathbb{P}(\Sigma_1, \Sigma_2) = \mathbb{P}(\Sigma_1)\mathbb{P}(\Sigma_2)$
- Σ_1 and Σ_2 are independent given any complete data set $\mathcal{D}$, $\mathbb{P}(\Sigma_1, \Sigma_2|\mathcal{D}) = \mathbb{P}(\Sigma_1|\mathcal{D})\mathbb{P}(\Sigma_2|\mathcal{D})$

Parameter Independence



Parameter independence is sometimes classified as either global or local. Global parameter independence refers to the independence between two parameter sets, $\theta_{X|\mathbf{u}}$ and $\theta_{Y|\mathbf{v}}$, corresponding to distinct variables $X \neq Y$. Local parameter independence refers to the independence between parameter sets, $\theta_{X|\mathbf{u}}$ and $\theta_{X|\mathbf{u}^\star}$, $\mathbf{u} \neq \mathbf{u}^\star$, corresponding to the same variable X

Learning with Discrete Parameter Sets



Prior knowledge

$$
\begin{aligned}
\theta_{S|h} &= (.1, .9) \\
\theta_{E|h} &= (.8, .2) \\
\theta_{H} &\in \{(.75, .25), (.90, .10)\} \\
\theta_{S|\bar{h}} &\in \{(.25, .75), (.50, .50)\} \\
\theta_{E|\bar{h}} &\in \{(.50, .50), (.75, .25)\}
\end{aligned}
$$

where each of the two values are considered equally likely.

Learning with Discrete Parameter Sets

Suppose now that our goal is to compute the probability of observing a smoker who exercises regularly, that is, s, e.

According to the ML approach, we first need to find the ML estimates θ^{ml} based on the given data, and then use them to compute this probability.

Among the eight possible parameterizations in this case, the one with maximum likelihood is:

$$\theta^{ml}: \quad \theta_H = (.75, .25),\ \theta_{S|\bar{h}} = (.25, .75),\ \theta_{E|\bar{h}} = (.50, .50)$$

If we plug in these parameter values in the base network:

$$\Pr_{\theta^{ml}}(s, e) \approx 9.13\%$$

Learning with Discrete Parameter Sets

The Bayesian approach, however, treats this problem differently.

It views the data set $\mathcal{D}$ as evidence on variables $H_1, S_1, E_1, \ldots, H_5, S_5, E_5$ in the meta network.

It then computes the posterior on variables S_6 and E_6 by performing inference on this meta network, leading to:

$$\mathbb{P}(S_6 = s, E_6 = e|\mathcal{D}) \approx 11.06\%$$

Learning with Discrete Parameter Sets

The Bayesian approach is therefore not estimating any parameters as is done in the ML approach.

Given discrete parameter sets, and a data set $\mathcal{D}$ of size N, we have

$$\mathbb{P}(\alpha_{N+1}|\mathcal{D}) = \sum_{\theta} \Pr_{\theta}(\alpha)\mathbb{P}(\theta|\mathcal{D}).$$

Here, event α_{N+1} is obtained from α by replacing every occurrence of variable X by its instance X_{N+1}.

Learning with Discrete Parameter Sets

For example, if α is $S=s, E=e$, then α_6 is $S_6=s, E_6=e$ We then have:

$$\mathbb{P}(S_6=s, E_6=e|\mathcal{D}) = \sum_{\theta} \Pr_\theta(S=s, E=e)\mathbb{P}(\theta|\mathcal{D}).$$

The Bayesian approach is therefore considering every possible parametrization θ, computing the probability $\Pr_\theta(S=s, E=e)$ using the base network, and then taking a weighted average of the computed probabilities. In other words, the Bayesian approach is computing the expected value of $\Pr_\theta(S=s, E=e)$.

Learning with Discrete Parameter Sets

If the data set is complete, one can compute Bayesian estimates by performing inference on the base network.

Let $\theta_{X|\mathbf{u}}$ be a discrete parameter set

The Bayesian estimate for parameter $\theta_{x|\mathbf{u}}$ given data set $\mathcal{D}$ is defined as follows:

$$\theta^{be}_{x|\mathbf{u}} \stackrel{def}{=} \sum_{\theta_{X|\mathbf{u}}} \theta_{x|\mathbf{u}} \cdot \mathbb{P}(\theta_{X|\mathbf{u}}|\mathcal{D})$$

The set of all Bayesian estimates $\theta^{be}_{x|\mathbf{u}}$ will be denoted by θ^{be}.

Learning with Discrete Parameter Sets

Given discrete parameter sets, and a complete data set $\mathcal{D}$ of size N, we have

$$\mathbb{P}(\alpha_{N+1}|\mathcal{D}) = \Pr_{\theta^{be}}(\alpha),$$

where θ^{be} are the Bayesian estimates given data set $\mathcal{D}$

$\mathbb{P}(\alpha_{N+1}|\mathcal{D})$ is an expectation of the probability $\Pr_\theta(\alpha)$. Hence, we can compute this expectation by performing inference on a base network that is parameterized by the Bayesian estimates.

Computing Bayesian Estimates

Bayesian estimates are easy to compute.

Let $\theta_{X|\mathbf{u}}$ be a discrete parameter set, and let $\mathcal{D}$ be a complete data set. We then have

$$\mathbb{P}(\theta_{X|\mathbf{u}}|\mathcal{D}) = \eta\, \mathbb{P}(\theta_{X|\mathbf{u}}) \prod_x \left[\theta_{x|\mathbf{u}}\right]^{\mathcal{D}\#(x\mathbf{u})},$$

where η is a normalizing constant.

Computing Bayesian Estimates

Consider now the parameter set $\theta_{E|\bar{h}}$ with values $\{(.50, .50), (.75, .25)\}$ and a uniform prior.

Case	H	S	E
1	F	F	T
2	T	F	T
3	T	F	T
4	F	F	F
5	F	T	F

We then have the following posterior:

$$\mathbb{P}(\theta_{E|\bar{h}} = (.50, .50)|\mathcal{D}) = \eta \times .50 \times \Big[.50\Big]^1 \Big[.50\Big]^2$$

$$\mathbb{P}(\theta_{E|\bar{h}} = (.75, .25)|\mathcal{D}) = \eta \times .50 \times \Big[.75\Big]^1 \Big[.25\Big]^2$$

Normalizing:

$$\mathbb{P}(\theta_{E|\bar{h}} = (.50, .50)|\mathcal{D}) \approx 72.73\%$$

$$\mathbb{P}(\theta_{E|\bar{h}} = (.75, .25)|\mathcal{D}) \approx 27.27\%$$

Computing Bayesian Estimates

Given:

$$\begin{aligned}\mathbb{P}(\theta_{E|\bar{h}} = (.50, .50)|\mathcal{D}) &\approx 72.73\% \\ \mathbb{P}(\theta_{E|\bar{h}} = (.75, .25)|\mathcal{D}) &\approx 27.27\%\end{aligned}$$

We can now compute the Bayesian estimate for every parameter by taking its expectation according to the above posterior:

$$\begin{aligned}\theta^{be}_{e|\bar{h}} &= .50 \times 72.73\% + .75 \times 27.27\% \approx .57 \\ \theta^{be}_{\bar{e}|\bar{h}} &= .50 \times 72.73\% + .25 \times 27.27\% \approx .43\end{aligned}$$

The Bayesian estimate for parameter set $\theta_{E|\bar{h}} = (\theta_{e|\bar{h}}, \theta_{\bar{e}|\bar{h}})$ is then $(.57, .43)$ in this case.

Closed Forms for Complete Data

Assuming here a base network with families $X\mathbf{U}$ and a complete data set $\mathcal{D}$ of size N:

- The prior probability of network parameters:

$$\mathbb{P}(\theta) = \prod_{X\mathbf{U}} \prod_{\mathbf{u}} \mathbb{P}(\theta_{X|\mathbf{u}})$$

- The posterior probability of network parameters:

$$\mathbb{P}(\theta|\mathcal{D}) = \prod_{X\mathbf{U}} \prod_{\mathbf{u}} \mathbb{P}(\theta_{X|\mathbf{u}}|\mathcal{D})$$

- The likelihood of network parameters:

$$\mathbb{P}(\mathcal{D}|\theta) = \prod_{i=1}^{N} \mathbb{P}(\mathbf{d}_i|\theta) = \prod_{i=1}^{N} \Pr_\theta(\mathbf{d}_i)$$

Closed Forms for Complete Data

Assuming here a base network with families $X\mathbf{U}$ and a complete data set $\mathcal{D}$ of size N:

- The marginal likelihood:[2]

$$\mathbb{P}(\mathcal{D}) = \prod_{i=1}^{N} \mathbb{P}(\mathbf{d}_i|\mathbf{d}_1, \ldots, \mathbf{d}_{i-1}) = \prod_{i=1}^{N} \Pr_{\theta_i^{be}}(\mathbf{d}_i),$$

where θ_i^{be} are the Bayesian estimates for data set $\mathbf{d}_1, \ldots, \mathbf{d}_{i-1}$

[2] Since $\mathbb{P}(\mathcal{D}|\theta)$ is called the likelihood of parameters θ, the quantity $\mathbb{P}(\mathcal{D})$ is called the marginal likelihood since it equals $\sum_\theta \mathbb{P}(\mathcal{D}|\theta)\mathbb{P}(\theta)$.

Closed Forms for Complete Data

One can easily compute MAP estimates under complete data.

$$\theta^{ma} = \operatorname*{argmax}_{\theta} \mathbb{P}(\theta|\mathcal{D})$$

Given parameter independence, we then have:

$$\theta^{ma}_{X|\mathbf{u}} = \operatorname*{argmax}_{\theta_{X|\mathbf{u}}} \mathbb{P}(\theta_{X|\mathbf{u}}|\mathcal{D})$$

Since,

$$\mathbb{P}(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathcal{D})} \propto \mathbb{P}(\mathcal{D}|\theta)\mathbb{P}(\theta),$$

the only difference between MAP and ML parameters is in the prior $\mathbb{P}(\theta)$. If all network parameterizations are equally likely, that is, $\mathbb{P}(\theta)$ is a uniform distribution, MAP and ML parameters coincide:

$$\operatorname*{argmax}_{\theta} \mathbb{P}(\theta|\mathcal{D}) = \operatorname*{argmax}_{\theta} \mathbb{P}(\mathcal{D}|\theta).$$

Learning with Continuous Parameter Sets

Assumed that the parameter set $\theta_{S|\bar{h}}$ has only two values $(.25, .75)$ and $(.50, .50)$ since our prior knowledge precluded all other values for network parameters $(\theta_{s|\bar{h}}, \theta_{\bar{s}|\bar{h}})$.

If on the other hand we allow all possible values for these parameters, the parameter set $\theta_{S|\bar{h}}$ will then be continuous (i.e., having an infinite number of values).

To apply Bayesian learning in this context, we need a method for capturing prior knowledge on continuous parameter sets (CPTs are only appropriate for discrete parameter sets).

Dirichlet Priors

Consider the parameter set $\theta_H = (\theta_h, \theta_{\bar{h}})$ in and suppose that we expect it will have the value $(.75, .25)$, yet we are not ruling out other values, such as $(.90, .10)$ and $(.40, .60)$

Suppose further that our belief in other values will decrease as they deviate more from the expected value $(.75, .25)$

Dirichlet Priors

One way to specify this knowledge is using a Dirichlet distribution, which requires two numbers ψ_h and $\psi_{\bar{h}}$, called exponents, where

$$\frac{\psi_h}{\psi_h + \psi_{\bar{h}}}$$

is the expected value of parameter θ_h and

$$\frac{\psi_{\bar{h}}}{\psi_h + \psi_{\bar{h}}}$$

is the expected value of parameter $\theta_{\bar{h}}$.

Exponents $\psi_h = 7.5$ and $\psi_{\bar{h}} = 2.5$ give expectation $(\frac{7.5}{7.5+2.5}, \frac{2.5}{7.5+2.5}) = (.75, .25)$
Exponents $\psi_h = 75$ and $\psi_{\bar{h}} = 25$ give same expectation
There is an infinite number of exponents that one can use.

Dirichlet Priors

The sum of these exponents, $\psi_h + \psi_{\bar{h}}$, is interpreted as a measure of confidence in the expectations they lead to.

This sum is called the equivalent sample size of the Dirichlet distribution, where a larger equivalent sample size is interpreted as providing more confidence in the corresponding expectations.

Think of the exponent ψ_h as the number of health-aware individuals we have observed before having encountered the current data set, and similarly for the exponent $\psi_{\bar{h}}$.

Accordingly, the exponents $(\psi_h = 7.5, \psi_{\bar{h}} = 2.5)$ and $(\psi_h = 75, \psi_{\bar{h}} = 25)$ can both be used to encode the belief that 75% of the individuals are health-aware, yet the second set of exponents imply a stronger belief as they are based on a larger sample.

Dirichlet Priors

Consider now variable E and suppose that it takes three values:

- e_1: the individual does not exercise at all.
- e_2: the individual exercises but not regularly.
- e_3: the individual exercises regularly.

Suppose now that we wish to encode our prior knowledge about the parameter set $\theta_{E|h} = (\theta_{e_1|h}, \theta_{e_2|h}, \theta_{e_3|h})$. If we expect this set to have the value $(.10, .60, .30)$, we can then use the exponents:

$$\psi_{e_1|h} = 10, \quad \psi_{e_2|h} = 60, \quad \psi_{e_3|h} = 30$$

which lead to the expectations:

$$\frac{10}{10+60+30}, \quad \frac{60}{10+60+30}, \quad \frac{30}{10+60+30}$$

Dirichlet Priors

A Dirichlet distribution for a continuous parameter set $\theta_{X|\mathbf{u}}$ is specified by a set of exponents, $\psi_{x|\mathbf{u}} \geq 1$.[a] The equivalent sample size of the distribution is defined as:

$$\psi_{X|\mathbf{u}} \stackrel{def}{=} \sum_x \psi_{x|\mathbf{u}}.$$

The Dirichlet distribution has the following density:

$$\rho(\theta_{X|\mathbf{u}}) \stackrel{def}{=} \eta \prod_x \left[\theta_{x|\mathbf{u}}\right]^{\psi_{x|\mathbf{u}}-1},$$

where η is a normalizing constant:

$$\eta \stackrel{def}{=} \frac{\Gamma(\psi_{X|\mathbf{u}})}{\prod_x \Gamma(\psi_{x|\mathbf{u}})}.$$

Here, $\Gamma(.)$ is the Gamma function, which is an extension of the factorial function to real numbers.[b]

[a] The Dirichlet distribution can be defined for exponents $0 < \psi_{x|\mathbf{u}} < 1$, but its behavior for these exponents will lead to mathematical complications that we try to avoid here. For example, Equation **??** will not hold in this case.

[b] The Gamma function is generally defined as $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$. We have $\Gamma(1) = 1$ and $\Gamma(a+1) = a\Gamma(a)$, which means that $\Gamma(a) = (a-1)!$ when a is an integer ≥ 1.

Dirichlet Priors

The expected value of network parameter $\theta_{x|\mathbf{u}}$ is given by:

$$\mathrm{Ex}(\theta_{x|\mathbf{u}}) = \frac{\psi_{x|\mathbf{u}}}{\psi_{X|\mathbf{u}}}$$

The variance of this parameter is given by:

$$\mathrm{Va}\left(\theta_{x|\mathbf{u}}\right) = \frac{\mathrm{Ex}(\theta_{x|\mathbf{u}})(1 - \mathrm{Ex}(\theta_{x|\mathbf{u}}))}{\psi_{X|\mathbf{u}} + 1}$$

The larger the equivalent sample size, $\psi_{X|\mathbf{u}}$, the smaller the variance and, hence, the more confidence we have in the expected values of network parameters.

Dirichlet Priors

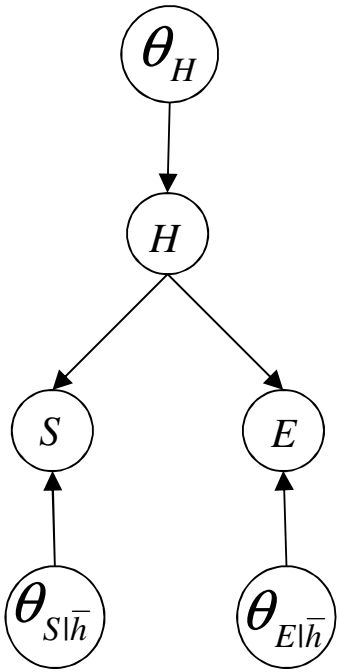
The mode of a parameter set is the value having the largest density:

$$\mathrm{Md}\left(\theta_{x|\mathbf{u}}\right) = \frac{\psi_{x|\mathbf{u}} - 1}{\psi_{X|\mathbf{u}} - |X|},$$

where $|X|$ is the number of values for variable X.

A Dirichlet distribution with two exponents is also known as the Beta distribution.

The Semantics of Continuous Parameter Sets



A meta network with discrete parameter sets induces a probability distribution, but a meta network with continuous parameter sets induces a density function.

The density function specified by this meta network:

$$\begin{aligned}
&\rho(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}, H, S, E) \\
&\quad = \quad \rho(\theta_H)\rho(\theta_{S|\bar{h}})\rho(\theta_{E|\bar{h}})\mathbb{P}(H|\theta_H)\mathbb{P}(S|H, \theta_{S|\bar{h}})\mathbb{P}(E|H, \theta_{E|\bar{h}})
\end{aligned}$$

The Semantics of Continuous Parameter Sets

The semantics of a network with continuous variables is defined by the chain rule, except that we now have a product of densities (for continuous variables) and probabilities (for discrete variables).

Discrete variables are summed out. Continuous variables are integrated over.

The marginal over parameter sets is a density given by:

$$\rho(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}) = \sum_{h,s,e} \rho(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}, H=h, S=s, E=e)$$

The marginal over discrete variables is a distribution given by:[3]

$$\mathbb{P}(H, S, E) = \int\int\int \rho(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}, H, S, E) d\theta_H d\theta_{S|\bar{h}} d\theta_{E|\bar{h}}$$

[3] Suppose that $\theta_{X|\mathbf{u}} = (\theta_{x_1|\mathbf{u}}, \ldots, \theta_{x_k|\mathbf{u}})$. Integrating over a parameter set $\theta_{X|\mathbf{u}}$ is a shorthand notation for successively integrating over parameters $\theta_{x_1|\mathbf{u}}, \ldots, \theta_{x_{k-1}|\mathbf{u}}$, while fixing the value of $\theta_{x_k|\mathbf{u}}$ to $1 - \sum_{i=1}^{k-1} \theta_{x_i|\mathbf{u}}$.

The Semantics of Continuous Parameter Sets

The result is a probability only if all continuous variables are integrated over; otherwise, the result is a density.

For example, the marginal over parameter set $\theta_{S|\bar{h}}$ is a density given by:

$$\rho(\theta_{S|\bar{h}}) = \int \int \Big[\sum_{h,s,e} \rho(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}, H=h, S=s, E=e) \Big] d\theta_H d\theta_{E|\bar{h}}$$

The Semantics of Continuous Parameter Sets

Density behaves like probability as far as independence is concerned.

For example, since the meta network satisfies parameter independence, we have:

$$\rho(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}) = \rho(\theta_H)\rho(\theta_{S|\bar{h}})\rho(\theta_{E|\bar{h}}),$$

and

$$\rho(\theta_H, \theta_{S|\bar{h}}, \theta_{E|\bar{h}}|\mathcal{D}) = \rho(\theta_H|\mathcal{D})\rho(\theta_{S|\bar{h}}|\mathcal{D})\rho(\theta_{E|\bar{h}}|\mathcal{D}),$$

when the data set $\mathcal{D}$ is complete.

The Semantics of Continuous Parameter Sets

Density also behaves like probability as far as conditioning is concerned.

For example,

$$\rho(H|\theta_H) = \frac{\rho(\theta_H, H)}{\rho(\theta_H)}$$

and

$$\rho(\theta_H|H) = \frac{\rho(\theta_H, H)}{\mathbb{P}(H)}$$

Bayesian Learning

Main result for Bayesian learning with continuous parameter sets

Given continuous parameter sets, and a data set $\mathcal{D}$ of size N, we have:[a]

$$\mathbb{P}(\alpha_{N+1}|\mathcal{D}) = \int \Pr_\theta(\alpha)\rho(\theta|\mathcal{D})d\theta$$

[a] Integrating over a parametrization θ is a shorthand notation for successively integrating over each of its parameter sets.

The quantity $\mathbb{P}(\alpha_{N+1}|\mathcal{D})$ is an expectation of the probability $\Pr_\theta(\alpha)$, which is defined with respect to the base network.

Bayesian Learning

Let $\theta_{X|\mathbf{u}}$ be a continuous parameter set

The Bayesian estimate for network parameter $\theta_{x|\mathbf{u}}$ given data set $\mathcal{D}$ is defined as follows:

$$\theta^{be}_{x|\mathbf{u}} \stackrel{def}{=} \int \theta_{x|\mathbf{u}} \cdot \rho(\theta_{X|\mathbf{u}}|\mathcal{D}) d\theta_{X|\mathbf{u}}$$

Bayesian Learning

As in the discrete case, we can sometimes reduce inference on a meta network to inference on a base network using the Bayesian estimates θ^{be}

Given continuous parameter sets, and a complete data set $\mathcal{D}$ of size N, we have

$$\mathbb{P}(\alpha_{N+1}|\mathcal{D}) = \Pr_{\theta^{be}}(\alpha),$$

where θ^{be} are the Bayesian estimates given data set $\mathcal{D}$

Computing Bayesian Estimates

The Bayesian estimates are at the heart of the Bayesian approach to learning, when the data set is complete.

The computation of these estimates, however, hinges on an ability to compute posterior marginals over parameter sets.

Consider a meta network where each parameter set $\theta_{X|\mathbf{u}}$ has a prior Dirichlet density $\rho(\theta_{X|\mathbf{u}})$ specified by exponents $\psi_{x|\mathbf{u}}$

Let $\mathcal{D}$ be a complete data set. The posterior density $\rho(\theta_{X|\mathbf{u}}|\mathcal{D})$ is then a Dirichlet density, specified by the following exponents:

$$\psi'_{x|\mathbf{u}} = \psi_{x|\mathbf{u}} + \mathcal{D}\#(x\mathbf{u})$$

Computing Bayesian Estimates

Consider now the parameter set $\theta_{S|h} = (\theta_{s|h}, \theta_{\bar{s}|h})$ with a prior density $\rho(\theta_{S|h})$ specified by the exponents

$$\psi_{s|h} = 1 \text{ and } \psi_{\bar{s}|h} = 9$$

The prior expectation of parameter $\theta_{s|h}$ is then .1

Case	H	S	E
1	F	F	T
2	T	F	T
3	T	F	T
4	F	F	F
5	F	T	F

The posterior density $\rho(\theta_{S|h}|\mathcal{D})$ is also Dirichlet, specified by the exponents

$$\psi'_{s|h} = 1 + 0 = 1 \text{ and } \psi'_{\bar{s},h} = 9 + 2 = 11$$

The posterior expectation of parameter $\theta_{s|h}$ is now $1/12$

Computing Bayesian Estimates

More generally, the posterior expectation of parameter $\theta_{x|\mathbf{u}}$ given complete data is given by:

$$\theta^{be}_{x|\mathbf{u}} = \frac{\psi_{x|\mathbf{u}} + \mathcal{D}\#(x\mathbf{u})}{\psi_{X|\mathbf{u}} + \mathcal{D}\#(\mathbf{u})}$$

where $\psi_{x|\mathbf{u}}$ are the exponents of the prior Dirichlet distribution and $\psi_{X|\mathbf{u}}$ is its equivalent sample size. This is the Bayesian estimate in the context of Dirichlet distributions.

The MAP estimate given complete data is:

$$\theta^{ma}_{x|\mathbf{u}} = \frac{\psi_{x|\mathbf{u}} + \mathcal{D}\#(x\mathbf{u}) - 1}{\psi_{X|\mathbf{u}} + \mathcal{D}\#(\mathbf{u}) - |X|}$$

In the above example, the MAP estimate for parameter $\theta_{s|h}$ is 0.

Computing Bayesian Estimates

Let us now compare these estimates with the ML estimate given in Chapter 17:

$$\theta^{ml}_{x|\mathbf{u}} = \frac{\mathcal{D}\#(x\mathbf{u})}{\mathcal{D}\#(\mathbf{u})}$$

Contrary to ML estimates, the Bayesian (and sometimes MAP) estimates do not suffer from the problem of zero counts.

These estimates are well defined even when $\mathcal{D}\#(\mathbf{u}) = 0$.

The Bayesian and MAP estimates will converge to the ML estimates as the data set size tends to infinity, assuming the data set is generated by a strictly positive distribution.

Computing Bayesian Estimates

A Dirichlet distribution is called non-informative if all exponents are equal to one: $\psi_{x|\mathbf{u}} = 1$

The expectation of parameter $\theta_{x|\mathbf{u}}$ is $1/|X|$ under this prior, leading to a uniform distribution for variable X given any parent instantiation $\mathbf{u}$

Under this prior, the Bayesian estimate given complete data is:

$$\theta^{be}_{x|\mathbf{u}} = \frac{1 + \mathcal{D}\#(x\mathbf{u})}{|X| + \mathcal{D}\#(\mathbf{u})}$$

Moreover, the MAP estimate is (coincides with the ML estimate[4])

$$\theta^{ma}_{x|\mathbf{u}} = \frac{\mathcal{D}\#(x\mathbf{u})}{\mathcal{D}\#(\mathbf{u})}$$

[4] This equality is not implied by the fact that parameters $\theta_{x|\mathbf{u}}$ have equal expectations. If all exponents are equal to, say, 10, all parameters will have equal expectations, yet the MAP and ML estimates will not coincide.

Closed Forms for Complete Data

Assuming a base network with families $X\mathbf{U}$ and a complete data set $\mathcal{D}$ of size N:

- The prior density of network parameters:

$$\rho(\theta) = \prod_{X\mathbf{U}} \prod_{\mathbf{u}} \rho(\theta_{X|\mathbf{u}})$$

- The posterior density of network parameters:

$$\rho(\theta|\mathcal{D}) = \prod_{X\mathbf{U}} \prod_{\mathbf{u}} \rho(\theta_{X|\mathbf{u}}|\mathcal{D})$$

- The likelihood of network parameters:

$$\mathbb{P}(\mathcal{D}|\theta) = \prod_{i=1}^{N} \Pr_\theta(\mathbf{d}_i)$$

Closed Forms for Complete Data

- The marginal likelihood:

$$\mathbb{P}(\mathcal{D}) = \prod_{X\mathbf{U}} \prod_{\mathbf{u}} \frac{\Gamma(\psi_{X|\mathbf{u}})}{\Gamma(\psi_{X|\mathbf{u}} + \mathcal{D}\#(\mathbf{u}))} \prod_{x} \frac{\Gamma(\psi_{x|\mathbf{u}} + \mathcal{D}\#(x\mathbf{u}))}{\Gamma(\psi_{x|\mathbf{u}})}$$

The proof of this closed form provides an alternative form that does not use the Gamma function, but the above form, which may seem surprising at first, is more commonly cited in the literature.