A Probabilistic Graphical Model Approach to Uncertainty Quantification

Nicholas Zabaras
Materials Process Design and Control Laboratory
Sibley School of Mechanical and Aerospace Engineering
101 Frank H. T. Rhodes Hall
Cornell University
Ithaca, NY 14853-3801

Email: nzabaras@gmail.com
URL: http://mpdc.mae.cornell.edu/

March 29, 2014

Numerical Methods for Uncertainty Quantification, Hausdorff Center for Mathematics, Universität Bonn, May 13-17, 2013
Content

- Introduction to graphical models
- Constructing a graphical model representation for a multiscale PDE system with stochastic input
- Factorizing the joint input/output PDF on a graph
- Introducing hidden variables (coarse graining)
- Efficient inference algorithm on a graphical model: belief propagation
- Using a probabilistic graphical model as a UQ surrogate

Examples

- C.M. Bishop, Pattern recognition and machine learning, Springer, 2007, Chapter 8
Why Graphical Models?

- Probabilistic graphical models have several useful properties:
  - 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, in which underlying mathematical expressions are carried along implicitly (possibly with linear scaling).
A graph comprises of nodes connected by links

- **Nodes** (also called vertices): represents a random variable (or group of random variables)

- **Links** (also called edges or arcs): express probabilistic relationships between these variables.

The graph then captures the way in which the joint distribution over all of the random variables can be decomposed into a product of factors each depending only on a subset of the variables.
Terminology

- **Parent and child**: whenever we have a directed edge, $X_i \rightarrow X_j$, we say that $X_j$ is the child of $X_i$, and $X_i$ is the parent of $X_j$.

- **Neighbor**: whenever we have an undirected edge, $X_i - X_j$, we say that $X_i$ is a neighbor of $X_j$.

- **Ancestor and descendant**: $X$ is an ancestor of $Y$ and that $Y$ is a descendant of $X$, if there exist a directed path $X_1, \ldots, X_k$ with $X_1 = X$ and $X_k = Y$.

- **Cycle and acyclic**: a cycle is a directed path $X_1, \ldots, X_k$ where $X_1 = X_k$. A graph is acyclic if it contains no cycles.
Probabilistic graphical models use a graph-based representation as the foundation for encoding a complete distribution over a multi-dimensional space and a graph that is a compact or factorized representation of a set of independences that hold in the specific distribution.

Two main types of probabilistic graphical model

Directed acyclic graph (Bayesian network)

Undirected graph (Markov random field)
Markov random fields, also known as undirected graphical models, in which the links do not carry arrows and have no directional significance.

- **Clique**: a subset of the nodes in a graph such that all the pair of nodes are connected.

- **Maximal clique**: a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.
Let us denote a clique by $C$ and the set of variables in that clique by $x_C$. Then the joint distribution is

$$p(x) = \frac{1}{Z} \prod_c \psi_c(x_C)$$

where $\psi_c(x_c)$ is the potential over clique $C$ and

$$Z = \sum_x \prod_c \psi_c(x_C)$$

is the normalization coefficient.

Without loss of generality, we can consider only maximal cliques which cannot be extended with other nodes without losing the fully connected property.
By considering only potential functions which satisfy $\psi_c(x_c) \geq 0$, we ensure that the joint probability $p(x) \geq 0$.

We can make sure the potentials are non-negative by using the exponential function,

$$\psi_c(x_c) = \exp\{-E_c(x_c)\}$$

Now we have to define $E_c(x_c)$ which can be anything!

Moreover, the joint also has a nice form

$$p(x) = \frac{1}{Z} \prod_c \exp(-E_c(x_c)) = \frac{1}{Z} \exp\left(-\sum_c E_c(x_c)\right)$$

In this way, the joint probability is represented using a **Boltzmann distribution**.
Illustration: Image De-Noising

Original Image

Noisy Image
The Markov random field model is shown in the figure. There are two types of cliques in the model, \( \{x_i, x_j\} \) and \( \{x_i, y_i\} \).

In this example, we add an extra term \( h x_i \) for each pixel \( i \), in order to bias the model towards pixel values that have one particular sign in preference to the other.

The energy function:

\[
E(x, y) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i
\]

The joint distribution:

\[
p(x, y) = \frac{1}{Z} \exp\left\{-E(x, y)\right\}
\]
Both directed and undirected graphs allow a global function of several variables to be expressed as a product of factors over subsets of those variables.

Factor graphs make this decomposition explicit by introducing additional nodes for the factors themselves in addition to the nodes for variables.

They also allow us to be more explicit about the details of the factorization.
Let us write the joint distribution over a set of variables in the form of a product of factors

\[ p(x) = \prod_s f_s(x_s) \]

For example, a distribution below can be expressed as a factor graph shown in the figure.

\[ p(x) = f_a(x_1, x_2) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_3) \]
An undirected graph can be readily converted to a factor graph.

\[ \psi(x_1, x_2, x_3) \]

\[ f(x_1, x_2, x_3) = \psi(x_1, x_2, x_3) \]

\[ f_a(x_1, x_2, x_3) f_b(x_2, x_3) = \psi(x_1, x_2, x_3) \]

Note that there may be several different factor graphs that correspond to the same undirected graph, as illustrated in the Figure above.
In many cases, the objective is to obtain marginal distributions or the joint probability of several random variables in a graphical model.

An important inference algorithm: belief propagation also known as sum-product algorithm.

- It is a technique invented to calculate marginals in Bayesian networks.
- It also works with undirected graphical models, factor graphs.
- BP provides exact solution when there are no loops in a graph, e.g. chain, tree.
- Otherwise, "loopy" BP provides approximate (but often very good) solutions.
To introduce generalized belief propagation, let us consider a pairwise undirected graphical model such that the joint probability is

\[ p(x) \propto \prod_{i,j} \psi_{ij}(x_i, x_j) \prod_i \phi_i(x_i) \]

In the BP algorithm, we introduce variables such as \( m_{ij}(x_j) \), which can intuitively be understood as a “message” from node \( i \) to node \( j \) about what state node \( j \) should be in.
Belief Propagation

- The messages $m_{ij}(x_j)$ will be a vector of the same dimensionality as $x_j$.

- In the BP algorithm, the belief (marginal) at node $i$ is proportional to the product of the local evidence at the node (i.e. $\phi_i(x_i)$), and all the messages coming into node $i$, i.e.

$$b_i(x_i) \propto \phi_i(x_i) \prod_{j \in N(i)} m_{ji}(x_i)$$
The messages are determined self-consistently by the message update rules:

\[ m_{ij}(x_j) \leftarrow \sum_{x_i} \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \setminus j} m_{ki}(x_i) \]
Belief propagation (BP) algorithm is the most commonly used technique for inference in probabilistic graphical models.

Two type of messages (update rules)

Variable node to factor node

\[ m_{x_i \rightarrow \mu}^{(n)} (x_i) \leftarrow \mu_i (x_i) \prod_{\mu_{pi} \in \Gamma(x_i) \setminus \mu} m_{\mu_{pi} \rightarrow x_i}^{(n-1)} (x_i) \]

Factor node to variable node

\[ m_{\mu \rightarrow x_i}^{(n)} (x_i) \leftarrow \int_{x_{\mu \setminus x_i}} \mu(x_{\mu}) \prod_{x_i \in x_{\mu \setminus x_i}} m_{x_i \rightarrow \mu}^{(n)} (x_i) \, dx_{\mu \setminus x_i} \]

Marginal distribution

\[ p^{(n)} (x_i) \propto \mu_i (x_i) \prod_{\mu_{pi} \in \Gamma(x_i)} m_{\mu_{pi} \rightarrow x_i}^{(n)} (x_i) \]

\[ p^{(n)} (x) \propto \mu(x) \prod_{x_j \in x} \prod_{\mu_{pi} \in \Gamma(x_j)} m_{\mu \rightarrow x_j}^{(n)} (x) \]
Solving Multiscale SPDEs Using Probabilistic Graphical Models
Define a probability space \((\Omega, F, P)\), we are interested in uncertainty quantification in a system represented by stochastic partial differential equations (SPDE) on a domain \(D \subset R^d\)

\[
L(x, \omega; u) = 0, \quad \forall x \in D \\
B(x, \omega; u) = 0, \quad \forall x \in \partial D
\]

Many challenges arise when both high-dimensional stochastic and multiscale phenomena are taken into account. Consider the fluid flow in porous media

\[
\mathbf{u}(x, \omega) = -K(x, \omega) \nabla p(x, \omega), \quad \forall x \in D \\
\nabla \cdot \mathbf{u}(x, \omega) = f(x), \quad a(x, \omega) = \log K
\]

with boundary conditions

\[
p = \bar{p} \quad \text{on } \partial D_p, \quad \mathbf{u} \cdot \mathbf{n} = \bar{\mathbf{u}} \quad \text{on } \partial D_u, \quad \partial D = \partial D_p \cup \partial D_u
\]

- **stochastic input**: permeability \(K(x, \omega)\)
- **Multiscale scales**:
  length scale of the system >> length scale of property variation
Problem Definition

- Due to the multiscale features, the spatial domain $D$ is discretized into:
  - fine grid $T_h = \bigcup_{i=1}^{N_h} e_i$ where $N_h$ is the number of fine elements.
  - coarse grid $T_c = \bigcup_{i=1}^{N_c} E_i$ where $N_c$ is the number of coarse elements.

- Stochastic input:
  - assume a constant property on each fine element, i.e. $a = (a_1, \ldots, a_{N_h})$
  - local property on coarse element $E_k$: $a_k = \{a_j \mid e_j \subset E_k\}$ such that $a_k \subset a$

![Diagram showing fine and coarse grids with a local property on a coarse element $E_k$]
Given the probability of stochastic input, \( p(a) \), the objective is to estimate the statistics of model responses, e.g. mean, variance and/or marginal distribution of physical response at a fixed point.

\[
\begin{align*}
\text{Samples of input} & : a_1 \cdots a_N \\
\text{Call deterministic solver} & : Y_i = f(a_i) \\
\text{Compute statistics of response} Y & : p(Y)
\end{align*}
\]
We construct a graphical model to represent the joint probability of the stochastic multiscale input and output.

However, the relationships between these random variables are not explicit!

It is possible to construct a graphical model from observation data. But this suffers from the curse of dimensionality (Jiang & Zabaras, 2013)

In this work, we construct the probabilistic graphical model based on prior knowledge and assumptions (that can be fully relaxed at a cost!)
Suppose we are interested in physical responses on a set of regularly distributed points on a coarse grid. The corresponding physical responses are denoted by \( Y = (y_1 \ldots y_n) \).

**Tasks**

- Apply probability theory to infer statistics of physical responses, e.g. marginal distributions of \( y_i \in Y, p(y_i) \).

In a probabilistic framework, the multivariate joint distribution

\[
p(Y) = \int p(Y, a) da = \int p(Y|a)p(a) da
\]

- stochastic input model \( p(a) \) is generally learned from observation data
- approximate the conditional distribution \( p(Y|a) \)
Consider a conditional random field (CRF) representing $p(Y|a)$ with a Gibbs distribution as [Nowozin etc. 2010]

$$p(Y|a) \propto \exp(-\mathcal{E}(Y;a))$$

where $\mathcal{E}(Y;a)$ is an energy function with the general form

$$\mathcal{E}(Y;a) \approx \sum_{i \in I} \phi^{(1)}_i(y_i, a) + \sum_{(i,j) \in I \times I} \phi^{(2)}_{i,j}(y_i, y_j, a) + \cdots + \phi^{(|I|)}_{I}(Y, a)$$

$\phi^{(n)}$ are feature functions measuring the interactions among $n$ variables in $Y$.

In the approximation used of energy functions, high-order interactions among physical responses are ignored; hence only $\phi^{(1)}$ and $\phi^{(2)}$ are retained

$$\mathcal{E}(Y;a) \approx \sum_{i \in I} \phi_i(y_i, a) + \sum_{(i,j) \in I \times I} \phi_{i,j}(y_i, y_j, a)$$

This can be fully relaxed (and potentially an adaptive expansion is possible within a Bayesian framework)
Since only up to pairwise interactions between variables in \( Y \) are considered, we introduce the Gaussian Markov random field (GMRF) [Wainwright etc. 2008].

\[
p(Y | a) \propto \exp \left( -\sum_i f_i(a) y_i - \sum_i \sum_j f_{ij}(a) y_i y_j \right)
\]

where \( f() \) are functions of stochastic input \( a \). The feature functions are defined as,

\[
\phi_i (y_i, a) = f_i(a) y_i \quad \phi_{i,j} (y_i, y_j, a) = f_{ij}(a) y_i y_j
\]

However, this global approximation as of now makes little sense in practice.

- Due to the curse of dimensionality, it is difficult to estimate the coefficients \( \zeta \) (feature functions) in terms of \( a \).
- It is impractical to make inference directly from this probabilistic model of \( Y \).
The decomposition of the global distribution $p(Y|a)$ is proposed through construction of an undirected graphical model of the multiscale system.

In a multiscale system with stochastic input, the corresponding graph contains two types of nodes: input $a$ and output $Y$.
Let $G(\mathcal{V}, \mathcal{E})$ be an undirected graph

- **nodes** (random variables)
- **edges** (correlations)

Let $\mathcal{C}$ denote a collection of cliques of the graph (i.e., fully connected subsets of nodes)

The joint distribution of $X_\mathcal{V}$ can be represented by

$$p(X_\mathcal{V}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(X_c)$$

**Why graphical model?**

It makes the inference problem easier.
A factor graph expresses the structure of factorization of a joint distribution $p(X_1, X_2, \ldots)$ into a product of factor nodes (factor functions) as

$$p(X_v) \propto \prod_c f_c(X_c)$$

$$f_1(X_1) = f_1(x_1, x_2, x_3)$$

Convert undirected graph into factor graph: the potential functions of cliques are represented by factor functions.
A probabilistic model for $p(Y|a)$ is constructed with the help of graphical model.

As in mixed finite element method, we assume a constant pressure, $h_k$, on each coarse element. Besides, the flow in coarse elements interact with each other through edges. Thus the flux on middle points of edges of coarse elements are also of interest.

Finally, the target physical responses include pressure and flux, i.e. $Y = (u,h)$ where $h = \{h_1,\ldots, h_{Nc}\}$ denotes the pressure and $u = \{u_1,\ldots, u_{Mc}\}$ the flux.
As there is little prior information on relationships between random variables \{u, h\} and \(a\), all nodes in each coarse element are assumed correlated and thus are linked in the graph.

The spatial correlations between physical responses are considered.

**Assumption**
- Each response \(u_i, h_k\) is only correlated to its neighboring nodes (including local feature \(a_k\)) within the same coarse element.
- Long distance interactions among variables are ignored.
Maximal clique: potential \[ q_k(u_{I_k}, h_k; a_k) \]

Joint probability \[ p(u, h | a) \propto \prod_k q_k(u_{I_k}, h_k; a_k) \]
In this way, the nodes on a coarse element $E_k : \{u_{i \in I_k}, h_k, a_k\}$, form a maximal clique in an undirected graphical model. $I_k$ is the index set of responses $u_i$ in element $E_k$.

$$p(u, h | a) \propto \prod_k q_k(u_{I_k}, h_k; a_k)$$

where $q_k()$ is the potential function of the maximal clique on $E_k$. Thus the global joint distribution is factorized into local potential functions on coarse elements.

$$p(u, h | a) \propto \exp \left( - \sum_k \mathcal{E}_k (u_{I_k}, h_k; a_k) \right)$$

such that

$$q_k(u_{I_k}, h_k; a_k) = \exp \left( - \mathcal{E}_k (u_{I_k}, h_k; a_k) \right)$$
Apply the definition of energy function in these sub-problems, the local energy functions are expressed by

\[ \mathcal{E}_k (u_{I_k}, h_k; a_k) \approx \sum_{i \in I_k} \phi_{k,i} (u_i, a_k) + \sum_{(i,j) \in I_k \times I_k, i \neq j} \phi_{k,ij} (u_i, u_j, a_k) + \phi_{k,0} (h_k, a_k) + \sum_{i \in I_k} \phi_{k,i0} (u_i, h_k, a_k) \]

where

\[ \phi_{k,i} (u_i, a_k) = f_{k,i} (a_k) u_i + f_{k,ii} (a_k) u_i^2, \quad \phi_{k,ij} (u_i, u_j, a_k) = f_{k,ij} (a_k) u_i u_j \]

\[ \phi_{k,0} (h_k, a_k) = f_{k,0} (a_k) h_k + f_{k,00} (a_k) h_k^2, \quad \phi_{k,i0} (u_i, h_k, a_k) = f_{k,i0} (a_k) u_i h_k \]

Since the functions of local features \( a_k \) in the energy functions are unknown, a nonparametric model is adopted

\[ f_{k,\cdot} (a_k) \equiv f_{k,\cdot} (a_k; \theta_k) = \theta_{k,1} + \sum_{t=2}^{r} \theta_{k,t} \zeta_t (a_k) \]

with unnormalized Gaussian kernels

\[ \zeta_t (a_k) = \exp \left( - \frac{\| a_k - \bar{a}_t \|^2}{\sigma_{\zeta}^2} \right) \]
Maximal clique: potential

\[ q_k \left( u_{I_k}, h_k; a_k \right) = \exp \left( -\mathcal{E}_k \left( u_{I_k}, h_k; a_k \right) \right) \]

\[ \mathcal{E}_k \left( u_{I_k}, h_k; a_k \right) \approx \sum_{i \in I_k} \phi_{k,i} (u_i, a_k) + \sum_{(i,j) \in I_k \times I_k, i \neq j} \phi_{k,ij} (u_i, u_j, a_k) + \phi_{k,0} (h_k, a_k) + \sum_{i \in I_k} \phi_{k,i0} (u_i, h_k, a_k) \]
Given a set of samples of input \( \{ a_k^{(n)} \}_{n=1}^{N} \) and specifying the number of kernels, the centers of Gaussian kernels are determined using K-means clustering.

A typical choice of the kernel width \( \sigma_\zeta \) is the average minimum distance between two realizations in the input space, i.e.

\[
\sigma_\zeta^2 = \frac{1}{N} \sum_{i=1}^{N} \min_{i \neq j} \left| a_k^{(i)} - a_k^{(j)} \right|^2
\]

The function \( f_{k,.} \) is a mapping from local features \( a_k \) to a scalar variable \( \zeta_k \). In other words, \( f_{k,.} \) projects the high-dimensional input into a low-dimensional space.

Since these variables \( \{\zeta_k\} \) are not directly observable, we call them hidden variables in the probabilistic model.
Conditional probability distribution:

\[ p(u, h | a, \Theta) \propto \prod_k \prod_{i \in I_k} \exp\left(-f_{k,i} (a_k ; \theta_k) u_i - f_{k,ii} (a_k ; \theta_k) u_i^2\right) \cdot \prod_{(i,j) \in I_k \times I_k, i \neq j} \exp\left(-f_{k,ij} (a_k ; \theta_k) u_i u_j \right) \]

\[ \cdot \prod_{i \in I_k} \exp\left(-f_{k,i0} (a_k ; \theta_k) u_i h_k \right) \exp\left(-f_{k,0} (a_k ; \theta_k) h_k - f_{k,00} (a_k ; \theta_k) h_k^2 \right) \]

The relationships between hidden variables and local features are

\[ \xi_{k,i} = f_{k,i} (a_k ; \theta_k), \xi_{k,ii} = f_{k,ii} (a_k ; \theta_k), \xi_{k,ij} = f_{k,ij} (a_k ; \theta_k), \]

\[ \xi_{k,i0} = f_{k,i0} (a_k ; \theta_k), \xi_{k,0} = f_{k,0} (a_k ; \theta_k), \xi_{k,00} = f_{k,00} (a_k ; \theta_k) \]

Then the conditional probability can be formulated in terms of hidden variables \( \xi \)

\[ p(u, h | \xi) \propto \prod_k \prod_{i \in I_k} \exp\left(-\xi_{k,i} u_i - \xi_{k,ii} u_i^2\right) \cdot \prod_{(i,j) \in I_k \times I_k, i \neq j} \exp\left(-\xi_{k,ij} u_i u_j \right) \]

\[ \cdot \prod_{i \in I_k} \exp\left(-\xi_{k,i0} u_i h_k \right) \exp\left(-\xi_{k,0} h_k - \xi_{k,00} h_k^2 \right) \]
According to the definition of hidden variables, each hidden variable is completely fixed given corresponding local features \( a_k \) and hyperparameters, i.e.

\[
p(\xi_{i,k} | \xi_{j,k}, a_k, \theta_k) = p(\xi_{i,k} | a_k, \theta_k)
\]

In other words, the hidden variables are conditionally independent on local features, thus

\[
p(\xi | a, \Theta) \propto \prod_k p(\xi_{k,0} | a_k, \theta_k) \cdot p(\xi_{0,0} | a_k, \theta_k) \cdot \prod_{(i,j) \in I_k \times I_k} p(\xi_{k,ij} | a_k, \theta_k)
\]

\[
\cdot \prod_{i \in I_k} p(\xi_{k,i} | a_k, \theta_k) \cdot p(\xi_{k,i0} | a_k, \theta_k)
\]

where \( \Theta = \bigcup_k \theta_k \).

Since there exist deterministic relationships between any \( \xi_{k,\cdot} \) and \( a_k \), \( \xi_{k,\cdot} \) takes value at \( f_{k,\cdot}(a_k) \) with probability 1 given \( a_k \). Then the conditional probability of \( \xi_{k,\cdot} \) on \( a_k \) is a Delta function

\[
p(\xi_{k,\cdot} | a_k, \theta_k) = \delta(\xi_{k,\cdot} - f_{k,\cdot}(a_k; \theta_k))
\]
The hidden variables $\zeta$ capture fine-scale effects on a coarse-scale. In other words, the influence of high-dimensional stochastic input on responses is represented by $\zeta$. The undirected graphical model and the corresponding factor graph with hidden variables are depicted as
In order to factorize the joint probability $p(u,h|a)$ explicitly via the graphical model, the undirected graph is transformed into a factor graph in which the factor nodes are exactly the potential functions:

- $\mu_{k,i} = \exp(-\phi_{k,i}(u_i, \xi_{k,i}, \xi_{k,ii}))$
- $\mu_{k,ij} = \exp(-\phi_{k,ij}(u_i, u_j, \xi_{k,ij}))$
- $\mu_{k,0} = \exp(-\phi_{k,0}(u_i, \xi_{k,0}, \xi_{k,00}))$
- $\mu_{k,i0} = \exp(-\phi_{k,i0}(u_i, h_k, \xi_{k,i0}))$

The potential functions for any of the hidden variables $\xi$ and local features $a_k$ are simply the conditional distributions (delta functions).
The complete probabilistic model for $p(u,h|a)$ is then factorized as a product of potential functions measuring the interactions between random variables

$$
\mu_{k,0}(h_k, \xi_{k,0}, \xi_{k,00}) = \exp\left(-\xi_{k,0} h_k - \xi_{k,00} h_k^2\right)
$$

$$
\mu_{k,l}(h_k, u_l, \xi_{k,l0}) = \exp\left(-\xi_{k,l0} u_l h_k\right)
$$

$$
p(\xi_{k,ii} | a_k) = \delta_{\xi}(f_{k,ii}(a_k))
$$

$$
\mu_{k,i}(u_i, \xi_{k,i}, \xi_{k,ii}) = \exp\left(-\xi_{k,i0} u_i - \xi_{k,i0} u_i^2\right)
$$

$$
\mu_{k,ij}(u_i, u_j, \xi_{k,ij}) = \exp\left(-\xi_{k,ij} u_i u_j\right)
$$
Remark

- The hidden variables $\xi$ on different coarse elements $k$ are different. For a stationary permeability random field $a$, one can assume that the hyperparameters $\theta_k$ are the same on different coarse elements. This is because local features $a_k$ have the same distribution on coarse elements and one should expect the same relationships between hidden variables and local input on different elements. Thus one can consider that the hidden variables on different elements have the same marginal distribution.

- However, they cannot be treated as the same variables as they are associated with local features $a_k$. Given a realization of stationary stochastic input $a^{(i)}$, the local features on elements $E_k$ and $E_l$ are generally different. Even though the hyperparameters could be identical, the realizations of the hidden variables are still different.

- If we have a nonstationary random field, the hidden variables on different elements are different variables with different marginal distributions.
Remark
- The subscripts of feature functions, potential functions, hidden variables as well as coefficient functions are defined according to the following rules:
  - $(k,i)$ denotes the flux component with flux index $i$,
  - $(k,ij)$ denotes interaction terms between the fluxes $u_i$ and $u_j$ ($i \neq j$),
  - $(k,0)$ denotes the pressure $h_k$ on the coarse element $k$,
  - $(k,i0)$ denotes interaction terms between flux $u_i$ and pressure $h_k$. 
We now need to learn the various parameters that define the probabilistic graphical model. Suppose we have a training set 
\( D = \{a^{(t)}, u^{(t)}, h^{(t)}\}_{t=1}^{N} \). The likelihood function of training data is formulated as

\[
p(D | \Theta) \equiv \prod_{t=1}^{N} p\left(u^{(t)}, h^{(t)} | a^{(t)}, \Theta\right) p\left(a^{(t)}\right) \]

\[
\propto \prod_{t=1}^{N} \exp\left(-\sum_{k} \mathcal{E}_{k}\left(a^{(t)}_{i}, h^{(t)}_{k}; a^{(t)}_{k}, \theta_{k}\right)\right)
\]

\[
\propto \prod_{k} \exp\left(-\sum_{t=1}^{N} \mathcal{E}_{k}\left(a^{(t)}_{i}, h^{(t)}_{k}; a^{(t)}_{k}, \theta_{k}\right)\right)
\]

Set the prior distribution for \( \Theta \), the Bayesian posterior is

\[
p(\Theta | D) \propto p(D | \Theta) p(\Theta)
\]

where \( p(\Theta) \) is chosen to be a multivariate Gaussian with mean zero and an identity covariance matrix.
As the prior leads to $p(\Theta) = \prod_k p(\theta_k)$, the posterior distribution can be decomposed as

$$p(\Theta | D) = \prod_k p(\theta_k | D_k)$$

$$\propto \prod_k \exp \left( -\sum_{t=1}^{N} \mathcal{E}_k \left( u_{i_k}^{(t)}, h_{k}^{(t)}; a_{k}^{(t)}, \theta_k \right) \right) p(\theta_k)$$

As a result, $\Theta$ in the probabilistic graphical model can be estimated locally!!

We will use a special Monte Carlo method --- **Sequential Monte Carlo** (SMC) to estimate the hyperparameters $\theta_k$ on each coarse element through local posterior $p(\theta_k | D_k)$. The samples that maximize the posterior will be taken as fixed parameter values in the probabilistic model for inference [de Moral etc. 2006].
Let the posterior \( \pi_n(\theta_k) \equiv p(\theta_k | D_k) \) be the target distribution. A set of auxiliary distributions are defined as
\[
\pi_t(\theta_k) \propto p^{\gamma_t}(D_k | \theta_k) p(\theta_k)
\]
where \( t = 0, 1, \ldots, n \) and \( 0 = \gamma_0 < \gamma_1 < \ldots < \gamma_n = 1 \) are tempering parameters and
\[
p(\theta_k | D_k) \propto \prod_k \exp \left( -\sum_{t=1}^{N} \mathcal{E}_k \left( u^{(t)}_{I_k}, h^{(t)}_{I_k}; a^{(t)}_k, \theta_k \right) \right)
\]

Remark

- When the stochastic input is a stationary random field, the local features have the same joint distribution. Thus we assume that the hidden variables have the same relationships with the local features, i.e. \( \Theta = \theta_1 \cdots = \theta_{N_e} \). Thus the global posterior \( p(\Theta | D) \) is directly used to infer the hyperparameters. The auxiliary distributions are defined as
\[
\pi_t(\Theta) \propto p^{\gamma_t}(D | \Theta) p(\Theta)
\]
Inference on the Probabilistic Graphical Model

Computing Marginal Distributions
Belief propagation (BP) algorithm is the most commonly used techniques for the inference problem in graphical model.

\[
\begin{align*}
&x_p \quad \text{..} \quad \mu_i \\
&\text{..} \\
&x_i \quad \text{..} \quad \mu \\
&x_j \\
&x_k \\
&\text{..} \\
&x_p \quad \text{..} \quad \mu_i \\
&\text{..} \\
&x_i \quad \text{..} \quad \mu \\
&x_j \\
&x_k
\end{align*}
\]

Two type of messages (update rules)

variable node to factor node
\[
m_{x_i \rightarrow \mu}^{(n)} (x_i) \leftarrow \mu_i (x_i) \prod_{\mu_{pi} \in \Gamma(x_i) \setminus \mu} m_{\mu_{pi} \rightarrow x_i}^{(n-1)} (x_i)
\]

factor node to variable node
\[
m_{\mu \rightarrow x_i}^{(n)} (x_i) \leftarrow \int_{x_{\mu \setminus x_i}} \mu (x_{\mu}) \prod_{x_i \in x_{\mu \setminus x_i}} m_{x_i \rightarrow \mu}^{(n)} (x_i) d x_{\mu \setminus x_i}
\]

Marginal distribution
\[
p^{(n)} (x_i) \propto \mu_i (x_i) \prod_{\mu \in \Gamma(x_i)} m_{\mu \rightarrow x_i}^{(n)} (x_i)
\]
\[
p^{(n)} (x) \propto \mu (x) \prod_{x_j \in x, \mu \in \Gamma(x_j)} m_{\mu \rightarrow x_j}^{(n)} (x)
\]
Inference on Graphs

Task
- Given stochastic input model $p(a)$, predict the marginal distribution of response $p(u_i)$

- In theory, this marginalization is obtained from

$$p(y_i) = \int p(Y | a)p(a) da dY_{-i}$$

We need to run belief propagation algorithms on a complete graphical model including variables $(u,h,a)$. However, this is challenging due to the high-dimensionality of stochastic input $a$.

- As an alternative, we perform marginalization on coarse-scale variables with the joint distribution

$$p(u,h,\xi) = \int p(u,h | \xi)p(\xi | a)p(a) da$$

$$= p(u,h | \xi)p(\xi)$$
In the BP algorithm used in this work, the messages are updated in parallel. At each iteration, we calculate the messages from each factor node to its neighboring variable nodes as well as the messages from each variable node to its neighboring factor nodes.

The messages are considered as converged if their change is less than a threshold in two successive iterations. There exists a unique message between any factor node and one of its arguments.
For all messages except those between hidden variables, since there is no prior information, they are represented non parametrically (as weighted Gaussian mixtures).

Without loss of generality, consider the message from factor node $\mu_{k,ij}(u_i, u_j, \xi_{k,ij})$ to variable node $u_i$

$$m_{\mu_{k,ij} \rightarrow u_i}(u_i) \approx \sum_{t=1}^{T} l_t \mathcal{N}(u_i; \bar{u}_i^t, \sigma_i^2)$$

At iteration $n$ of the BP algorithm, the messages between factor nodes and variables are updated by

$$m_{\mu_{k,ij} \rightarrow u_i}^{(n)}(u_i) \leftarrow \int \mu_{k,ij}(u_i, u_j, \xi_{k,ij}) m_{u_j \rightarrow \mu_{k,ij}}^{(n)}(u_j) m_{\xi_{k,ij} \rightarrow \mu_{k,ij}}^{(n)}(\xi_{k,ij}) d\xi_{k,ij} du_j$$
However, a BP update which multiplies $d$ Gaussian mixtures, each with $T$ components can produce a Gaussian mixture with $T^d$ components.

Therefore, when updating the messages, we draw samples of $(u_i, u_j, \xi_{k,ij})$ from

$$
\mu_{k,ij}(u_i, u_j, \xi_{k,ij}) m_{u_j}^{(n)}(u_j) m_{\xi_{k,ij}}^{(n)}(\xi_{k,ij})
$$

using MCMC. Then the message is approximated using a Gaussian mixture model with $T$ kernels from the samples.
Another challenge lies in the update of messages between hidden variables. Although analytic expressions of $p(a)$ and $p(\zeta|a)$ are explicit, the joint distribution of hidden variables $\zeta$ could be complicated such that the links between them are implicit when stochastic input has been removed from the graph.

To bypass the difficulties in passing messages between hidden variables, the graphical model is transformed into the following one
Inference in Graphs

- In this factor graph, each hidden variable $\zeta$ is assigned with an incoming message $m(\zeta)$ measuring the information propagated from the other hidden variables.

- Suppose the message from a factor node, e.g. $\mu_{k,ij}$ to $\zeta_{k,ij}$ is denoted by $m_{\mu_{k,ij} \rightarrow \zeta_{k,ij}}(\xi_{k,ij})$, we have

$$p(\xi_{k,ij}) \propto m(\xi_{k,ij}) \cdot m_{\mu_{k,ij} \rightarrow \zeta_{k,ij}}(\xi_{k,ij})$$

- The incoming messages are also updated in the BP iteration. At iteration $n$, the input message is updated according to

$$m^{(n+1)}(\xi_{k,ij}) = p(\xi_{k,ij}) / m^{(n)}_{\mu_{k,ij} \rightarrow \zeta_{k,ij}}(\xi_{k,ij})$$
Step 0: Initialization

- initialize the input messages as the marginal distributions of hidden variables
  \[ m^{(0)}(\xi_{k,ij}) = p(\xi_{k,ij}), m^{(0)}(\xi_{k,i}) = p(\xi_{k,i}) \]

- initialize all messages between \( u \) and factor node as
  \[
m^{(0)}_{\mu_{k,ij} \to u_i}(u_i) = \sum_{t=1}^{n} l_t N(u_i; \bar{u}_i^{(t)}, \sigma_i^2)\]
  with \( l_t = 1/n, \bar{u}_i^{(t)} = t/n, \sigma_i^2 = 1 \)
At step $t$

- Input messages $m^{(t)}(\xi_{k,i})$ have been fixed in step $t-1$

- Update the message from $\mu_{k,i}$ to $u_i$

$$m^{(t)}_{\mu_{k,i} \to u_i} \leftarrow \int \mu_{k,i}(u_i, \xi_{k,i}, \xi_{k,ii}) m^{(t)}(\xi_{k,i}) m^{(t)}(\xi_{k,ii}) d\xi_{k,i} d\xi_{k,ii}$$

This is obtained by making samples $(u_i, \xi_{k,i}, \xi_{k,ii})$ of function $\mu_{k,i}(u_i, \xi_{k,i}, \xi_{k,ii}) m^{(t)}(\xi_{k,i}) m^{(t)}(\xi_{k,ii})$ using standard Metropolis-Hasting sampler

$m^{(t)}_{\mu_{k,j} \to u_i}$ are obtained by Gaussian mixture approximation of samples $(u_i)$, i.e.

$$m^{(t)}_{\mu_{k,j} \to u_i}(u_i) \approx \sum_{t=1}^{n} l_i N(u_i; \bar{u}_{i,t}^{(t)}, \sigma_i^2)$$

- Update the message from $\mu_{k,ij}$ to $u_i$ and the message from $\mu_{k,ij}$ to $u_j$ in the same way

$$m^{(t)}_{\mu_{k,ij} \to u_i} \leftarrow \int \mu_{k,ij}(u_i, u_i, \xi_{k,ij}) m^{(t)}(\xi_{k,ij}) m^{(t-1)}_{u_j \to \mu_{k,ij}}(u_j) d\xi_{k,ij} du_j$$

$$m^{(t)}_{\mu_{k,ij} \to u_j} \leftarrow \int \mu_{k,ij}(u_i, u_i, \xi_{k,ij}) m^{(t)}(\xi_{k,ij}) m^{(t-1)}_{u_i \to \mu_{k,ij}}(u_i) d\xi_{k,ij} du_i$$
At step $t$:

- Update the input message at step $t + 1$

\[
m_{\mu_{k,ij} \rightarrow \xi_{k,ij}}^{(t)}(\xi_{k,ij}) \leftarrow \int \mu_{k,ij}(u_i, u_i, \xi_{k,ij}) m_{u_i \rightarrow \mu_{k,ij}}^{(t)}(u_i) m_{u_j \rightarrow \mu_{k,ij}}^{(t)}(u_j) du_i du_j
\]

\[
m_{\mu_{k,ij} \rightarrow \xi_{k,ij}}^{(t+1)}(\xi_{k,ij}) = p(\xi_{k,ij}) / m_{\mu_{k,ij} \rightarrow \xi_{k,ij}}^{(t)}(\xi_{k,ij})
\]
At step $t$, compute the marginal distributions of each physical responses by multiplying all incoming messages from neighboring factor nodes

$$
p(u_i) \propto \prod_{\mu \in \Gamma(u_i)} m_{\mu \rightarrow i}(u_i) \approx \sum_{t=1}^{n} l_t N\left(u_i; \bar{u}_i^{(t)}, \sigma_i^2\right)$$

if $\max_i \left| u_i^{(t)} - u_i^{(t-1)} \right| < \varepsilon$, the marginal distribution of $u_i$ converges. The same rule is applied to pressure $h_k$.

Stop iteration until all marginal distributions converge.
Given a realization of stochastic input, $a^{(n)}$, the values of hidden variables can be directly obtained through functions $f(.)$.

As the hidden variables are observed in this case, there is no message between them. Then the factor graph corresponds to the conditional distribution $p(u,h|\xi^{(n)})$. The unobserved variables are $(u,h)$.

When belief propagation is performed, we get the marginals of physical responses conditioned on the input. Let the expectation $E(u_i|a^{(n)})$ and $E(h_k|a^{(n)})$ be predicted values of physical responses, we can get a surrogate model by running the belief propagation algorithm on a factor graph given a realization of stochastic input.
In this section, we construct probabilistic graphical model based solutions to predict fluid flow in random heterogeneous porous media.

The domain is a unit square $[0,1]^2$. The permeability is defined on a 64x64 fine grid and we are interested in flux at the middle point of edges of coarse elements as well as pressure on a 8x8 coarse grid.

The training data are generated using mixed multiscale FEM on the coarse grid.

We choose $r = 4$ kernels to approximate the relationships between hidden variables and local features.
The log-permeability is taken as a Gaussian Markov random field with the covariance kernel

\[ \text{Cov}(x, y) = \sigma^2 \exp\left(-\frac{|x_1 - y_1|}{L_1} - \frac{|x_2 - y_2|}{L_2}\right) \]

and zero mean. (\(\sigma=1.0, L_1=0.1, L_2=0.1\))

Governing equations

\[ u(x, \omega) = -K(x, \omega) \nabla p(x, \omega) \quad \forall x \in D \]
\[ \nabla \cdot u(x, \omega) = f, \quad a(x, \omega) = \log K \]

with source term

\[ f(x) = \begin{cases} 
  -r, & \text{if } 0 \leq x_i < w, \text{ for } i = 1, 2 \\
  r, & \text{if } 1 - w \leq x_i < 1, \text{ for } i = 1, 2 \\
  0 & \text{otherwise}
\end{cases} \]

where the parameters are chosen to be \(r = 10\) and \(w = 1/8\). No-flow Neumann boundary conditions are applied on all boundaries.
Predicted physical responses given a realization of stochastic input (a)-(c) x-velocity, y-velocity and pressure obtained from direct simulation, and (d)-(f) x-velocity, y-velocity and pressure predicted by the probabilistic graphical model (trained with 60 data points)
Predicted mean of x-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 20, (c) 40, (d) 60 data.
Predicted variance of x-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 20, (c) 40, (d) 60 data.
Predicted mean of y-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 20, (c) 40, (d) 60 data.
Predicted variance of y-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 20, (c) 40, (d) 60 data.
Example 1

Predicted mean of pressure (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 20, (c) 40, (d) 60 data.
Predicted variance of pressure (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 20, (c) 40, (d) 60 data.
Example 1

Predicted marginal PDF of x-velocity at point (0.5, 0.4375) with (a) 2 and (b) 4 Gaussian components in nonparametric messages.

Predicted marginal PDF of y-velocity at point (0.4375, 0.5) with (a) 2 and (b) 4 Gaussian components in nonparametric messages.
Predicted marginal PDF of pressure at an element centered at point (0.4375,0.4375) with (a) 2 and (b) 4 Gaussian components in nonparametric messages.

The $L_2$ norm of the error in variance as a function of the observed samples for MC simulation and graphical model prediction.
k-fold cross-validation error (k = 10) of x-velocity, y-velocity and pressure predicted by the probabilistic graphical model with (a)-(c) 40 samples, and (d)-(f) 60 samples
The joint PDF of the x-velocity $u_1$ at $(0.5, 0.4375)$ and $u_2$ at $(0.375, 0.4375)$: (a) direct simulation (b) probabilistic graphical model; the joint PDF of y-velocity $v_1$ at $(0.4375, 0.5)$ and $v_2$ at $(0.4375, 0.375)$: (c) direct simulation (d) probabilistic graphical model.
Example 2

- The log-permeability is taken as a Gaussian Markov random field with the covariance kernel

\[ \text{Cov}(x, y) = \sigma^2 \exp\left( -\frac{|x_1 - y_1|}{L_1} - \frac{|x_2 - y_2|}{L_2} \right) \]

and zero mean (\(\sigma=1.0, L_1=0.1, L_2=0.2\)).

- Governing equations

\[ u(x, \omega) = -K(x, \omega) \nabla p(x, \omega) \quad \forall x \in D \]

\[ \nabla \cdot u(x, \omega) = f, \quad a(x, \omega) = \log K \]

with source term \(f = 0\) and boundary conditions

\[ p = 1 \text{ on } x = 0, \quad p = 0 \text{ on } x = 1 \]

and no-flow Neumann boundary conditions on the other boundaries.

- The stochastic input, log-permeability, is generated on a 64x64 fine grid using KL expansion with 100 first terms retained. Training data are generated on a 8x8 grid using mixed multiscale FEM.
Predicted physical responses given a realization of stochastic input (a)-(c) x-velocity, y-velocity and pressure obtained from direct simulation, and (d)-(f) x-velocity, y-velocity and pressure predicted by the probabilistic graphical model (trained with 2400 data points).
Example 2

Predicted mean of x-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Predicted variance of x-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Example 2

Predicted mean of y-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Example 2

Predicted variance of y-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Example 2

Predicted mean of pressure (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Example 2

Predicted variance of pressure (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Predicted marginal PDF of x-velocity at point (0.5, 0.4375) with (a) 2 and (b) 4 Gaussian components in nonparametric messages

Predicted marginal PDF of y-velocity at point (0.4375, 0.5) with (a) 2 and (b) 4 Gaussian components in nonparametric messages
Predicted marginal $p$ PDF of pressure at an element centered at point $(0.4375, 0.4375)$ with (a) 2 and (b) 4 Gaussian components in nonparametric messages.

The $L_2$ norm of the error in variance as a function of the observed samples for MC simulation and graphical model prediction.
k-fold cross-validation error (k = 10) of x-velocity, y-velocity and pressure predicted by the probabilistic graphical model with (a)-(c) 40 samples, and (d)-(f) 60 samples
The joint PDF of the x-velocity $u_1$ at (0.5, 0.4375) and $u_2$ at (0.375, 0.4375): (a) direct simulation (b) probabilistic graphical model; the joint PDF of y-velocity $v_1$ at (0.4375, 0.5) and $v_2$ at (0.4375, 0.375): (c) direct simulation (d) probabilistic graphical model.
In previous examples, it was assumed that the porous media are stationary such that the covariance between any two points in the domain depends on their distance rather than their actual locations.

Here we extend the probabilistic graphical model to nonstationary permeability random fields.

In this example, we use a nonstationary random field as stochastic input. The log-permeability on the k-th coarse element is a Gaussian random field with mean zero and an exponential covariance function:

$$\text{Cov}(x, x^*) = \sigma^2 \exp \left( - \frac{|x_1 - x_1^*|}{L_{k,1}} - \frac{|x_2 - x_2^*|}{L_{k,2}} \right)$$

where

$$L_{k,1} = 0.1 + \frac{0.4}{N_y - 1} j_k, \quad L_{k,2} = 0.1 + \frac{0.4}{N_x - 1} i_k$$

and $$N_x = N_y = 8$$. The boundary conditions and source terms are the same as Example 2.
Predicted physical responses given a realization of stochastic input (a)-(c) x-velocity, y-velocity and pressure obtained from direct simulation, and (d)-(f) x-velocity, y-velocity and pressure predicted by the probabilistic graphical model (trained with 2400 data points)
Example 3

Predicted mean of x-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Example 3

Predicted variance of x-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data
Example 3

Predicted mean of y-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data
Example 3

Predicted variance of y-velocity (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Predicted mean of pressure (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Example 3

(a) (b) (c) (d)

Predicted variance of pressure (a) MC simulation with $10^6$ samples, and probabilistic model trained by (b) 800, (c) 1600, (d) 2400 data.
Predicted marginal PDF of x-velocity at point (0.5, 0.4375) with (a) 2 and (b) 4 Gaussian components in nonparametric messages.

Predicted marginal PDF of y-velocity at point (0.4375, 0.5) with (a) 2 and (b) 4 Gaussian components in nonparametric messages.
Predicted marginal PDF of pressure at an element centered at point \( (0.4375, 0.4375) \) with (a) 2 and (b) 4 Gaussian components in nonparametric messages.

The \( L_2 \) norm of the error in variance as a function of the observed samples for MC simulation and graphical model prediction.
k-fold cross-validation error (k = 10) of x-velocity, y-velocity and pressure predicted by the probabilistic graphical model with (a)-(c) 40 samples, and (d)-(f) 60 samples
The joint PDF of the x-velocity $u_1$ at (0.5, 0.4375) and $u_2$ at (0.375, 0.4375): (a) direct simulation (b) probabilistic graphical model; the joint PDF of y-velocity $v_1$ at (0.4375, 0.5) and $v_2$ at (0.4375, 0.375): (c) direct simulation (d) probabilistic graphical model.
Conclusions

- By treating stochastic input and physical responses as random variables, a probabilistic graphical model is constructed to efficiently represent their probabilistic relationship.

- The high-dimensional global joint distribution of input and output (responses) is modeled by decomposing it into lower-dimensional (spatially) local problems.

- Based on the probabilistic graphical model, the belief propagation algorithm can directly capture the uncertainty propagation from fine-scale to coarse-scale.

- The statistics of physical responses are accurately obtained through the probabilistic graphical model.