Inference in Multiparameter Models, Conditional Posterior, Local Conjugacy

Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Feb 4, 2019
Today’s Plan

- Foray into models with several parameters
- Goal will be to infer the posterior over all of them (not posterior for some, MLE-II for others)
- Idea of conditional/local posteriors in such problems
- **Local conjugacy** (which helps in computing conditional posteriors)
- **Gibbs sampling** (an algorithm that infer the joint posterior via conditional posteriors)
- An example: Bayesian matrix factorization (model with many parameters)
- Note: Conditional/local posterior, local conjugacy, etc are important ideas (will appear in many inference algorithms that we will see later)
So far we've usually seen models with one "main" parameter and maybe a few hyperparams.
Moving Beyond Simple Models..

- So far we've usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,
  - Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
Moving Beyond Simple Models..

- So far we've usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,
  - Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
  - Bayesian linear regression with weight vector $\mathbf{w}$ and noise/prior precision hyperparams $\beta, \lambda$

Hyperparams, if desired, can be also estimated via MLE-II

Note however that MLE-II would only give a point estimate of hyperparams

What if we have a model with lots of parameters/hyperparams and want posteriors over all of those?

Intractable in general but today we will look at a way of doing approx. inference in such models
So far we’ve usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,

- Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
- Bayesian linear regression with weight vector \( w \) and noise/prior precision hyperparams \( \beta, \lambda \)
- GP regression with one function to be learned
Moving Beyond Simple Models..

- So far we've usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,
  - Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
  - Bayesian linear regression with weight vector $w$ and noise/prior precision hyperparams $\beta, \lambda$
  - GP regression with one function to be learned

- Easy posterior inference if the likelihood and prior are conjugate to each other
Moving Beyond Simple Models..

- So far we've usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,
  - Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
  - Bayesian linear regression with weight vector $w$ and noise/prior precision hyperparams $\beta, \lambda$
  - GP regression with one function to be learned

- Easy posterior inference if the likelihood and prior are conjugate to each other
  - Otherwise have to approx. the posterior (e.g., Laplace approx. - recall Bayesian logistic regression)
So far we’ve usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,
- Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
- Bayesian linear regression with weight vector \( w \) and noise/prior precision hyperparams \( \beta, \lambda \)
- GP regression with one function to be learned

Easy posterior inference if the likelihood and prior are conjugate to each other
- Otherwise have to approx. the posterior (e.g., Laplace approx. - recall Bayesian logistic regression)

Hyperparams, if desired, can be also estimated via MLE-II
Moving Beyond Simple Models..

- So far we’ve usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,
  - Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
  - Bayesian linear regression with weight vector $w$ and noise/prior precision hyperparams $\beta, \lambda$
  - GP regression with one function to be learned

- Easy posterior inference if the likelihood and prior are conjugate to each other
  - Otherwise have to approx. the posterior (e.g., Laplace approx. - recall Bayesian logistic regression)

- Hyperparams, if desired, can be also estimated via MLE-II
  - Note however that MLE-II would only give a point estimate of hyperparams
Moving Beyond Simple Models..

- So far we've usually seen models with one “main” parameter and maybe a few hyperparams, e.g.,
  - Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
  - Bayesian linear regression with weight vector $w$ and noise/prior precision hyperparams $\beta, \lambda$
  - GP regression with one function to be learned

- Easy posterior inference if the likelihood and prior are conjugate to each other
  - Otherwise have to approx. the posterior (e.g., Laplace approx. - recall Bayesian logistic regression)

- Hyperparams, if desired, can be also estimated via MLE-II
  - Note however that MLE-II would only give a point estimate of hyperparams

- What if we have a model with lots of parameters/hyperparams and want posteriors over all of those?
Moving Beyond Simple Models..

- So far we've usually seen models with one "main" parameter and maybe a few hyperparams, e.g.,
  - Given data assumed to be from a Gaussian, infer the mean assuming variance known (or vice-versa)
  - Bayesian linear regression with weight vector $\mathbf{w}$ and noise/prior precision hyperparams $\beta, \lambda$
  - GP regression with one function to be learned

- Easy posterior inference if the likelihood and prior are conjugate to each other
  - Otherwise have to approx. the posterior (e.g., Laplace approx. - recall Bayesian logistic regression)

- Hyperparams, if desired, can be also estimated via MLE-II
  - Note however that MLE-II would only give a point estimate of hyperparams

- What if we have a model with lots of parameters/hyperparams and want posteriors over all of those?
  - Intractable in general but today we will look at a way of doing approx. inference in such models
Multiparameter Models

- Multiparameter models consist of two or more unknowns, say $\theta_1$ and $\theta_2$
- Given data $y$, some examples for the simple two parameter case
Multiparameter Models

- Multiparameter models consist of two or more unknowns, say $\theta_1$ and $\theta_2$
- Given data $y$, some examples for the simple two parameter case

Assume the likelihood model to be of the form $p(y|\theta_1, \theta_2)$ (e.g., case 1 and 3 above)
Multiparameter Models

- Multiparameter models consist of two or more unknowns, say $\theta_1$ and $\theta_2$
- Given data $y$, some examples for the simple two parameter case

\[ p(y|\theta_1, \theta_2) \] (e.g., case 1 and 3 above)

Assume a joint prior distribution $p(\theta_1, \theta_2)$
Multiparameter Models

- Multiparameter models consist of two or more unknowns, say $\theta_1$ and $\theta_2$
- Given data $y$, some examples for the simple two parameter case

Assume the likelihood model to be of the form $p(y|\theta_1, \theta_2)$ (e.g., case 1 and 3 above)
Assume a joint prior distribution $p(\theta_1, \theta_2)$
The joint posterior $p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$
Multiparameter Models

- Multiparameter models consist of two or more unknowns, say $\theta_1$ and $\theta_2$
- Given data $y$, some examples for the simple two parameter case

![Diagram showing relationships between parameters and data]

- Assume the likelihood model to be of the form $p(y|\theta_1, \theta_2)$ (e.g., case 1 and 3 above)
- Assume a joint prior distribution $p(\theta_1, \theta_2)$
- The joint posterior $p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$
  - Easy the joint prior is conjugate to the likelihood (e.g., NIW prior for Gaussian likelihood)
Multiparameter Models

- Multiparameter models consist of two or more unknowns, say $\theta_1$ and $\theta_2$
- Given data $y$, some examples for the simple two parameter case

Assume the likelihood model to be of the form $p(y|\theta_1, \theta_2)$ (e.g., case 1 and 3 above)

Assume a joint prior distribution $p(\theta_1, \theta_2)$

The joint posterior $p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$

- Easy the joint prior is conjugate to the likelihood (e.g., NIW prior for Gaussian likelihood)
- Otherwise needs more work, e.g., MLE-II, MCMC, VB, etc. (already saw MLE-II, will see more later)
Multiparameter Models: Some Examples

- Multiparameter models arise in many situations, e.g.,
  - Probabilistic models with unknown hyperparameters (e.g., Bayesian linear regression we just saw)
Multiparameter Models: Some Examples

- Multiparameter models arise in many situations, e.g.,
  - Probabilistic models with unknown hyperparameters (e.g., Bayesian linear regression we just saw)
  - Joint analysis of data from multiple (and possibly related) groups: Hierarchical models

![Diagram of a Bayesian model with parameters $\mu_w$, $\Sigma_w$, $y_{ij}$, $x_{ij}$, $w_j$, and $\sigma^2$. The diagram shows the relationship between the response variable $y_{ij}$ and the feature vector $x_{ij}$, with parameters $\mu_w$ and $\Sigma_w$ specifying a Gaussian prior $\mathcal{N}(\mu_w, \Sigma_w)$ for all $J$ regression models.]
Multiparameter Models: Some Examples

- Multiparameter models arise in many situations, e.g.,
  - Probabilistic models with unknown hyperparameters (e.g., Bayesian linear regression we just saw)
  - Joint analysis of data from multiple (and possibly related) groups: Hierarchical models

... and in fact, pretty much in any non-toy example of probabilistic model :)
Another Example Problem: Matrix Factorization/Completion

Given: Data \( D = \{ r_{ij} \} \) of “interactions” (e.g., ratings) of users \( i = 1, \ldots, N \) on \( j = 1, \ldots, M \) items.
Another Example Problem: Matrix Factorization/Completion

Given: Data $D = \{r_{ij}\}$ of “interactions” (e.g., ratings) of users $i = 1, \ldots, N$ on $j = 1, \ldots, M$ items

Note: “users” and “items” could mean other things too (depends on the data)
Another Example Problem: Matrix Factorization/Completion

Given: Data \( D = \{ r_{ij} \} \) of “interactions” (e.g., ratings) of users \( i = 1, \ldots, N \) on \( j = 1, \ldots, M \) items

Note: “users” and “items” could mean other things too (depends on the data)

\( (i,j) \in \Omega \) denotes an observed user-item pair. \( \Omega \) is the set of all such pairs
Given: Data $D = \{r_{ij}\}$ of “interactions” (e.g., ratings) of users $i = 1, \ldots, N$ on $j = 1, \ldots, M$ items

- Note: “users” and “items” could mean other things too (depends on the data)

$(i, j) \in \Omega$ denotes an observed user-item pair. $\Omega$ is the set of all such pairs

Only a small number of user-item ratings observed, i.e., $|\Omega| \ll NM$
Given: Data $\mathcal{D} = \{r_{ij}\}$ of “interactions” (e.g., ratings) of users $i = 1, \ldots, N$ on $j = 1, \ldots, M$ items
- Note: “users” and “items” could mean other things too (depends on the data)

$(i,j) \in \Omega$ denotes an observed user-item pair. $\Omega$ is the set of all such pairs

Only a small number of user-item ratings observed, i.e., $|\Omega| \ll NM$

We would like to predict the unobserved values $r_{ij} \notin \mathcal{D}$
Let’s call the full matrix $R$ and assume it to be an \textit{approximately low-rank} matrix

$$R = UV^T + E$$
Matrix Completion via Matrix Factorization

Let’s call the full matrix $R$ and assume it to be an approximately low-rank matrix

$$R = UV^\top + E$$

- $U = [u_1 \ldots u_N]^\top$ is $N \times K$ and consists of the latent factors of the $N$ users
  - $u_i \in \mathbb{R}^K$ denotes the latent factors (or learned features) of user $i$

- $V = [v_1 \ldots v_M]^\top$ is $M \times K$ and consists of the latent factors of the $M$ items
  - $v_j \in \mathbb{R}^K$ denotes the latent factors (or learned features) of item $j$

- $E = \{\epsilon_{ij}\}$ consists of the “noise” in $R$ (not captured by the low-rank assumption)

We can write each element of matrix $R$ as $r_{ij} = u_i ^\top v_j + \epsilon_{ij}$ ($i = 1, \ldots, N$, $j = 1, \ldots, M$)
Matrix Completion via Matrix Factorization

Let’s call the full matrix $R$ and assume it to be an approximately low-rank matrix

$$R = UV^\top + E$$

- $U = [u_1 \ldots u_N]^{\top}$ is $N \times K$ and consists of the latent factors of the $N$ users
  - $u_i \in \mathbb{R}^K$ denotes the latent factors (or learned features) of user $i$
- $V = [v_1 \ldots v_M]^{\top}$ is $M \times K$ and consists of the latent factors of the $M$ items
  - $v_j \in \mathbb{R}^K$ denotes the latent factors (or learned features) of item $j$
Matrix Completion via Matrix Factorization

Let’s call the full matrix $\mathbf{R}$ and assume it to be an approximately low-rank matrix

$$
\mathbf{R} = \mathbf{U}\mathbf{V}^\top + \mathbf{E}
$$

- $\mathbf{U} = [\mathbf{u}_1 \ldots \mathbf{u}_N]^\top$ is $N \times K$ and consists of the latent factors of the $N$ users
  - $\mathbf{u}_i \in \mathbb{R}^K$ denotes the latent factors (or learned features) of user $i$
- $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_M]^\top$ is $M \times K$ and consists of the latent factors of the $M$ items
  - $\mathbf{v}_j \in \mathbb{R}^K$ denotes the latent factors (or learned features) of item $j$
- $\mathbf{E} = \{\epsilon_{ij}\}$ consists of the “noise” in $\mathbf{R}$ (not captured by the low-rank assumption)
Matrix Completion via Matrix Factorization

- Let’s call the full matrix $R$ and assume it to be an approximately low-rank matrix

  \[ R = UV^\top + E \]

- $U = [u_1 \ldots u_N]^\top$ is $N \times K$ and consists of the latent factors of the $N$ users
  - $u_i \in \mathbb{R}^K$ denotes the latent factors (or learned features) of user $i$

- $V = [v_1 \ldots v_M]^\top$ is $M \times K$ and consists of the latent factors of the $M$ items
  - $v_j \in \mathbb{R}^K$ denotes the latent factors (or learned features) of item $j$

- $E = \{\epsilon_{ij}\}$ consists of the “noise” in $R$ (not captured by the low-rank assumption)

- We can write each element of matrix $R$ as

  \[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \quad (i = 1, \ldots, N, \quad j = 1, \ldots, M) \]
Matrix Completion via Matrix Factorization

- Let’s call the full matrix $R$ and assume it to be an approximately low-rank matrix
  \[ R = UV^\top + E \]

- $U = [u_1 \ldots u_N]^\top$ is $N \times K$ and consists of the latent factors of the $N$ users
  - $u_i \in \mathbb{R}^K$ denotes the latent factors (or learned features) of user $i$

- $V = [v_1 \ldots v_M]^\top$ is $M \times K$ and consists of the latent factors of the $M$ items
  - $v_j \in \mathbb{R}^K$ denotes the latent factors (or learned features) of item $j$

- $E = \{\epsilon_{ij}\}$ consists of the “noise” in $R$ (not captured by the low-rank assumption)

- We can write each element of matrix $R$ as
  \[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \quad (i = 1, \ldots, N, \quad j = 1, \ldots, M) \]

- Given $u_i$ and $v_j$, any unobserved element in $R$ can be predicted using the above
The low-rank matrix factorization model assumes

\[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \]
The low-rank matrix factorization model assumes

\[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \]

Let’s assume the noise to be Gaussian \( \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij} \mid 0, \beta^{-1}) \)
A Bayesian Model for Matrix Factorization

- The low-rank matrix factorization model assumes

\[ r_{ij} = u_i^T v_j + \epsilon_{ij} \]

- Let’s assume the noise to be Gaussian \( \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1}) \)

- This results in the following Gaussian likelihood for each observation
A Bayesian Model for Matrix Factorization

- The low-rank matrix factorization model assumes
  \[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \]

- Let’s assume the noise to be Gaussian \( \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1}) \)

- This results in the following **Gaussian likelihood** for each observation
  \[ p(r_{ij}|u_i, v_j) = \]
A Bayesian Model for Matrix Factorization

• The low-rank matrix factorization model assumes

\[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \]

• Let’s assume the noise to be Gaussian

\[ \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1}) \]

• This results in the following Gaussian likelihood for each observation

\[ p(r_{ij}|u_i, v_j) = \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1}) \]
A Bayesian Model for Matrix Factorization

- The low-rank matrix factorization model assumes

\[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \]

- Let’s assume the noise to be Gaussian \( \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1}) \)

- This results in the following Gaussian likelihood for each observation

\[ p(r_{ij}|u_i, v_j) = \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1}) \]

- Assume Gaussian priors on the user and item latent factors

\[ p(u_i) = \mathcal{N}(u_i|0, \lambda_u^{-1}I_K) \quad \text{and} \quad p(v_j) = \mathcal{N}(v_j|0, \lambda_v^{-1}I_K) \]
A Bayesian Model for Matrix Factorization

- The low-rank matrix factorization model assumes
  \[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \]
- Let’s assume the noise to be Gaussian \( \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1}) \)
- This results in the following Gaussian likelihood for each observation
  \[ p(r_{ij}|u_i, v_j) = \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1}) \]
- Assume Gaussian priors on the user and item latent factors
  \[ p(u_i) = \mathcal{N}(u_i|0, \lambda_u^{-1} I_K) \quad \text{and} \quad p(v_j) = \mathcal{N}(v_j|0, \lambda_v^{-1} I_K) \]
- The goal is to infer latent factors \( U = \{u_i\}_{i=1}^N \) and \( V = \{v_j\}_{j=1}^M \), given observed ratings from \( R \)
A Bayesian Model for Matrix Factorization

- The low-rank matrix factorization model assumes
  \[ r_{ij} = u_i^\top v_j + \epsilon_{ij} \]

- Let's assume the noise to be Gaussian \( \epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij} | 0, \beta^{-1}) \)

- This results in the following **Gaussian likelihood** for each observation
  \[ p(r_{ij} | u_i, v_j) = \mathcal{N}(r_{ij} | u_i^\top v_j, \beta^{-1}) \]

- Assume **Gaussian priors** on the user and item latent factors
  \[ p(u_i) = \mathcal{N}(u_i | 0, \lambda_u^{-1} I_K) \quad \text{and} \quad p(v_j) = \mathcal{N}(v_j | 0, \lambda_v^{-1} I_K) \]

- The goal is to infer latent factors \( U = \{u_i\}_{i=1}^N \) and \( V = \{v_j\}_{j=1}^M \), given observed ratings from \( R \)

- For simplicity, we will assume the hyperparams \( \beta, \lambda_u, \lambda_v \) to be fixed and not to be learned
Our target posterior distribution for this model will be

\[ p(U, V | R) = \frac{p(R | U, V)p(U, V)}{\int \int p(R | U, V)p(U, V) dU dV} \]

The denominator (and hence the posterior) is intractable! Therefore, the posterior must be approximated somehow. One way to approximate is to compute Conditional Posterior (CP) over individual variables, e.g.,

\[ p(u_i | R, V, U_{-i}) \] and \[ p(v_j | R, U, V_{-j}) \]

\( U_{-i} \) denotes all of \( U \) except \( u_i \). Note: \( V_{-j} \) denotes all of \( V \) except \( v_j \).

Caveat: Each CP should be "computable" (but this is possible for models with "local conjugacy") since CP of each var. depends on all other vars, inference algos based on computing CPs usually work in alternating fashion, until each CP converges (e.g., Gibbs sampling which we’ll look at later).
The Posterior

- Our target posterior distribution for this model will be

\[
p(U, V|R) = \frac{p(R|U, V)p(U, V)}{\int \int p(R|U, V)p(U, V)\,dU\,dV} = \prod_{(i, j) \in \Omega} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) \prod_{i, j} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) \, du_1 \ldots du_N \, dv_1 \ldots dv_M
\]

The denominator (and hence the posterior) is intractable! Therefore, the posterior must be approximated somehow. One way to approximate is to compute Conditional Posterior (CP) over individual variables, e.g., \( p(u_i|\mathcal{R}, V, U_{-i}) \) and \( p(v_j|\mathcal{R}, U, V_{-j}) \). Note: \( U_{-i} \) denotes all of \( U \) except \( u_i \). Note: \( V_{-j} \) denotes all of \( V \) except \( v_j \). Caveat: Each CP should be "computable" (but this is possible for models with "local conjugacy"). Since CP of each var. depends on all other vars, inference algos based on computing CPs usually work in alternating fashion, until each CP converges (e.g., Gibbs sampling which we'll look at later).
The Posterior

Our target posterior distribution for this model will be

$$p(U, V | R) = \frac{p(R | U, V) p(U, V)}{\int \int p(R | U, V) p(U, V) dU dV} = \frac{\prod_{(i,j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j)}{\int \ldots \int \prod_{(i,j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) du_1 \ldots du_N dv_1 \ldots dv_M}$$

The denominator (and hence the posterior) is intractable!
The Posterior

- Our target posterior distribution for this model will be

\[ p(U, V|R) = \frac{p(R|U, V)p(U, V)}{\int \int p(R|U, V)p(U, V)du dv} = \frac{\prod_{(i, j) \in \Omega} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j)}{\int \cdots \int \prod_{(i, j) \in \Omega} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) du_1 \cdots du_N dv_1 \cdots dv_M} \]

- The denominator (and hence the posterior) is intractable!
- Therefore, the posterior must be approximated somehow
The Posterior

Our target posterior distribution for this model will be

\[
p(U, V | R) = \frac{p(R | U, V) p(U, V)}{\int \int p(R | U, V) p(U, V) dU dV} = \frac{\prod_{(i, j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j)}{\int \cdots \int \prod_{(i, j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) d_{u_1} \cdots d_{u_N} d_{v_1} \cdots d_{v_M}}
\]

The denominator (and hence the posterior) is intractable!

Therefore, the posterior must be approximated somehow.

One way to approximate is to compute **Conditional Posterior (CP)** over individual variables, e.g.,

\[
p(u_i | R, V, U_{-i}) \quad \text{and} \quad p(v_j | R, U, V_{-j})
\]
Our target posterior distribution for this model will be

\[
p(U, V | R) = \frac{p(R|U, V)p(U, V)}{\int \int p(R|U, V)p(U, V) dU dV} = \frac{\prod_{(i, j) \in \Omega} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j)}{\int \ldots \int \prod_{(i, j) \in \Omega} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) du_1 \ldots du_N dv_1 \ldots dv_M}
\]

The denominator (and hence the posterior) is intractable!

Therefore, the posterior must be approximated somehow

One way to approximate is to compute Conditional Posterior (CP) over individual variables, e.g.,

\[
p(u_i|R, V, U_{-i}) \quad \text{and} \quad p(v_j|R, U, V_{-j})
\]

\(U_{-i}\) denotes all of \(U\) except \(u_i\). Note: \(V, U_{-i}\) is the set of all unknowns except \(u_i\)
The Posterior

Our target posterior distribution for this model will be

\[ p(U, V | R) = \frac{p(R | U, V) p(U, V)}{\int \int p(R | U, V) p(U, V) dU dV} = \frac{\prod_{(i,j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j)}{\int \cdots \int \prod_{(i,j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) du_1 \cdots du_N dv_1 \cdots dv_M} \]

The denominator (and hence the posterior) is intractable!

Therefore, the posterior must be approximated somehow

One way to approximate is to compute Conditional Posterior (CP) over individual variables, e.g.,

\[ p(u_i | R, V, U_{-i}) \text{ and } p(v_j | R, U, V_{-j}) \]

\( U_{-i} \) denotes all of \( U \) except \( u_i \). Note: \( V, U_{-i} \) is the set of all unknowns except \( u_i \)

\( V_{-j} \) denotes all of \( V \) except \( v_j \). Note: \( U, V_{-j} \) is the set of all unknowns except \( v_j \)
The Posterior

Our target posterior distribution for this model will be

\[
p(U, V|R) = \frac{p(R|U, V)p(U, V)}{\int \int p(R|U, V)p(U, V)dUdV} = \frac{\prod_{(i,j) \in \Omega} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j)}{\int \ldots \int \prod_{(i,j) \in \Omega} p(r_{ij}|u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) dU_1 \ldots dU_N dV_1 \ldots dV_M}
\]

The denominator (and hence the posterior) is intractable!

Therefore, the posterior must be approximated somehow

One way to approximate is to compute Conditional Posterior (CP) over individual variables, e.g.,

\[
p(u_i|R, V, U_{-i}) \quad \text{and} \quad p(v_j|R, U, V_{-j})
\]

\(U_{-i}\) denotes all of \(U\) except \(u_i\). Note: \(V, U_{-i}\) is the set of all unknowns except \(u_i\)

\(V_{-j}\) denotes all of \(V\) except \(v_j\). Note: \(U, V_{-j}\) is the set of all unknowns except \(v_j\)

Caveat: Each CP should be “computable” (but this is possible for models with “local conjugacy”)
Our target posterior distribution for this model will be

\[ p(U, V | R) = \frac{p(R | U, V) p(U, V)}{\int \int p(R | U, V) p(U, V) dU dV} = \frac{\prod_{(i,j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j)}{\int \dots \int \prod_{(i,j) \in \Omega} p(r_{ij} | u_i, v_j) \prod_i p(u_i) \prod_j p(v_j) dU_1 \ldots dU_N dV_1 \ldots dV_M} \]

The denominator (and hence the posterior) is intractable!

Therefore, the posterior must be approximated somehow.

One way to approximate is to compute \textit{Conditional Posterior} (CP) over individual variables, e.g.,

\[ p(u_i | R, V, U_{-i}) \text{ and } p(v_j | R, U, V_{-j}) \]

\(U_{-i}\) denotes all of \(U\) except \(u_i\). Note: \(V, U_{-i}\) is the set of all unknowns except \(u_i\)

\(V_{-j}\) denotes all of \(V\) except \(v_j\). Note: \(U, V_{-j}\) is the set of all unknowns except \(v_j\)

\textbf{Caveat:} Each CP should be “computable” (but this is possible for models with “local conjugacy”)

Since CP of each var. depends on all other vars, inference algos based on computing CPs usually work in \textit{alternating fashion}, until each CP converges (e.g., Gibbs sampling which we’ll look at later)
Conditional Posterior and Local Conjugacy
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy.
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data $\mathbf{X}$ and $K$ unknown params/hyperparams $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits **local conjugacy**
- Note: Some researchers also call CP as **Complete Conditional** or **Local Posterior**

Consider a general model with data \( X \) and \( K \) unknown params/hyperparams \( \Theta = (\theta_1, \theta_2, \ldots, \theta_K) \)

- Suppose posterior \( p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)} \) is intractable (will be the case if \( p(\Theta) \) isn't conjugate)
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data $X$ and $K$ unknown params/hyperparams $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$

- Suppose posterior $p(\Theta | X) = \frac{p(X | \Theta)p(\Theta)}{p(X)}$ is intractable (will be the case if $p(\Theta)$ isn’t conjugate)

- However suppose we can compute the following conditional posterior tractably

$$p(\theta_k | X, \Theta_{-k}) = \frac{p(X | \theta_k, \Theta_{-k})p(\theta_k)}{\int p(X | \theta_k, \Theta_{-k})p(\theta_k)d\theta_k}$$
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data $X$ and $K$ unknown params/hyperparams $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$

- Suppose posterior $p(\Theta | X) = \frac{p(X | \Theta)p(\Theta)}{p(X)}$ is intractable (will be the case if $p(\Theta)$ isn't conjugate)

- However suppose we can compute the following conditional posterior tractably

$$p(\theta_k | X, \Theta_{-k}) = \frac{p(X | \theta_k, \Theta_{-k})p(\theta_k)}{\int p(X | \theta_k, \Theta_{-k})p(\theta_k)d\theta_k} \propto p(X | \theta_k, \Theta_{-k})p(\theta_k)$$

.. which would be possible if $p(X | \theta_k, \Theta_{-k})$ and $p(\theta_k)$ are conjugate to each other
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data $\mathbf{X}$ and $K$ unknown params/hyperparams $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$

- Suppose posterior $p(\Theta | \mathbf{X}) = \frac{p(\mathbf{X} | \Theta)p(\Theta)}{p(\mathbf{X})}$ is intractable (will be the case if $p(\Theta)$ isn’t conjugate)

- However suppose we can compute the following conditional posterior tractably

$$p(\theta_k | \mathbf{X}, \Theta_{-k}) = \frac{p(\mathbf{X} | \theta_k, \Theta_{-k})p(\theta_k)}{\int p(\mathbf{X} | \theta_k, \Theta_{-k})p(\theta_k) d\theta_k} \propto p(\mathbf{X} | \theta_k, \Theta_{-k})p(\theta_k)$$

.. which would be possible if $p(\mathbf{X} | \theta_k, \Theta_{-k})$ and $p(\theta_k)$ are conjugate to each other

- Such models are called “locally conjugate” models
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data \( X \) and \( K \) unknown params/hyperparams \( \Theta = (\theta_1, \theta_2, \ldots, \theta_K) \)

- Suppose posterior \( p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)} \) is intractable (will be the case if \( p(\Theta) \) isn’t conjugate)

- However suppose we can compute the following conditional posterior tractably
  \[
p(\theta_k|X, \Theta_{-k}) = \frac{p(X|\theta_k, \Theta_{-k})p(\theta_k)}{\int p(X|\theta_k, \Theta_{-k})p(\theta_k)\,d\theta_k} \propto p(X|\theta_k, \Theta_{-k})p(\theta_k)
  \]
  .. which would be possible if \( p(X|\theta_k, \Theta_{-k}) \) and \( p(\theta_k) \) are conjugate to each other

- Such models are called “locally conjugate” models

- Important: In the above context, when considering the likelihood \( p(X|\theta_k, \Theta_{-k}) \)
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data $\mathbf{X}$ and $K$ unknown params/hyperparams $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$

- Suppose posterior $p(\Theta | \mathbf{X}) = \frac{p(\mathbf{X} | \Theta)p(\Theta)}{p(\mathbf{X})}$ is intractable (will be the case if $p(\Theta)$ isn't conjugate)

- However suppose we can compute the following conditional posterior tractably

  $$p(\theta_k | \mathbf{X}, \Theta_{-k}) = \frac{p(\mathbf{X} | \theta_k, \Theta_{-k}) p(\theta_k)}{\int p(\mathbf{X} | \theta_k, \Theta_{-k}) p(\theta_k) d\theta_k} \propto p(\mathbf{X} | \theta_k, \Theta_{-k}) p(\theta_k)$$

  .. which would be possible if $p(\mathbf{X} | \theta_k, \Theta_{-k})$ and $p(\theta_k)$ are conjugate to each other

- Such models are called “locally conjugate” models

- Important: In the above context, when considering the likelihood $p(\mathbf{X} | \theta_k, \Theta_{-k})$
  - $\mathbf{X}$ actually refers to only that part of data $\mathbf{X}$ that depends on $\theta_k$
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data $X$ and $K$ unknown params/hyperparams $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$

- Suppose posterior $p(\Theta | X) = \frac{p(X|\Theta)p(\Theta)}{p(X)}$ is intractable (will be the case if $p(\Theta)$ isn't conjugate)

- However suppose we can compute the following conditional posterior tractably

  $$p(\theta_k | X, \Theta_{-k}) = \frac{p(X|\theta_k, \Theta_{-k})p(\theta_k)}{\int p(X|\theta_k, \Theta_{-k})p(\theta_k)d\theta_k} \propto p(X|\theta_k, \Theta_{-k})p(\theta_k)$$

  .. which would be possible if $p(X|\theta_k, \Theta_{-k})$ and $p(\theta_k)$ are conjugate to each other

- Such models are called "locally conjugate" models

- **Important:** In the above context, when considering the likelihood $p(X|\theta_k, \Theta_{-k})$
  - $X$ actually refers to only that part of data $X$ that depends on $\theta_k$
  - $\Theta_{-k}$ refers to only those unknowns that "interact" with $\theta_k$ in generating that part of data
Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute if the model admits local conjugacy
  - Note: Some researchers also call CP as Complete Conditional or Local Posterior

- Consider a general model with data $X$ and $K$ unknown params/hyperparams $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$

- Suppose posterior $p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)}$ is intractable (will be the case if $p(\Theta)$ isn't conjugate)

- However suppose we can compute the following conditional posterior tractably

$$p(\theta_k|X, \Theta_{-k}) = \frac{p(X|\theta_k, \Theta_{-k})p(\theta_k)}{\int p(X|\theta_k, \Theta_{-k})p(\theta_k)d\theta_k} \propto p(X|\theta_k, \Theta_{-k})p(\theta_k)$$

  .. which would be possible if $p(X|\theta_k, \Theta_{-k})$ and $p(\theta_k)$ are conjugate to each other

- Such models are called “locally conjugate” models

- Important: In the above context, when considering the likelihood $p(X|\theta_k, \Theta_{-k})$
  - $X$ actually refers to only that part of data $X$ that depends on $\theta_k$
  - $\Theta_{-k}$ refers to only those unknowns that “interact” with $\theta_k$ in generating that part of data
With the conditional posterior based approximation, the target posterior

\[ p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)} \]

.. is represented by several conditional posteriors \( p(\theta_k|X, \Theta_{-k}), k = 1, \ldots, K \)
Representation of Posterior

- With the conditional posterior based approximation, the target posterior

\[ p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)} \]

.. is represented by several conditional posteriors \( p(\theta_k|X, \Theta_{-k}), k = 1, \ldots, K \)

- Each of the conditional posterior is a distribution over one unknown \( \theta_k \), given all other unknowns
With the conditional posterior based approximation, the target posterior

\[ p(\Theta | X) = \frac{p(X | \Theta)p(\Theta)}{p(X)} \]

.. is represented by several conditional posteriors \( p(\theta_k | X, \Theta_{-k}) \), \( k = 1, \ldots, K \)

Each of the conditional posterior is a distribution over one unknown \( \theta_k \), given all other unknowns

Need a way to “combine” these CPs to get the overall posterior
With the conditional posterior based approximation, the target posterior

\[ p(\Theta | X) = \frac{p(X | \Theta) p(\Theta)}{p(X)} \]

is represented by several conditional posteriors \( p(\theta_k | X, \Theta_{-k}), k = 1, \ldots, K \).

Each of the conditional posterior is a distribution over one unknown \( \theta_k \), given all other unknowns.

Need a way to “combine” these CPs to get the overall posterior.

One way to get the overall representation of the posterior can be using sampling based inference algorithms like Gibbs sampling or MCMC (more on this later).
Detour: Gibbs Sampling (Geman and Geman, 1982)

- A general sampling algorithm to simulate samples from multivariate distributions
Detour: Gibbs Sampling (Geman and Geman, 1982)

- A general sampling algorithm to simulate samples from multivariate distributions
- Samples one component at a time from its conditional, conditioned on all other components
Detour: Gibbs Sampling (Geman and Geman, 1982)

- A general **sampling algorithm** to simulate samples from multivariate distributions
- Samples one component at a time from its conditional, conditioned on all other components
  - Assumes that the conditional distributions are available in a closed form
Detour: Gibbs Sampling (Geman and Geman, 1982)

- A general **sampling algorithm** to simulate samples from multivariate distributions
- Samples one component at a time from its conditional, conditioned on all other components
  - Assumes that the conditional distributions are available in a closed form

Suppose
\[
\theta \sim N_2(0, \Sigma) \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.
\]

Then
\[
\begin{align*}
\theta_1 | \theta_2 & \sim N(\rho \theta_2, [1 - \rho^2]) \\
\theta_2 | \theta_1 & \sim N(\rho \theta_1, [1 - \rho^2])
\end{align*}
\]

are the conditional distributions.
Detour: Gibbs Sampling (Geman and Geman, 1982)

- A general sampling algorithm to simulate samples from multivariate distributions
- Samples one component at a time from its conditional, conditioned on all other components
  - Assumes that the conditional distributions are available in a closed form

\[
\begin{align*}
\theta &\sim N_2(0, \Sigma) \\
\Sigma &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.
\end{align*}
\]

Then

\[
\begin{align*}
\theta_1|\theta_2 &\sim N(\rho \theta_2, [1 - \rho^2]) \\
\theta_2|\theta_1 &\sim N(\rho \theta_1, [1 - \rho^2])
\end{align*}
\]

are the conditional distributions.

- The generated samples give a sample-based approximation of the multivariate distribution
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a sampling-based approximation of a multiparameter posterior distribution

Gibbs sampler iteratively draws random samples from conditional posteriors. When run long enough, the sampler produces samples from the joint posterior.

For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibbs sampler looks like this:

1. Initialize $\theta(0)$
2. For $s = 1, \ldots, S$
   - Draw a random sample for $\theta_1$ as $\theta_1(s) \sim p(\theta_1 | \theta_2(s-1), y)$
   - Draw a random sample for $\theta_2$ as $\theta_2(s) \sim p(\theta_2 | \theta_1(s), y)$

The set of $S$ random samples $\{\theta_1(s), \theta_2(s)\}_{s=1}^S$ represent the joint posterior distribution $p(\theta_1, \theta_2 | y)$.

More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea).
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a **sampling-based approximation** of a multiparameter posterior distribution.
- Gibbs sampler iteratively draws random samples from conditional posteriors.

When run long enough, the sampler produces samples from the joint posterior. For the simple two-parameter case \( \theta = (\theta_1, \theta_2) \), the Gibbs sampler looks like this:

1. Initialize \( \theta(0) \).
2. For \( s = 1, \ldots, S \):
   - Draw a random sample for \( \theta_1 \) as \( \theta_1(s) \sim p(\theta_1 | \theta_1(s-1), y) \).
   - Draw a random sample for \( \theta_2 \) as \( \theta_2(s) \sim p(\theta_2 | \theta_2(s), y) \).

The set of \( S \) random samples \( \{\theta_1(s), \theta_2(s)\}_{s=1}^S \) represents the joint posterior distribution \( p(\theta_1, \theta_2 | y) \).

More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea).
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a **sampling-based approximation** of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior

For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibbs sampler looks like this:

1. Initialize $\theta(0)$
2. For $s = 1, \ldots, S$
   - Draw a random sample for $\theta_1$ as $\theta_1 (s) \sim p(\theta_1 | \theta_{1-(s-1)}, y)$
   - Draw a random sample for $\theta_2$ as $\theta_2 (s) \sim p(\theta_2 | \theta_1 (s), y)$

The set of $S$ random samples $\{\theta_1 (s), \theta_2 (s)\}_{s=1}^S$ represent the joint posterior distribution $p(\theta_1, \theta_2 | y)$.

More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea).
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a **sampling-based approximation** of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibb sampler looks like this

\[
\text{Initialize } \theta^{(0)}_1, \theta^{(0)}_2 \\
\text{For } s = 1, \ldots, S \\
\quad \text{Draw a random sample for } \theta_1 \text{ as } \theta_1^{(s)} \sim p(\theta_1 | \theta^{(s-1)}_1, y) \\
\quad \text{Draw a random sample for } \theta_2 \text{ as } \theta_2^{(s)} \sim p(\theta_2 | \theta^{(s)}_1, y) \\
\text{The set of } S \text{ random samples } \{\theta_1^{(s)}, \theta_2^{(s)}\}_{s=1}^S \text{ represent the joint posterior distribution } p(\theta_1, \theta_2 | y) \\
\]

More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea)
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a **sampling-based approximation** of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibb sampler looks like this
  - Initialize $\theta_2^{(0)}$
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a sampling-based approximation of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibb sampler looks like this
  - Initialize $\theta_2^{(0)}$
  - For $s = 1, \ldots, S$
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a sampling-based approximation of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibbs sampler looks like this
  - Initialize $\theta_2^{(0)}$
  - For $s = 1, \ldots, S$
    - Draw a random sample for $\theta_1$ as $\theta_1^{(s)} \sim p(\theta_1 | \theta_2^{(s-1)}, y)$

The set of $S$ random samples $\{\theta_1^{(s)}, \theta_2^{(s)}\}_{s=1}^S$ represent the joint posterior distribution $p(\theta_1, \theta_2 | y)$

More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea)
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a sampling-based approximation of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional Posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case \( \theta = (\theta_1, \theta_2) \), the Gibb sampler looks like this
  - Initialize \( \theta_2^{(0)} \)
  - For \( s = 1, \ldots, S \)
    - Draw a random sample for \( \theta_1 \) as \( \theta_1^{(s)} \sim p(\theta_1|\theta_2^{(s-1)}, y) \)
    - Draw a random sample for \( \theta_2 \) as \( \theta_2^{(s)} \sim p(\theta_2|\theta_1^{(s)}, y) \)

The set of \( S \) random samples \( \{\theta_1^{(s)}, \theta_2^{(s)}\}_{s=1}^S \) represent the joint posterior distribution \( p(\theta_1, \theta_2 | y) \)

More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea)
Can be used to get a sampling-based approximation of a multiparameter posterior distribution.

Gibbs sampler iteratively draws random samples from conditional posteriors.

When run long enough, the sampler produces samples from the joint posterior.

For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibb sampler looks like this:

- Initialize $\theta_2^{(0)}$.
- For $s = 1, \ldots, S$:
  - Draw a random sample for $\theta_1$ as $\theta_1^{(s)} \sim p(\theta_1 | \theta_2^{(s-1)}, y)$.
  - Draw a random sample for $\theta_2$ as $\theta_2^{(s)} \sim p(\theta_2 | \theta_1^{(s)}, y)$.

The set of $S$ random samples $\{\theta_1^{(s)}, \theta_2^{(s)}\}_{s=1}^S$ represent the joint posterior distribution $p(\theta_1, \theta_2 | y)$. 
Detour: Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a **sampling-based approximation** of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case \( \theta = (\theta_1, \theta_2) \), the Gibb sampler looks like this
  - Initialize \( \theta_2^{(0)} \)
  - For \( s = 1, \ldots, S \)
    - Draw a random sample for \( \theta_1 \) as \( \theta_1^{(s)} \sim p(\theta_1|\theta_2^{(s-1)}, y) \)
    - Draw a random sample for \( \theta_2 \) as \( \theta_2^{(s)} \sim p(\theta_2|\theta_1^{(s)}, y) \)

The set of \( S \) random samples \( \{\theta_1^{(s)}, \theta_2^{(s)}\}_{s=1}^S \) represent the joint posterior distribution \( p(\theta_1, \theta_2|y) \)

- More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea)
Back to Bayesian Matrix Factorization..
The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy.

To see this, note that the conditional posterior for user latent factor $u_i$ is:

$$p(u_i | R, V, U_{-i}) \propto \prod_{j: (i, j) \in \Omega} p(r_{ij} | u_i, v_j) p(u_i)$$

Note: the posterior of $u_i$ doesn't actually depend on $U_{-i}$ and rows of $R$ except row $i$.

After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is:

$$p(u_i | R, V) \propto \prod_{j: (i, j) \in \Omega} N(r_{ij} | u_i^\top v_j, \beta^{-1}) N(u_i | 0, \lambda^{-1} u_I K)$$

Since $V$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression.

Linear regression analogy:

- $\{v_j\}_j: (i, j) \in \Omega$: inputs,
- $\{r_{ij}\}_j: (i, j) \in \Omega$: responses,
- $u_i$: unknown weight vector

Likewise, the conditional posterior of $v_j$ will be:

$$p(v_j | R, U) \propto \prod_{i: (i, j) \in \Omega} N(r_{ij} | u_i^\top v_j, \beta^{-1}) N(v_j | 0, \lambda^{-1} v_I K)$$

.. like Bayesian lin. reg. with

- $\{u_i\}_i: (i, j) \in \Omega$: inputs,
- $\{r_{ij}\}_i: (i, j) \in \Omega$: responses,
- $v_j$: unknown weight vec
Bayesian Matrix Factorization: Conditional Posteriors

- The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy
- To see this, note that the conditional posterior for user latent factor $u_i$ is

$$p(u_i | R, V, U_{−i}) \propto \prod_{j : (i, j) \in \Omega} p(r_{ij} | u_i, v_j) p(u_i)$$

Note: the posterior of $u_i$ doesn't actually depend on $U_{−i}$ and rows of $R$ except row $i$.

After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is

$$p(u_i | R, V) \propto \prod_{j : (i, j) \in \Omega} N(r_{ij} | u_i^\top v_j, \beta_{−1}) N(u_i | 0, \lambda_{−1} u_I K)$$

Since $V$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression.

Linear regression analogy:

- $\{v_j\}_{j : (i, j) \in \Omega}$: inputs,
- $\{r_{ij}\}_{j : (i, j) \in \Omega}$: responses,
- $u_i$: unknown weight vector

Likewise, the conditional posterior of $v_j$ will be

$$p(v_j | R, U) \propto \prod_{i : (i, j) \in \Omega} N(r_{ij} | u_i^\top v_j, \beta_{−1}) N(v_j | 0, \lambda_{−1} v_I K)$$

.. like Bayesian lin. reg. with

- $\{u_i\}_{i : (i, j) \in \Omega}$: inputs,
- $\{r_{ij}\}_{i : (i, j) \in \Omega}$: responses,
- $v_j$: unknown weight vec
Bayesian Matrix Factorization: Conditional Posteriors

- The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy.
- To see this, note that the conditional posterior for user latent factor $u_i$ is

$$p(u_i | R, V, U_{-i}) \propto \prod_{j : (i,j) \in \Omega} p(r_{ij} | u_i, v_j) p(u_i)$$

Note: the posterior of $u_i$ doesn’t actually depend on $U_{-i}$ and rows of $R$ except row $i$.

After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is

$$p(u_i | R, V) \propto \prod_{j : (i,j) \in \Omega} N(r_{ij} | u_i^\top v_j, \beta^{-1}) N(u_i | 0, \lambda^{-1} u I_K)$$

Since $V$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression.

Linear regression analogy:

- $\{v_j\}_{j : (i,j) \in \Omega}$: inputs,
- $\{r_{ij}\}_{j : (i,j) \in \Omega}$: responses,
- $u_i$: unknown weight vector

Likewise, the conditional posterior of $v_j$ will be

$$p(v_j | R, U) \propto \prod_{i : (i,j) \in \Omega} N(r_{ij} | u_i^\top v_j, \beta^{-1}) N(v_j | 0, \lambda^{-1} v I_K)$$

.. like Bayesian lin. reg. with

- $\{u_i\}_{i : (i,j) \in \Omega}$: inputs,
- $\{r_{ij}\}_{i : (i,j) \in \Omega}$: responses,
- $v_j$: unknown weight vec.
The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy.

To see this, note that the conditional posterior for user latent factor $u_i$ is

$$p(u_i | R, V, U_{-i}) \propto \prod_{j: (i,j) \in \Omega} p(r_{ij} | u_i, v_j) p(u_i)$$

Note: the posterior of $u_i$ doesn’t actually depend on $U_{-i}$ and rows of $R$ except row $i$.

After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is
Bayesian Matrix Factorization: Conditional Posteriors

- The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy.
- To see this, note that the conditional posterior for user latent factor $u_i$ is

$$p(u_i|R, V, U_{-i}) \propto \prod_{j: (i,j) \in \Omega} p(r_{ij}|u_i, v_j)p(u_i)$$

- Note: the posterior of $u_i$ doesn’t actually depend on $U_{-i}$ and rows of $R$ except row $i$.
- After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is

$$p(u_i|R, V) \propto \prod_{j: (i,j) \in \Omega} \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1}) \mathcal{N}(u_i|0, \lambda_u^{-1}I_K)$$
Bayesian Matrix Factorization: Conditional Posteriors

- The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy.
- To see this, note that the conditional posterior for user latent factor $u_i$ is

$$p(u_i|\mathbf{R}, \mathbf{V}, U_{-i}) \propto \prod_{j:(i,j) \in \Omega} p(r_{ij}|u_i, v_j)p(u_i)$$

- Note: the posterior of $u_i$ doesn't actually depend on $U_{-i}$ and rows of $\mathbf{R}$ except row $i$.
- After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is

$$p(u_i|\mathbf{R}, \mathbf{V}) \propto \prod_{j:(i,j) \in \Omega} \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1})\mathcal{N}(u_i|0, \lambda_u^{-1}I_K)$$

- Since $\mathbf{V}$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression.
Bayesian Matrix Factorization: Conditional Posteriors

- The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy
- To see this, note that the conditional posterior for user latent factor $u_i$ is

$$p(u_i|R, V, U_{-i}) \propto \prod_{j:(i,j)\in \Omega} p(r_{ij}|u_i, v_j)p(u_i)$$

- Note: the posterior of $u_i$ doesn’t actually depend on $U_{-i}$ and rows of $R$ except row $i$
- After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is

$$p(u_i|R, V) \propto \prod_{j:(i,j)\in \Omega} N(r_{ij}|u_i^\top v_j, \beta^{-1}) N(u_i|0, \lambda_u^{-1}I_K)$$

- Since $V$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression
  - Linear regression analogy: $\{v_j\}_{j:(i,j)\in \Omega}$: inputs, $\{r_{ij}\}_{j:(i,j)\in \Omega}$: responses, $u_i$: unknown weight vector
Bayesian Matrix Factorization: Conditional Posteriors

- The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy.
- To see this, note that the conditional posterior for user latent factor $u_i$ is
  \[ p(u_i|R, V, U_{-i}) \propto \prod_{j:(i,j)\in\Omega} p(r_{ij}|u_i, v_j)p(u_i) \]

  - Note: the posterior of $u_i$ doesn't actually depend on $U_{-i}$ and rows of $R$ except row $i$.
  - After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is
    \[ p(u_i|R, V) \propto \prod_{j:(i,j)\in\Omega} \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1})\mathcal{N}(u_i|0, \lambda_u^{-1}I_K) \]

- Since $V$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression.
  - Linear regression analogy: $\{v_j\}_{j:(i,j)\in\Omega}$: inputs, $\{r_{ij}\}_{j:(i,j)\in\Omega}$: responses, $u_i$: unknown weight vector.
- Likewise, the conditional posterior of $v_j$ will be
Bayesian Matrix Factorization: Conditional Posteriors

- The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy.
- To see this, note that the conditional posterior for user latent factor $u_i$ is
  \[ p(u_i|R, V, U_{-i}) \propto \prod_{j:(i,j) \in \Omega} p(r_{ij}|u_i, v_j)p(u_i) \]
- Note: the posterior of $u_i$ doesn’t actually depend on $U_{-i}$ and rows of $R$ except row $i$.
- After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is
  \[ p(u_i|R, V) \propto \prod_{j:(i,j) \in \Omega} \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1})\mathcal{N}(u_i|0, \lambda_u^{-1}I_K) \]
- Since $V$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression.
  - Linear regression analogy: $\{v_j\}_{j:(i,j) \in \Omega}$: inputs, $\{r_{ij}\}_{j:(i,j) \in \Omega}$: responses, $u_i$: unknown weight vector.
- Likewise, the conditional posterior of $v_j$ will be
  \[ p(v_j|R, U) \propto \prod_{i:(i,j) \in \Omega} \mathcal{N}(r_{ij}|u_i^\top v_j, \beta^{-1})\mathcal{N}(v_j|0, \lambda_v^{-1}I_K) \]
The BMF model with Gaussian likelihood and Gaussian prior has local conjugacy. To see this, note that the conditional posterior for user latent factor $u_i$ is

$$p(u_i | R, V, U_{-i}) \propto \prod_{j: (i,j) \in \Omega} p(r_{ij} | u_i, v_j) p(u_i)$$

Note: the posterior of $u_i$ doesn’t actually depend on $U_{-i}$ and rows of $R$ except row $i$

After substituting the likelihood and prior (both Gaussians), the conditional posterior of $u_i$ is

$$p(u_i | R, V) \propto \prod_{j: (i,j) \in \Omega} \mathcal{N}(r_{ij} | u_i^\top v_j, \beta^{-1}) \mathcal{N}(u_i | 0, \lambda_u^{-1} I_K)$$

Since $V$ fixed (remember we are computing conditional posteriors alternating fashion), the likelihood and prior are conjugate. This is just like Bayesian linear regression

- Linear regression analogy: \{v_j\}_{j: (i,j) \in \Omega}: inputs, \{r_{ij}\}_{j: (i,j) \in \Omega}: responses, $u_i$: unknown weight vector

Likewise, the conditional posterior of $v_j$ will be

$$p(v_j | R, U) \propto \prod_{i: (i,j) \in \Omega} \mathcal{N}(r_{ij} | u_i^\top v_j, \beta^{-1}) \mathcal{N}(v_j | 0, \lambda_v^{-1} I_K)$$

.. like Bayesian lin. reg. with \{u_i\}_{i: (i,j) \in \Omega}: inputs, \{r_{ij}\}_{i: (i,j) \in \Omega}: responses, $v_j$: unknown weight vec
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression

For each $u_i$, its conditional posterior, given $V$ and ratings $p(u_i|\mathbf{R}, V) = N(\mu_{ui}, \Sigma_{ui})$

where

$$\Sigma_{ui} = (\lambda_u I + \beta \sum_{j: (i, j) \in \Omega} v_j v_j^\top)^{-1}$$

and

$$\mu_{ui} = \Sigma_{ui} (\beta \sum_{j: (i, j) \in \Omega} r_{ij} v_j)$$

For each $v_j$, its conditional posterior, given $U$ and ratings $p(v_j|\mathbf{R}, U) = N(\mu_{vj}, \Sigma_{vj})$

where

$$\Sigma_{vj} = (\lambda_v I + \beta \sum_{i: (i, j) \in \Omega} u_i u_i^\top)^{-1}$$

and

$$\mu_{vj} = \Sigma_{vj} (\beta \sum_{i: (i, j) \in \Omega} r_{ij} u_i)$$

These conditional posteriors can be updated in an alternating fashion until convergence.

This can be implemented using a Gibbs sampler.

Note: Hyperparameters can also be inferred by computing their conditional posteriors (also see "Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo" by Salakhutdinov and Mnih (2008)).

Can extend Gaussian BMF easily to other exp. family distr. while maintaining local conjugacy.
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression.
- For each $u_i$, its conditional posterior, given $V$ and ratings $\mathbf{R}$,

\[
p(u_i | \mathbf{R}, V) = \mathcal{N}(u_i | \mu_{u_i}, \Sigma_{u_i})
\]

where $\Sigma_{u_i} = (\lambda_u I + \beta \sum_{j: (i, j) \in \Omega} v_j v_j^\top)^{-1}$ and $\mu_{u_i} = \Sigma_{u_i} (\beta \sum_{j: (i, j) \in \Omega} r_{ij} v_j)$.

These conditional posteriors can be updated in an alternating fashion until convergence. This can be implemented using a Gibbs sampler. Note: Hyperparameters can also be inferred by computing their conditional posteriors (also see "Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo" by Salakhutdinov and Mnih (2008)).

Can extend Gaussian BMF easily to other exponential family distributions while maintaining local conjugacy.
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression.
- For each $u_i$, its conditional posterior, given $V$ and ratings

$$p(u_i | R, V) = \mathcal{N}(u_i | \mu_{ui}, \Sigma_{ui})$$

where $\Sigma_{ui} = (\lambda_u I + \beta \sum_{j: (i,j) \in \Omega} v_j v_j^\top)^{-1}$ and $\mu_{ui} = \Sigma_{ui} (\beta \sum_{j: (i,j) \in \Omega} r_{ij} v_j)$.
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression.
- For each $u_i$, its conditional posterior, given $V$ and ratings:

$$p(u_i|R, V) = \mathcal{N}(u_i|\mu_{u_i}, \Sigma_{u_i})$$

where $\Sigma_{u_i} = (\lambda_u I + \beta \sum_{j: (i,j) \in \Omega} v_j v_j^\top)^{-1}$ and $\mu_{u_i} = \Sigma_{u_i} (\beta \sum_{j: (i,j) \in \Omega} r_{ij} v_j)$

- For each $v_j$, its conditional posterior, given $U$ and ratings:

$$p(v_j|R, U) = \mathcal{N}(v_j|\mu_{v_j}, \Sigma_{v_j})$$

where $\Sigma_{v_j} = (\lambda_v I + \beta \sum_{i: (i,j) \in \Omega} u_i u_i^\top)^{-1}$ and $\mu_{v_j} = \Sigma_{v_j} (\beta \sum_{i: (i,j) \in \Omega} r_{ij} u_i)$

These conditional posteriors can be updated in an alternating fashion until convergence. This can be be implemented using a Gibbs sampler. Note: Hyperparameters can also be inferred by computing their conditional posteriors (also see "Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo" by Salakhutdinov and Mnih (2008)). Can extend Gaussian BMF easily to other exp. family distr. while maintaining local conjugacy.
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression.
- For each $u_i$, its conditional posterior, given $V$ and ratings
  \[ p(u_i|R, V) = \mathcal{N}(u_i|\mu_{u_i}, \Sigma_{u_i}) \]
  where $\Sigma_{u_i} = (\lambda_u I + \beta \sum_{j:(i,j) \in \Omega} v_j v_j^T)^{-1}$ and $\mu_{u_i} = \Sigma_{u_i} (\beta \sum_{j:(i,j) \in \Omega} r_{ij} v_j)$

- For each $v_j$, its conditional posterior, given $U$ and ratings
  \[ p(v_j|R, U) = \mathcal{N}(v_j|\mu_{v_j}, \Sigma_{v_j}) \]
  where $\Sigma_{v_j} = (\lambda_v I + \beta \sum_{i:(i,j) \in \Omega} u_i u_i^T)^{-1}$ and $\mu_{v_j} = \Sigma_{v_j} (\beta \sum_{i:(i,j) \in \Omega} r_{ij} u_i)$

- These conditional posteriors can be updated in an alternating fashion until convergence.
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression

- For each $u_i$, its conditional posterior, given $V$ and ratings

  \[ p(u_i|R, V) = \mathcal{N}(u_i | \mu_{u_i}, \Sigma_{u_i}) \]

  where $\Sigma_{u_i} = (\lambda_u I + \beta \sum_{j: (i,j) \in \Omega} v_j v_j^T)^{-1}$ and $\mu_{u_i} = \Sigma_{u_i} (\beta \sum_{j: (i,j) \in \Omega} r_{ij} v_j)$

- For each $v_j$, its conditional posterior, given $U$ and ratings

  \[ p(v_j|R, U) = \mathcal{N}(v_j | \mu_{v_j}, \Sigma_{v_j}) \]

  where $\Sigma_{v_j} = (\lambda_v I + \beta \sum_{i: (i,j) \in \Omega} u_i u_i^T)^{-1}$ and $\mu_{v_j} = \Sigma_{v_j} (\beta \sum_{i: (i,j) \in \Omega} r_{ij} u_i)$

- These conditional posteriors can be updated in an alternating fashion until convergence

  - This can be be implemented using a Gibbs sampler
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression.

- For each \( u_i \), its conditional posterior, given \( V \) and ratings

\[
p(u_i | R, V) = \mathcal{N}(u_i | \mu_{u_i}, \Sigma_{u_i})
\]

where \( \Sigma_{u_i} = (\lambda_u I + \beta \sum_{j: (i, j) \in \Omega} v_j v_j^\top)^{-1} \) and \( \mu_{u_i} = \Sigma_{u_i} (\beta \sum_{j: (i, j) \in \Omega} r_{ij} v_j) \)

- For each \( v_j \), its conditional posterior, given \( U \) and ratings

\[
p(v_j | R, U) = \mathcal{N}(v_j | \mu_{v_j}, \Sigma_{v_j})
\]

where \( \Sigma_{v_j} = (\lambda_v I + \beta \sum_{i: (i, j) \in \Omega} u_i u_i^\top)^{-1} \) and \( \mu_{v_j} = \Sigma_{v_j} (\beta \sum_{i: (i, j) \in \Omega} r_{ij} u_i) \)

- These conditional posteriors can be updated in an alternating fashion until convergence.

  - This can be implemented using a Gibbs sampler.

  - Note: Hyperparameters can also be inferred by computing their conditional posteriors (also see “Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo” by Salakhutdinov and Mnih (2008)).
Bayesian Matrix Factorization: Conditional Posteriors

- The conditional posteriors will have forms similar to solution of Bayesian linear regression

- For each $u_i$, its conditional posterior, given $V$ and ratings

$$p(u_i|R, V) = \mathcal{N}(u_i|\mu_{u_i}, \Sigma_{u_i})$$

where $\Sigma_{u_i} = (\lambda_u I + \beta \sum_{j:(i,j)\in\Omega} v_j v_j^\top)^{-1}$ and $\mu_{u_i} = \Sigma_{u_i}(\beta \sum_{j:(i,j)\in\Omega} r_{ij} v_j)$

- For each $v_j$, its conditional posterior, given $U$ and ratings

$$p(v_j|R, U) = \mathcal{N}(v_j|\mu_{v_j}, \Sigma_{v_j})$$

where $\Sigma_{v_j} = (\lambda_v I + \beta \sum_{i:(i,j)\in\Omega} u_i u_i^\top)^{-1}$ and $\mu_{v_j} = \Sigma_{v_j}(\beta \sum_{i:(i,j)\in\Omega} r_{ij} u_i)$

- These conditional posteriors can be updated in an alternating fashion until convergence

  - This can be implemented using a Gibbs sampler

  - Note: Hyperparameters can also be inferred by computing their conditional posteriors (also see “Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo” by Salakhutdinov and Mnih (2008))

  - Can extend Gaussian BMF easily to other exp. family distr. while maintaining local conjugacy
Bayesian Matrix Factorization

- The posterior predictive distribution for BMF (assuming other hyperparams known)

\[ p(r_{ij} | R) = \int \int p(r_{ij} | u_i, v_j) p(u_i, v_j | R) du_i dv_j \]

In general, this is hard and needs approximation

If we are using Gibbs sampling, we can use the \( S \) samples \( \{ u_i(1), v_j(1) \} \) to compute the mean

For the Gaussian likelihood case, the mean can be computed as

\[ E[r_{ij}] \approx \frac{1}{S} \sum_{s=1}^{S} u_i(s)^{\top} v_j(s) \] (Monte-Carlo averaging)

Can also compute the variance of \( r_{ij} \) (think how)

A comparison of Bayesian MF with other methods (from Salakhutdinov and Mnih (2008))
The posterior predictive distribution for BMF (assuming other hyperparams known)

\[ p(r_{ij} | R) = \int \int p(r_{ij} | u_i, v_j) p(u_i, v_j | R) du_i dv_j \]

In general, this is hard and needs approximation

- If we are using Gibbs sampling, we can use the \( S \) samples \( \{ u_i^{(s)}, v_j^{(s)} \}_{s=1}^S \) to compute the mean
- For the Gaussian likelihood case, the mean can be computed as

\[ \mathbb{E}[r_{ij}] \approx \frac{1}{S} \sum_{s=1}^S u_i^{(s)\top} v_j^{(s)} \]  

(Monte-Carlo averaging)

- Can also compute the variance of \( r_{ij} \) (think how)
Bayesian Matrix Factorization

- The posterior predictive distribution for BMF (assuming other hyperparams known)

\[ p(r_{ij}|R) = \int \int p(r_{ij}|u_i, v_j)p(u_i, v_j|R)du_idv_j \]

- In general, this is hard and needs approximation

  - If we are using Gibbs sampling, we can use the \( S \) samples \( \{u_i^{(s)}, v_j^{(s)}\}_{s=1}^S \) to compute the mean
  - For the Gaussian likelihood case, the mean can be computed as

\[
E[r_{ij}] \approx \frac{1}{S} \sum_{s=1}^S u_i^{(s)}v_j^{(s)} \quad \text{(Monte-Carlo averaging)}
\]

  - Can also compute the variance of \( r_{ij} \) (think how)

- A comparison of Bayesian MF with other methods (from Salakhutdinov and Mnih (2008))
Bayesian inference in even very complex probabilistic models can often be performed rather easily if the models have the local conjugacy property.
Bayesian inference in even very complex probabilistic models can often be performed rather easily if the models have the local conjugacy property.

It therefore helps to choose the likelihood model and priors on each parameter as exponential family distributions.
Bayesian inference in even very complex probabilistic models can often be performed rather easily if the models have the local conjugacy property.

It therefore helps to choose the likelihood model and priors on each parameter as exponential family distributions.

Even if we can’t get a globally conjugate model, we can still get a model with local conjugacy.
Summary and Some Comments

- Bayesian inference in even very complex probabilistic models can often be performed rather easily if the models have the local conjugacy property.
- It therefore helps to choose the likelihood model and priors on each param as exp. family distr.
  - Even if we can’t get a globally conjugacy model, we can still get a model with local conjugacy.
- Local conjugacy allows computing conditional posteriors that are needed in inference algos like Gibbs sampling, MCMC, EM, variational inference, etc.