Inference in Multi-parameter Models, Conditional Posterior, Local Conjugacy

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Plan

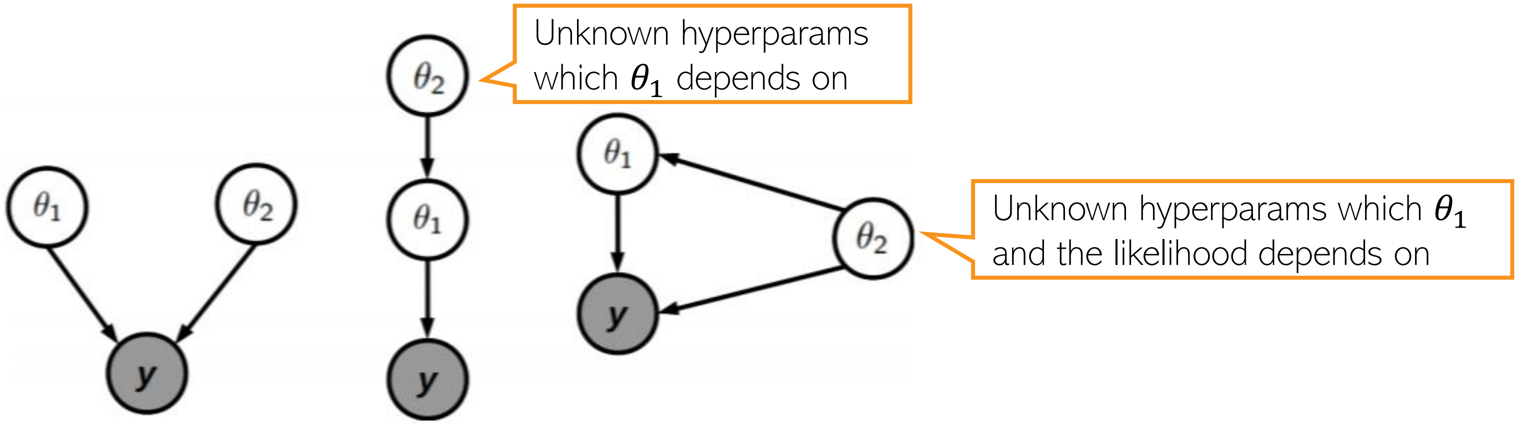
- Foray into models with several parameters
- Goal: Infer the posterior over all of them (not posterior for some, MLE-II for others)
- Idea of conditional/local posteriors in such problems
- Local conjugacy (which helps in computing conditional posteriors)
- Gibbs sampling (an algorithm that infer the joint posterior via conditional posteriors)
- An example: Bayesian matrix factorization (model with many parameters)
- Conditional/local posterior, local conjugacy, etc are important ideas (will appear in many inference algorithms that we will see later)

Moving Beyond Simple Models..

- So far, our models usually had a "main" parameter and maybe a few hyperparams, e.g.,
  - For a Gaussian, infer the mean assuming variance known (or vice-versa)
  - Bayesian linear regression with weight vector $\boldsymbol{w}$ and noise/prior precision hyperparams β, λ
  - GP regression with one function to be learned, and fixed hyperparams
- Easy posterior inference if the likelihood and prior are conjugate to each other
- Hyperparams usually estimated via MLE-II (full posterior much harder)
- For non-conjugate models or models with many parameters, need posterior approx
  - Can use Laplace approx but it has limitations (unimodality posterior, model should be differentiable)
- We will now look at more general inference schemes for "difficult" cases
  - Difficult = Models with many params/hyperparams, non-conjugacy, non-differ., etc
  - Will be intractable in general. We will study approx. inference methods to handle such cases

Multiparameter Models

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

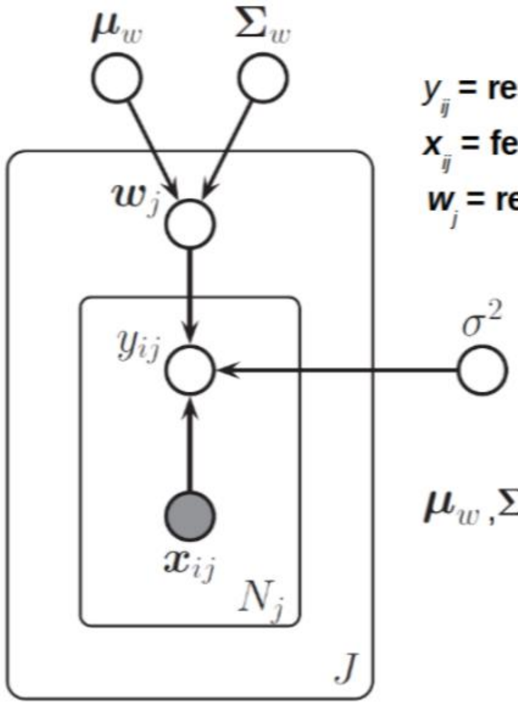


- Assume the likelihood model to be of the form $\boldsymbol{p(y|\theta_1,\theta_2)}$
- Assume a joint prior distribution $\boldsymbol{p(\theta_1,\theta_2)}$

  This prior may still be conditioned on some fixed hyperparams
- The joint posterior $\boldsymbol{p(\theta_1,\theta_2|y) \propto p(\theta_1,\theta_2)p(y|\theta_1,\theta_2)}$
  - 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)

Multiparameter Models: Some Examples

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



y_{ij} = response variable of student *i* in school *j*

$\mathbf{x}_{ij}$ = feature vector of student *i* in school *j*

$\mathbf{w}_j$ = regression model for school *j*

$\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w$ = parameters of the shared Gaussian prior $\mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ on all the *J* regression models

  - .. and in fact, pretty much in any non-toy example of probabilistic model

Another Example: Matrix Factorization/Completion

- Given: Data $\boldsymbol{R} = \{r_{ij}\}$ of "interactions" (e.g., ratings)
  - Here $i = 1,2,\ldots,N$ denotes users, $j = 1,2,\ldots,M$ denotes items

Assume each item $j = 1,2,\ldots,M$ modeled by an unknown parameter vector $\boldsymbol{v}_j \in \mathbb{R}^K$

Only a small fraction of user-item ratings are observed (say user-item index set Ω)

Need to predict the rest using these

Many methods exist

Assume each user $i = 1,2,\ldots,N$ modeled by an unknown parameter vector $\boldsymbol{u}_i \in \mathbb{R}^K$ for some small K

4	3			5	
5		4		4	
4		5	3	4	
	3				5
	4				4
		2	4		5

Similar to assuming a low-rank structure on the ratings matrix

Can assume zero mean Gaussian noise $\mathcal{N}(0, \beta^{-1})$ for real-valued ratings

$$r_{ij} = \boldsymbol{u}_i^\top \boldsymbol{v}_j + \epsilon_{ij}$$

Likelihood

$$p(r_{ij}|\boldsymbol{u}_i, \boldsymbol{v}_j) = \mathcal{N}(r_{ij}|\boldsymbol{u}_i^\top \boldsymbol{v}_j, \beta^{-1})$$

Individual priors for each user and item

$$p(\boldsymbol{u}_i) = \mathcal{N}(\boldsymbol{u}_i|\boldsymbol{0}, \lambda_u^{-1})$$

$$p(\boldsymbol{v}_j) = \mathcal{N}(\boldsymbol{v}_j|\boldsymbol{0}, \lambda_v^{-1})$$

$$p(\{\boldsymbol{u}_i\}_{i=1}^N, \{\boldsymbol{v}_j\}_{j=1}^M|\boldsymbol{R})$$

Posterior

For simplicity, hyperparams $\beta, \lambda_u, \lambda_v$ are fixed or unknow but we may only want to do point estimation for them

Bayesian Matrix Factorization (BMF): The Posterior

- Our target posterior distribution for this model is

Due to conditional independence of the observations given params

Assume that the joint prior factorizes into individual priors

$$p(\mathbf{U},\mathbf{V}|\mathbf{R}) = \frac{p(\mathbf{R}|\mathbf{U},\mathbf{V})p(\mathbf{U},\mathbf{V})}{\int\int p(\mathbf{R}|\mathbf{U},\mathbf{V})p(\mathbf{U},\mathbf{V})d\mathbf{U}d\mathbf{V}}$$

$$= \frac{\prod_{(i,j)\in\Omega} p(r_{ij}|\boldsymbol{u}_i,\mathbf{v}_j)\prod_i p(\boldsymbol{u}_i)\prod_j p(\mathbf{v}_j)}{\int\ldots\int \prod_{(i,j)\in\Omega} p(r_{ij}|\boldsymbol{u}_i,\mathbf{v}_j)\prod_i p(\boldsymbol{u}_i)\prod_j p(\mathbf{v}_j)d\boldsymbol{u}_1\ldots d\boldsymbol{u}_N d\mathbf{v}_1\ldots d\mathbf{v}_M}$$

- Posterior still tractable since integrals here are intractable
- Need to approx. the posterior. One way is via conditional posteriors (CP), e.g.,

All of $\boldsymbol{U}$ except $\boldsymbol{u}_i$

All of $\boldsymbol{V}$ except $\boldsymbol{v}_j$

$$p(\boldsymbol{u}_i|\boldsymbol{R},\boldsymbol{V},\boldsymbol{U}_{-i}) \qquad\qquad p(\boldsymbol{v}_j|\boldsymbol{R},\boldsymbol{U},\boldsymbol{V}_{-j})$$

- CP of each unknown is conditioned on fixed value of all other unknowns
  - The different CPs can be computed in an alternating fashion (like ALT-OPT/EM)
- Note: CP individually won't give us joint posterior. Need to combine them

E.g., using MCMC, variational inference, EM, etc

Conditional Posterior and Local Conjugacy

- Conditional Posteriors are easy to compute for model that are locally conjugacy
  - Note: Some researchers also call each CP as Complete Conditional or Local Posterior
- Consider a model with data $\boldsymbol{X}$ and K unknown params/h.p. $\boldsymbol{\Theta} = (\theta_1, \theta_2, \dots, \theta_K)$
- Suppose the joint posterior $p(\Theta|\boldsymbol{X}) = \frac{p(\Theta)p(\boldsymbol{X}|\Theta)}{p(\boldsymbol{X})}$ is intractable (like BMF)
- Local Conjugacy: If we can compute each CP tractably

Possible if the likelihood $p(\boldsymbol{X}|\theta_k, \Theta_{-k})$ and prior $p(\theta_k)$ are conjugate to each other

This is the notion of local conjugacy as opposed to full/joint conjugacy

Θ_{-k} is assumed known while computing this CP

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

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

In the likelihood model

Approximating Joint Posterior via CPs

- With the conditional posterior based approximation, the target joint posterior

$$p(\Theta|\mathbf{X}) = \frac{p(\mathbf{X}|\Theta)p(\Theta)}{p(\mathbf{X})}$$

... is represented by several conditional posteriors $p(\theta_k, |\boldsymbol{X}, \Theta_{-k})$, $k = 1,2, \dots, K$

- Each CP is a distribution over one unknown $\boldsymbol{\theta_k}$, given all other unknowns
- Need a way to "combine" these CPs to get the overall posterior
  - MCMC (e.g., Gibbs sampling): Based on generating samples from the CPs
  - Variational Inference (VI): Based on cyclic estimation of each CP
  - Note: Expectation Maximization also computes CP of latent variables in its E step
- More on this when we discuss MCMC, VI, EM, etc

Will revisit Gibbs sampling again when discussing MCMC algos

- But let's look at Gibbs sampling (an MCMC algo) right away as it is fairly simple

Gibbs Sampling (Geman and Geman, 1982)

- A general algo to generate samples from multivar. distr. <u>one component at a time</u>
  - Not limited to sampling from intractable posteriors only
  - Sometimes can be used even if we can draw from the multivar distr. directly

Note: If posterior, it will be conditioned on other stuff too (e.g., data, other param, etc)

- Assume we want to sample from a joint distribution $p(\theta_1, \theta_2, \ldots, \theta_K)$
- It generates one component θ_k at a time using its conditional $p(\theta_k | \Theta_{-k})$
- Each conditional is assumed to be available in closed form. An example below:

Suppose

A 2-dim Gaussian

$$\theta \sim N_2(0, \Sigma) \qquad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

Then

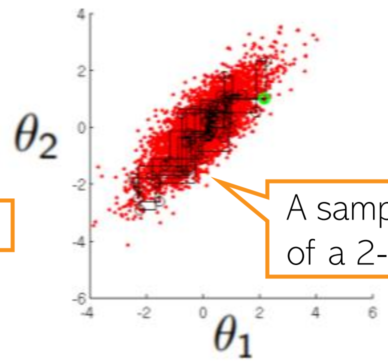
$$\theta_1 | \theta_2 \sim N\left(\rho\theta_2, [1 - \rho^2]\right)$$

A 1-dim Gaussian

$$\theta_2 | \theta_1 \sim N\left(\rho\theta_1, [1 - \rho^2]\right)$$

A 1-dim Gaussian

are the conditional distributions.



This example is a toy illustration of sampling from a 2-dim Gaussian by sampling from 1-dim Gaussians

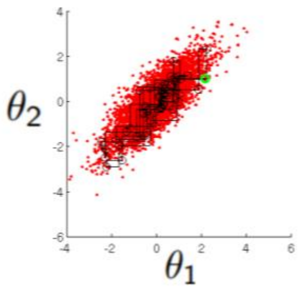
A sample based representation of a 2-dim Gaussian distribution

Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a sampling-based approx. of a multiparam. posterior
- Gibbs sampler iteratively draws random samples from CPs
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-param case $\theta = (\theta_1, \theta_2)$, the Gibb sampler looks like this
  - Initialize $\theta_2^{(0)}$
  - For $s = 1,2, \dots, S$
    - Draw a random sample for θ_1 as $\theta_1^{(s)} \sim p(\theta_1 | X, \theta_2^{(s-1)})$
    - Draw a random sample for θ_2 as $\theta_2^{(s)} \sim p(\theta_2 | X, \theta_1^{(s)})$

This CP uses the most recent value of θ_2

This CP uses the most recent value of θ_1



- These S random samples $\left(\theta_1^{(s)}, \theta_1^{(s)}\right)_{s=1}^{S}$ represent joint posterior $p(\theta_1, \theta_2 | X)$
- This is just a high-level idea. More on this when we discuss MCMC

Back to Bayesian Matrix Factorization

Bayesian Matrix Factorization: The CPs

- BMF with Gaussian likelihood and Gaussian prior on each user/item params is not fully conjugate but has local conjugacy
- To see this, note that the conditional posterior (CP) for user parameter $\boldsymbol{u}_i$

$$p(\boldsymbol{u}_i|\mathbf{R},\mathbf{V},\mathbf{U}_{-i}) \propto \prod_{j:(i,j)\in\Omega} p(r_{ij}|\boldsymbol{u}_i,\mathbf{v}_j)p(\boldsymbol{u}_i)$$

Only depends on the ratings of user $\boldsymbol{i}$. Also doesn't depend on $\mathbf{U}_{-i}$

This is due to the structure of the problem and the $\boldsymbol{u}_i$'s being independent a priori

$$= \prod_{j:(i,j)\in\Omega} \mathcal{N}(r_{ij}|\boldsymbol{u}_i^\top \mathbf{v}_j,\beta^{-1})\mathcal{N}(\boldsymbol{u}_i|\mathbf{0},\lambda_u^{-1}\mathbf{I}_K)$$

- The above is just like Bayesian linear regression, given $\boldsymbol{R}$ and fixed $\mathbf{V}$
  - Weight vector is $\boldsymbol{u}_i$, training data is $\{(\boldsymbol{v}_j, r_{ij})\}_{j:(i,j)\in\Omega}$, given
  - Also have local conjugacy since likelihood and prior are conjugate (assuming hyperparams fixed)
- Likewise, the CP for the item parameter $\boldsymbol{v}_j$ can be computed as

$$p(\mathbf{v}_j|\mathbf{R},\mathbf{U}) \propto \prod_{i:(i,j)\in\Omega} \mathcal{N}(r_{ij}|\boldsymbol{u}_i^\top \mathbf{v}_j,\beta^{-1})\mathcal{N}(\mathbf{v}_j|\mathbf{0},\lambda_v^{-1}\mathbf{I}_K)$$

Another Bayesian linear regression problem with weight vector $\boldsymbol{v}_j$

Bayesian Matrix Factorization: The CPs

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

$$p(\boldsymbol{u}_i|\mathbf{R},\mathbf{V}) = \mathcal{N}(\boldsymbol{u}_i|\boldsymbol{\mu}_{u_i},\boldsymbol{\Sigma}_{u_i})$$

$$\boldsymbol{\Sigma}_{u_i} = (\lambda_u \mathbf{I} + \beta \textstyle\sum_{j:(i,j)\in\Omega} \boldsymbol{v}_j \boldsymbol{v}_j^\top)^{-1}$$

$$\boldsymbol{\mu}_{u_i} = \boldsymbol{\Sigma}_{u_i}(\beta \textstyle\sum_{j:(i,j)\in\Omega} r_{ij} \boldsymbol{v}_j)$$

$$p(\boldsymbol{v}_j|\mathbf{R},\mathbf{U}) = \mathcal{N}(\boldsymbol{v}_j|\boldsymbol{\mu}_{v_j},\boldsymbol{\Sigma}_{v_j})$$

$$\boldsymbol{\Sigma}_{v_j} = (\lambda_v \mathbf{I} + \beta \textstyle\sum_{i:(i,j)\in\Omega} \boldsymbol{u}_i \boldsymbol{u}_i^\top)^{-1}$$

$$\boldsymbol{\mu}_{v_j} = \boldsymbol{\Sigma}_{v_j}(\beta \textstyle\sum_{i:(i,j)\in\Omega} r_{ij} \boldsymbol{u}_i)$$

- These CPs can be updated in an alternating fashion until convergence
  - Many ways. One popular way is to use Gibbs sampling
  - Note: Hyperparameters can also be inferred by computing their CPs[1]

- Can extend Gaussian BMF easily to other exp. family distr. while maintaining local conj.
  - Example: Poisson likelihood and gamma priors on user/item parameters[2]

[1] "Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo" by Salakhutdinov and Mnih (2008)

[2] "Scalable recommendation with Poisson factorization" by Gopalan et al (2013)

BMF: Making Predictions

- PPD for each missing entry of the matrix (assuming hyperparams known)

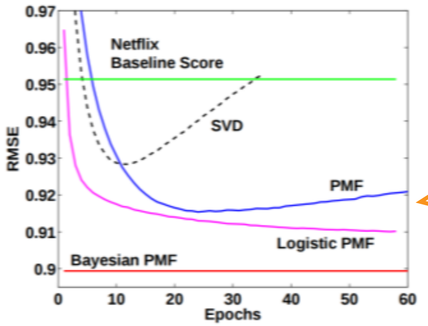
$$p(r_{ij}|\mathbf{R}) = \int\int p(r_{ij}|\boldsymbol{u}_i, \boldsymbol{v}_j)p(\boldsymbol{u}_i, \boldsymbol{v}_j|\mathbf{R})d\boldsymbol{u}_i d\boldsymbol{v}_j$$

- In general, this is intractable and needs approximation
  - If using Gibbs sampling, we can use S samples $(u_i^{(s)}, v_j^{(s)})_{s=1}^{S}$ to compute mean of r_{ij}
  - For the Gaussian likelihood case, the mean can be computed as

Can also compute the variance in the predicted r_{ij} using these S samples (think how)

$$\mathbb{E}[r_{ij}] \approx \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{u}_i^{(s)\top} \boldsymbol{v}_j^{(s)} \qquad \text{(Monte-Carlo averaging)}$$

- Comparison of Bayesian MF with others (from Salakhutdinov and Mnih (2008))



All other baselines are optimization based or point estimation based probabilistic models (PMF = probabilistic matrix factorization with point estimation)

Summary

- Local conjugacy is helpful even for complex prob. models with many params
  - CPs will have a closed form
  - Easy to implement Gibbs sampling can be used to be (approx.) posterior
  - Many other approx. inference algos (like variational inference) benefit from local conjugacy

- Helps to choose likelihood and priors on each param as exp. family distr.
  - So if we can't get a globally conjugate model, we can still get a model with local conjugacy

- Even if we can't have local conjugacy, the notion of CPs is applicable
  - We can break the inference problem into estimating CPs (exactly if we have local conjugacy, or approximately if we don't have local conjugacy)
  - Almost all approx. algorithms work by estimating CPs exactly or approximately

Coming Up

- Latent Variable Models
- Expectation Maximization
- Variational Inference