COS513 LECTURE 8
STATISTICAL CONCEPTS

NIKOLAI SLAVOV AND ANKUR PARIKH

1. Making meaningful statements from joint probability distributions.

A graphical model (GM) represents a family of probability distributions over its variables. The model has

- Variables that can be:
  - Observed – represented by shaded nodes
  - Hidden – Represented by unshaded nodes. Typical examples of models having unobserved variables are some clustering and classification models.
- Parameters. The parameters specify a particular member of the family of distributions defined by the graphical model. The parameters can be:
  - Potential functions
  - Tables of conditional distributions

Given a model and parameters, we have spent the last few lectures discussing how to infer marginal and conditional probabilites of random variables taking on particular values. However, what about the reverse question? When building the model, how do we infer the parameters from data? The goal of this lecture is to provide an overview of statistical concepts that can provide us with the tools to answer this type of question.

2. Bayesian versus Frequentist statistics.

Two main schools of statistics with a history of rivalry are the Bayesian and the Frequentist schools.

2.1. **Bayesian Statistics.** For a Bayesian statistician every question is formulated in terms of a joint probability distribution. For example, if we have selected a model and need to infer its parameters (that is select a particular member from the family of distributions specified by the model structure) we can easily express $p(\vec{x}|\theta)$, the probability of observing the data $\vec{x}$ given the parameters, θ. However, we would like to know about $p(\theta|\vec{x})$, the probability of the parameters given the data. Applying Bayes rule we express

$p(\theta|\vec{x})$ in terms of $p(\vec{x}|\theta)$ and the marginal distributions (Note that the numerator is a joint distribution).

$$p(\theta|\vec{x}) = \frac{p(\vec{x}|\theta)p(\theta)}{p(\vec{x})}. \tag{1}$$

The expression implies that θ is a random variable and requires us specify its marginal probability $p(\theta)$, which is known as the **prior**. As a result we can compute the conditional probability of the parameter set given the data $p(\theta|\vec{x})$ (known as the **posterior**). Thus, Bayesian inference results not in a single best estimate (as is the case with the maximum likelihood estimator, see the next section) but in a conditional probability distribution of the parameters given the data.

The requirement to set a prior in Bayesian statistics is one of the most frequent criticisms from the Frequentist school since computing the posterior is subject to bias coming from the initial belief (the prior) that is formed (at least in some cases) without evidence. One approach to avoid setting arbitrary parameters is to estimate $p(\theta)$ from data not used in the parameter estimation. We delve deeper into this problem later.

2.2. **Frequentist Statistics.** Frequentists do not want to be restricted to probabilistic descriptions and try to avoid the use of priors. Instead they seek to develop an "objective" statistical theory, in which two statisticians using the same methodology must draw the same conclusions from the data, regardless of their initial beliefs. In frequentist statistics, θ is considered a fixed property, and thus not given a probability distribution. To invert the relationship between parameters and data, frequentists often use estimators of θ and various criteria of how good the estimator is. Two common criteria are the bias and the variance of an estimator. The variance quantifies deviations from the expected estimate. The bias quantifies the difference between the true value of θ and the expected estimate of θ based on an infinite (very large) amount of data. The lower the variance and the bias of an estimator the better estimates of θ it is going to make.
A commonly used estimator is the maximum likelihood estimator (MLE). The MLE aims to identify the parameter set $\hat{\theta}_{ML}$ for which the probability of observing the data is maximized. This is accomplished by maximizing the likelihood $p(\vec{x}|\theta)$ with respect to the parameters. Equivalently, we can maximize the logarithm of the likelihood (called the log likelihood). This gives the same optimization result, since $\log(x)$ is a monotonic function.

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(\vec{x}|\theta) = \arg\max_{\theta} \log p(\vec{x}|\theta). \tag{2}$$

For linear models in which the errors are not correlated, have expectation zero, and have the same variance, the best unbiased estimator is the maximum likelihood estimator (MLE), see the **Gauss Markov Theorem**. However, in some cases it may be desirable to trade increased bias in the estimator for significant reduction in its variance. This will be discussed later in the context of regularized regression. One way to reduce the variance of the estimate of θ is to use the maximum a posteriori (MAP) estimator. The $\hat{\theta}_{MAP}$ is exactly the mode of the posterior distribution. To compute $\hat{\theta}_{MAP}$ we maximize the posterior:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|\vec{x}) = \arg\max_{\theta} \frac{p(\vec{x}|\theta)p(\theta)}{p(\vec{x})}. \tag{3}$$

However, since $p(x)$ is independent of θ we can simplify this to:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\vec{x}|\theta)p(\theta). \tag{4}$$

Equivalently, we can maximize the logarithm of a posterior:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} [\log p(\vec{x}|\theta) + \log p(\theta)]. \tag{5}$$

In the equation above, the prior serves as a "penalty term". One can view the MAP estimate as a maximization of penalized likelihood.

When calculating the $\hat{\theta}_{MAP}$ we need to know the prior. However, the prior also has parameters (called **hyperparameters**). We could treat these as random variables and endow them with distributions that have more hyperparameters. This is the concept of hierarchical Bayesian modeling. However, we must stop at some point to avoid infinite regress. At this point we have to estimate the hyperparameters. One common approach is to use the MLE to estimate these:

$$\hat{\alpha}_{ML} = \arg\max_{\alpha} \log p(\vec{\theta}|\alpha), \tag{6}$$

where α is a parameter of the prior (a hyperparameter).
The advantage of using a hierarchical model is that the hyperparameters that have to be set (at the top level) have less of an influence on the posterior than θ does. The higher the hierarchy, the smaller the influence (meaning less subjectivity). However, this benefit has diminishing returns as the number of levels increases, so 2-3 levels is usually sufficient. Increasing the hierarchy also increases the computational complexity of estimation.

The last estimator we mention is the Bayes estimate, which is the mean of the posterior.

$$\hat{\theta}_{Bayes} = \int \theta p(\theta|x)\, d\theta. \tag{7}$$

3. Density Estimation

Let us see how these two schools of thought tackle the major statistical problem of density estimation. Consider a random variable X. The goal of density estimation is to induce a probability density function (or probability mass function in the case of discrete variables) for X, based on a set of observations: $\{X_1, X_2, ...X_N\}$. $X_1, ...X_N$ are random variables all characterized by the same distribution (**identically distributed**).

For our examples below, let us assume that X is a univariate random variable with a Gaussian distribution.

$$p(x|\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}. \tag{8}$$

A univariate gaussian's parameters are its mean μ and its variance σ^2, so $\theta = \{\mu, \sigma^2\}$. Therefore our goal is to be able to characterize θ based on our set of observations.

3.1. The Frequentist Approach to Density Estimation.

We first present the IID sampling model, a frequentist approach to density estimation.In addition to assuming that the observations are identically distributed, the IID sampling model also assumes that the observations are **independent** (IID stands for independent and identically distributed). We can represent this as a graphical model, as shown in Figure 1. Note that θ is not a random variable in this case (as indicated by the dot in the node). It is easy to see that the variable nodes are independent.

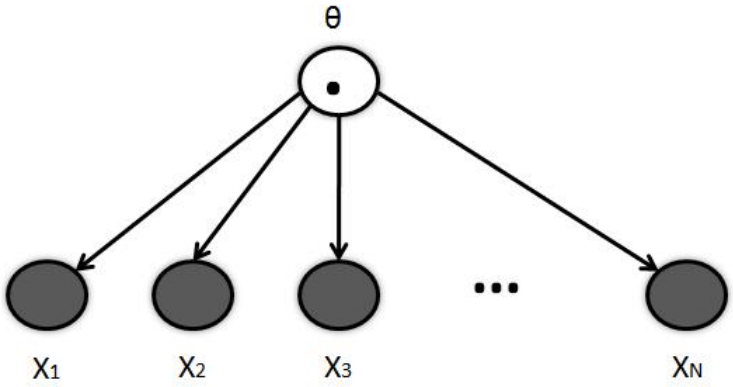

FIGURE 1. The IID Sampling Model

Since we are assuming each X_i to be independent:

$$p(x_{1:N}|\theta) = \prod_{n=1}^{N} p(x_n|\theta) \tag{9}$$

$$= \prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{\frac{1}{2\sigma^2}(x_n-\mu)^2\right\} \tag{10}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}. \tag{11}$$

To estimate the parameters we maximize the log likelihood with respect to θ:

$$l(\theta; x_{1:N}) = \log p(x_{1:N}|\theta) \tag{12}$$

$$= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2. \tag{13}$$

$$\frac{\partial l(\theta; x_{1:N})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu). \tag{14}$$

To maximize, we set the derivative to 0 and solve:

$$\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\hat{\mu}_{ML}) = 0. \tag{15}$$

$$\sum_{n=1}^{N} x_n - N\hat{\mu}_{ML} = 0. \tag{16}$$

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n. \tag{17}$$

Now we do the same for variance:

$$\frac{\partial l(\theta; x_{1:N})}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}(-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2) \tag{18}$$

$$= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n - \mu)^2. \tag{19}$$

$$0 = -\frac{N}{2\hat{\sigma}^2_{ML}} + \frac{1}{2\hat{\sigma}^4_{ML}}\sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2. \tag{20}$$

$$\hat{\sigma}^2_{ML} = \sum_{n=1}^{N}(x_n - \hat{\mu}_{ML})^2. \tag{21}$$

Thus we see that the maximum likelihood estimates of the mean and variance are simply the sample mean and sample variance respectively. (We are setting both partial derivatives to 0 and solving simultaneously. This explains the presence of $\hat{\mu}_{ML}$ when finding $\hat{\sigma}^2_{ML}$.)

3.2. **The Bayesian Approach to Density Estimation.** The Bayesian approach to density estimation is to form a posterior density $p(\theta|x)$. Consider a simpler version of the problem where the the variance σ^2 is a known parameter, and but the mean μ is unknown and thus a random variable. Thus, we seek to obtain $p(\theta|x)$ based on the the Gaussian likelihood $p(x|\mu)$ and the prior density $p(\mu)$.

Now the decision to make is what is the prior density of μ? For computational ease we take $p(\mu)$ to be a Gaussian distribution (since it will give us a Gaussian posterior). To be consistent with the Bayesian approach, we must consider the parameters of this prior distribution (the hyperparameters). We could treat these as random variables as well and add another level of hierarchy. However, in this example, we choose to stop and let the fixed constants μ_0 and τ^2 be the mean and variance of the prior distribution respectively. The situation is modeled in Figure 2. Here we see that μ is now a random variable, and that the observations are conditionally independent given μ. However, they are not marginally independent.

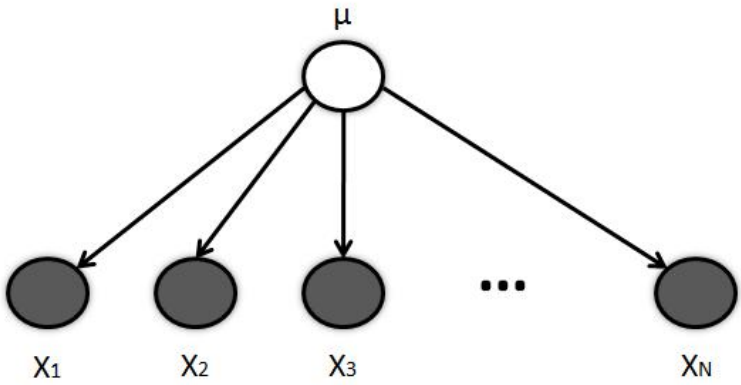

FIGURE 2. The Bayesian Model for Density Estimation

Thus the Bayesian model is making a weaker assumption than the frequentist model, which assumes marginal independence. We elaborate on this when we discuss De Finetti's Theorem and exchangability.

We can now calculate the prior density to be:

$$p(\mu) = \frac{1}{2\pi\tau^2} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\}, \tag{22}$$

and we obtain the joint probability:

(23)

$$p(x_{1:N}, \mu) = p(x_{1:N}|\mu)p(\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \frac{1}{2\pi\tau^2} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\}.$$

Normalizing the joint probability (dividing by $p(x_{1:N})$), yields the posterior:

$$p(\mu|x_{1:N}) = \frac{1}{2\pi\tilde{\sigma}^2} \exp\left\{-\frac{1}{2\tilde{\sigma}^2}(\mu - \tilde{\mu})^2\right\}. \tag{24}$$

$$\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/r^2}\bar{x} + \frac{1/r^2}{N/\sigma^2 + 1/r^2}\mu_0, \tag{25}$$

where $\bar{x}$ is the sample mean.

$$\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}. \tag{26}$$

Please see the Appendix for a derivation. Note for small N, the hyperparameters μ_0 and τ^2 have a significant influence on $\tilde{\mu}$ and $\tilde{\sigma}^2$. However, as N gets larger this influence diminishes and $\tilde{\mu}$ approaches the sample mean and

$\tilde{\sigma}^2$ approaches the sample variance. Thus, as N gets larger the Bayesian estimate approaches the maximum likelihood estimate. Ths intuitively makes sense since as we see more data points, we rely on the prior less and less.

4. De Finetti's Theorem

We had briefly mentioned that by assuming only conditional independence, the Bayesian model was making a weaker assumption than the IID sampling model. We now make this more precise. In fact the assumption of conditional independence is equal to that of exchangability. Exchangability essentially means that the order of our observations does not matter.

More formally, $X_1, X_2, ...X_N$ are exchangable if for any permutation π, $X_1, X_2, ...X_N$ and $X_{\pi(1)}, X_{\pi(2)}, ...X_{\pi(N)}$ have the same probability distribution.

De Finetti's theorem states that:

If $X_1, X_2, ...X_N$ are exchangeable, then

$$p(x_1, x_2, ...x_N) = \int (\prod_{n=1}^{N} p(x_n|\theta)p(\theta)\, d\theta, \tag{27}$$

and thus establishes that exchangability implies conditional independence.

This is a major advantage of Bayesian techniques, since exchangability is a very reasonable assumption for many applications, while marginal independence is often not.

5. Plates

Sometimes when describing graphical models, it is convenient to have a notation to indicate repeated structures. We use plates (or rectangular boxes) to enclose a repeated structure. We interpret this by copying the structure inside the plate N times, when N is indicated in the lower right corner of the plate. On the next page are the plate diagrams for both the IID sampling and Bayesian density estimation models.

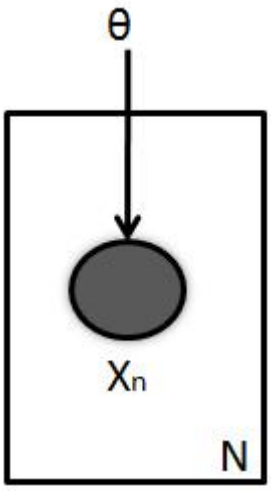

FIGURE 3. Plate notation for the IID sampling model

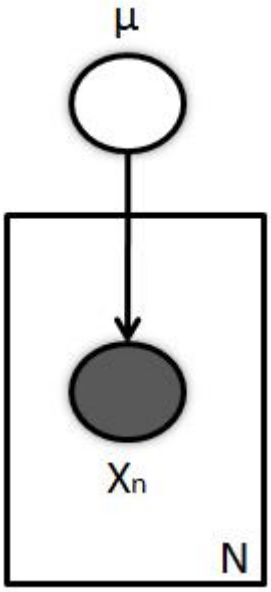

FIGURE 4. Plate notation for Bayesian density estimation

6. DENSITY ESTIMATION OF DISCRETE DATA

We next describe the problem of density estimation for discrete random variables where the goal is to estimate the probability mass function (instead of the density function).

6.1. **The Frequentist Approach.** First we describe the frequentist approach. As before we assume that the data is IID. We allow the observations (random variables) $X_1, X_2...X_N$ to take on M values. To represent this range of values, we find it convenient to use a vector representation. Let the range of X_n be the set of binary M-component vectors with exactly one component equal to 1 and the other components equal to 0. For example, if X_n could take on three possible values, we have

$$X_n \in \{[1,0,0],[0,1,0],[0,0,1]\} \tag{28}$$

We use superscripts to refer to the components of these vectors, so X_n^k refers to the k^{th} component of X_n. Note that since only of the components is equal to 1, we have $\sum_k X_n^k = 1$.

With this representation we write the probability mass function of X_n in the general form of a multinomial distribution with parameter vector $\theta = (\theta_1, \theta_2, ...\theta_N)$. The multinomial distribution is basically just a generalization of the binomial distribution to M outcomes (instead of just 2). Note that $\sum_i \theta_i = 1$.

$$p(x_n|\theta) = \theta_1^{x_n^1}\theta_2^{x_n^2} \cdots \theta_M^{x_n^M}. \tag{29}$$

Now to find $p(x_{1:N}|\theta)$ we take the product over the individual multinomial probabilities, (since the observations are assumed to be independent).

$$p(x_{1:N}|\theta) = \prod_{n=1}^{N} \theta_1^{x_n^1}\theta_2^{x_n^2} \cdots \theta_M^{x_n^M}. \tag{30}$$

$$= \theta_1^{\sum_{n=1}^N x_n^1}\theta_2^{\sum_{n=1}^N x_n^2} \cdots \theta_M^{\sum_{n=1}^N x_n^M}. \tag{31}$$

Just like in the case of the Gaussian, we estimate the parameters by maximizing the log-likelihood.

$$l(\theta; x_{1:N}) = \sum_{n=1}^{N}\sum_{k=1}^{M} x_n^k \log \theta_k. \tag{32}$$

To maximize the equation above with respect to θ, we use Lagrange multipliers.

$$\tilde{l}(\theta; x_{1:N}) = \sum_{n=1}^{N}\sum_{k=1}^{M} x_n^k \log \theta_k + \lambda(1 - \sum_{k=1}^{M} \theta_k). \tag{33}$$

We take partial derivatives with respect to θ_k.

$$\frac{\partial \tilde{l}(\theta; x_{1:N})}{\partial \theta_k} = \frac{\sum_{n=1}^{N} x_n^k}{\theta_k} - \lambda. \tag{34}$$

To maximize we set the derivatives equal to 0:

$$\lambda = \frac{\sum_{n=1}^{N} x_n^k}{\hat{\theta}_{ML,k}} \tag{35}$$

$$\lambda \hat{\theta}_{ML,k} = \sum_{n=1}^{N} x_n^k \tag{36}$$

$$\lambda \sum_{k=1}^{M} \hat{\theta}_{ML,k} = \sum_{k=1}^{M} \sum_{n=1}^{N} x_n^k \tag{37}$$

$$\lambda = \sum_{k=1}^{M} \sum_{n=1}^{N} x_n^k \tag{38}$$

$$\lambda = \sum_{n=1}^{N} \sum_{k=1}^{M} x_n^k \tag{39}$$

$$\lambda = N. \tag{40}$$

This yields:

$$\hat{\theta}_{ML,k} = \frac{1}{N} \sum_{n=1}^{N} x_n^k. \tag{41}$$

Just like in the case of the Gaussian, this is basically a sample proportion since $\sum_{n=1}^{N} x_n^k$ is a count of the number of times that the kth value is observed.

6.2. **Bayesian estimation of discrete data.** As in the Gaussian example, we must specify a prior distribution for the parameters before we can calculate the posterior. In the Gaussian setting, we chose the prior to be Gaussian so the posterior would be Gaussian. Here we seek the same property. In general, a family of distributions that when chosen as a prior gives a posterior in the same family (with respect to a particular likelihood), is called a **conjugate prior**. In this case, the conjugate prior of the multinomial distribution is the Dirichlet distribution:

$$p(\theta) = C(\alpha)\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1} .... \theta_M^{\alpha_M - 1}. \tag{42}$$

where $\sum_i \theta_i = 1$ and where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_M)$ are the hyperparameters ($\theta_i' s$ are the variables). $C(\alpha)$ is a normalizing constant:

$$C(\alpha) = \frac{\Gamma(\sum_{i=1}^{M} \alpha_i)}{\prod_{i=1}^{M} \Gamma(\alpha_i)}, \tag{43}$$

where $\Gamma(\cdot)$ is the gamma function. We then calculate the posterior:

$$p(\theta|x_{1:N}) \propto \theta_1^{\sum_{n=1}^{N} x_n^1} \theta_2^{\sum_{n=1}^{N} x_n^2} \cdots \theta_M^{\sum_{n=1}^{N} x_n^M} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_M^{\alpha_M - 1} \tag{44}$$

$$= \theta_1^{\sum_{n=1}^{N} x_n^1 + \alpha_1 - 1} \theta_2^{\sum_{n=1}^{N} x_n^2 + \alpha_2 - 1} \cdots \theta_M^{\sum_{n=1}^{N} x_n^M + \alpha_M - 1}. \tag{45}$$

This is a Dirichlet density with parameters $\sum_{n=1}^{N} x_n^k + \alpha_k$. Note that it is easy to to find the normalization factor, since we can just substitute into Equation 43.

7. Appendix

In this section we derive Equation 24. Recall that:

$$p(x_{1:N}, \mu) = p(x_{1:N}|\mu)p(\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\} \frac{1}{2\pi\tau^2} \exp\left\{-\frac{1}{2\tau^2}(\mu - \mu_0)^2\right\}, \tag{46}$$

and we seek to normalize this joint probability to obtain the posterior.

Let us focus on the terms involving μ, treating the other terms as "constants" and dropping them as we go. We have:

$$\begin{aligned}
p(x_{1:N}, \mu) &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n^2 - 2x_n\mu + \mu^2) - \frac{1}{2\tau^2}(\mu^2 - 2\mu\mu_0 + \mu^2)\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left[\frac{1}{\sigma^2}(x_n^2 - 2x_n\mu + \mu^2) + \frac{1}{\tau^2}\left(\frac{\mu^2}{N} - \frac{2\mu\mu_0}{N} + \frac{\mu^2}{N}\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\left[\left(\frac{1}{\sigma^2} + \frac{1}{N\tau^2}\right)\mu^2 - 2\left(\frac{x_n}{\sigma^2} + \frac{\mu_0}{N\tau^2}\right)\mu + C\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[\left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)\mu^2 - 2\left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)\mu + C\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)\left[\mu^2 - 2\left(\frac{N}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{N\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}\right)\mu\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}[\mu^2 - 2\tilde{\mu}\mu]\right\},
\end{aligned}$$

where

$$\tilde{\sigma}^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}, \tag{47}$$

and

$$\tilde{\mu} = \frac{N/\sigma^2}{N/\sigma^2 + 1/r^2} \bar{x} + \frac{1/r^2}{N/\sigma^2 + 1/r^2} \mu_0. \tag{48}$$

$\bar{x}$ is the sample mean.

This proves that the posterior has a Gaussian distribution with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$.