Special Topics in Computational Biology

November 8, 2006

Stanley Sawyer, Department of Mathematics, WashU

Nested Hypothesis Tests:
Likelihoods and Likelihood Ratios:

General framework:
H_1: Model with n_1 parameters
H_0: Model with n_0 parameters with $H_0 \subseteq H_1$

Are the extra parameters really necessary?
Does H_1 fit the data any better than H_0?

A too-easy example:
Toss a coin 100 times, get 56 heads, 44 tails.

H_1: Coin has $\Pr(H) = p$ for unknown p
H_0: Coin is fair: That is, $p = 0.50$

Is the extra parameter p really justified by the data?

A better example:

Suppose that we observe for 1000 nucleotides from one strand of DNA:

(A)212 (T)219 (C)253 (G)316

In particular, this implies:

(AT)431 (CG)569

This DNA strand appears CG-rich, but is there AT or CG bias within the strand? Can we test

$H_0 : p_A = p_T$ and $p_C = p_G$?

even though we are fairly sure that $p_{AT} < p_{CG}$?
If this is false, there are four parameters,

namely p_A, p_T, p_C, and p_G.

This is a nested hypothesis test with

H_1: 3 free parameters (p_A, p_T, p_C)

H_0: 1 free parameter $p_A = p_T$

since $p_A + p_A + p_C + p_C = 1$.

A general theory for testing nested hypotheses:
We go through four steps:

I. *Likelihoods of H_0 and H_1:* Let $L_{H_1}(p, X)$ be the probability of observing the nucleotide counts X. We assume for simplicity the probability that they are observed *in a particular order*, so that

$$\begin{aligned} L_{H_1}(p, X) &= L_{H_1}(p_A, p_T, p_C, p_G, X) \\ &= {p_A}^{n_A}\, {p_T}^{n_T}\, {p_C}^{n_C}\, {p_G}^{n_G} \end{aligned}$$

without any combinatorial coefficients, where

$$p_A + p_T + p_C + p_G = 1$$

We use the same likelihood function for H_0 and H_1 with restricted values for H_0, so that

$$\begin{aligned} L_{H_0}(p, X) &= L_{H_1}(p_A, p_A, p_C, p_C, X) \\ &= {p_A}^{n_A+n_T}\, {p_C}^{n_C+n_G} \end{aligned}$$

where $p_A + p_C = (1/2)(p_A + p_A + p_C + p_C) = 1/2$.

II. Define the *maximum likelihood estimator* (MLE) $\widehat{p} = \widehat{p}(X)$ of $p = (p_A, p_T, p_C, p_G)$ (for H_1) as the solution of

$$\max_p L_{H_1}(p, X) \;=\; L_{H_1}\big(\widehat{p}(X),\, X\big)$$

Since

$$\log L_{H_1}(p, X) = n_A \log p_A + \cdots + n_G \log p_G$$

This is the same as solving (for A)

$$\frac{\partial}{\partial p_A} \log L_{H_1}(p, X) \;=\; \frac{p_A}{n_A} - \frac{p_G}{n_G} \;=\; 0$$

The second terms is because we are assuming that p_A, p_C, p_T are free with $p_G = 1 - p_A - p_T - p_C$, so that $\frac{\partial}{\partial p_A} p_G = -1$. Similarly

$$\frac{\partial}{\partial p_T} \log L_{H_1}(p, X) \;=\; \frac{p_T}{n_T} - \frac{p_G}{n_G} \;=\; 0$$

$$\frac{\partial}{\partial p_C} \log L_{H_1}(p, X) \;=\; \frac{p_C}{n_C} - \frac{p_G}{n_G} \;=\; 0$$

This implies

$$\frac{p_A}{n_A} = \frac{p_T}{n_T} = \frac{p_C}{n_C} = \frac{p_G}{n_G} = \lambda$$

Thus $p_A+p_T+p_C+p_G = 1 = \lambda(n_A+n_T+n_C+n_G) = \lambda n$, from which it follows that $\lambda = 1/n$. Hence the MLEs are the sample proportions

$$\widehat{p}_A = \frac{n_A}{n}, \quad \widehat{p}_T = \frac{n_T}{n}, \quad \widehat{p}_C = \frac{n_C}{n}, \quad \widehat{p}_G = \frac{n_G}{n}$$

Similarly

$$\log L_{H_0}(p, X) = (n_A + n_T) \log p_A + (n_C + n_G) \log p_C$$

where $p_A + p_C = (1/2)$. Thus

$$\frac{\partial}{\partial p_A} \log L_0(p, X) = \frac{p_A}{n_A + n_T} - \frac{p_C}{n_C + n_G} = 0$$

or $p_A/(n_A + n_T) = p_C/(n_C + n_G) = \lambda_0$ for some λ_0.

This implies $p_A+p_C = 1/2 = \lambda_0(n_A+n_T+n_C+n_G) = \lambda_0 n$. Thus $\lambda_0 = 1/(2n)$ and

$$\widehat{p}_A = \widehat{p}_T = \frac{n_A + n_T}{2n}, \quad \widehat{p}_C = \widehat{p}_T = \frac{n_C + n_G}{2n}$$

III. We find the *estimated likelihood of the data* X for models H_1 and H_0: Our best guess for the likelihood of X for H_1 is

$$\begin{aligned}\widehat{L}(H_1, X) &= L_{H_1}(\widehat{p}_1(X), X) \\ &= \left(\frac{n_A}{n}\right)^{n_A} \left(\frac{n_T}{n}\right)^{n_T} \left(\frac{n_C}{n}\right)^{n_C} \left(\frac{n_G}{n}\right)^{n_G}\end{aligned}$$

By the same arguments, the estimated likelihood of X for H_0 is

$$\begin{aligned}\widehat{L}(H_0, X) &= L_{H_1}(\widehat{p}_0(X, H_0), X) \\ &= \left(\frac{n_A + n_T}{2n}\right)^{n_A+n_T} \left(\frac{n_C + n_G}{2n}\right)^{n_C+n_G}\end{aligned}$$

IV. Finally, the *Likelihood Ratio Test* (LRT) of H_0 with respect to accepting H_1 is to compute and compare

$$Q = 2 \log \left(\frac{\widehat{L}(H_1, X)}{\widehat{L}(H_0, X)} \right) \approx \chi^2_d$$

where $d = n_1 - n_0$ is the number of extra parameters in H_1. That is, $P = \Pr\left(\chi^2_d \geq Q_{\text{obs}}\right)$. Here $d = 3-1 = 2$, since H_1 has two more parameters. Here

$$\begin{aligned}
\log \widehat{L}(H_1, X) &= 212 \log \left(\frac{212}{1000} \right) + 219 \log \left(\frac{219}{1000} \right) \\
&\quad + 253 \log \left(\frac{253}{1000} \right) + 316 \log \left(\frac{316}{1000} \right) \\
&= -1373.190
\end{aligned}$$

$$\begin{aligned}
\log \widehat{L}(H_0, X) &= 431 \log \left(\frac{431}{2000} \right) + 569 \log \left(\frac{569}{2000} \right) \\
&= -1376.742
\end{aligned}$$

Thus

$$\begin{aligned} Q &= 2\log\left(\frac{\widehat{L}(H_1, X)}{\widehat{L}(H_0, X)}\right) \\ &= 2\Big(\log \widehat{L}(H_1, X) - \log \widehat{L}(H_0, X)\Big) \\ &= 2(1376.742 - 1373.190) = 7.1034 \end{aligned}$$

The P-value for rejecting H_0 in favor of H_1 is

$$P = P(\chi^2_2 \geq 7.1034) = 0.0287 < 0.05$$

so that we reject H_0. That is, the data show borderline evidence for strand asymmetry between A and T and/or between C and G, for data

(A)212 (T)219 (C)253 (G)316

and

(AT)431 (CG)569

Bayesian Analysis and Conjugate Priors:

Example: Toss a coin $n = 10$ times. Let X be the number of heads ($0 \leq X \leq 10$). How should we estimate $p = \Pr(\text{Head})$? Here, the likelihood and MLE are:

$$L(p, X) = p^X (1-p)^{10-X}, \qquad \widehat{p}(X) = \frac{X}{10}$$

But what if $X = 0$? Is $\widehat{p}(X) = 0$ a reasonable assertion based on just 10 coin tosses?

One problem may be that we are treating p and X differently (as a parameter and a random variable, respectively):

We might have a better idea of how to handle odd questions such as this (and many other odd questions more generally) if we could somehow put p and X on the same footing.

The first step is to force p to be a random variable by saying that it has a probability distribution $\pi_0(p)$ for $0 \leq p \leq 1$. This is called a *prior distribution* for p, since we have not yet observed X.

The prior distribution $\pi_0(p)$ makes p into a random variable. That is,

$$\sum_{X=0}^{n} \int_0^1 \pi_0(p) L(p, X) = 1$$

This means that we can view (p, X) as two random variables with joint probability distribution

$$\pi_1(p, X) = \pi_0(p) L(p, X) = \pi_0(p)\, p^X (1-p)^{10-X}$$

Finally, we notice that X is constant (because we have just observed it), so that we can form the *conditional* or *posterior distribution* of p given X:

$$\pi_1(p \mid X) = \frac{\pi_0(p)L(p, X)}{\int_0^1 \pi_0(z)L(z, X)\, dz}$$

We can then use the posterior distribution $\pi_1(p \mid X)$ to define the *Bayes estimator*

$$\widehat{p}_B(X) = \int_0^1 p\pi_1(p, X)\, dp = \frac{\int_0^1 p\pi_0(p)L(p, X)\, dp}{\int_0^1 \pi_0(p)L(p, X)\, dp}$$

Note that $\widehat{p}_B(X)$ makes sense even if $X = 0$, but depends on $\pi_0(p)$.

Also, we should be careful when we use this method, since we may end up having to evaluate not only one, but two nasty integrals. In contrast, finding MLEs only requires derivatives but not integrals.

Example: $\pi_0(p) = 1$. Then the full likelihood is

$$\pi_1(p, X) = \pi_0(p)L(p, X) = p^X(1-p)^{10-X}$$

so that the posterior distribution is

$$\pi_1(p \mid X) = C(X)p^X(1-p)^{10-X}$$

Note that, as a function of p, the posterior density

$$\begin{aligned}\pi_1(p \mid X) &= C(X)p^X(1-p)^{10-X} \\ &\approx \text{Beta}(X+1, 11-X)\end{aligned}$$

where $\text{Beta}(\alpha, \beta)$ is the beta distribution with density

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, p^{\alpha-1}(1-p)^{\beta-1}$$

Since

$$E\big(\text{Beta}(\alpha,\beta)\big) = \frac{\alpha}{\alpha+\beta}$$

it follows that

$$\begin{aligned}\widehat{p}_B(X) &= \int_0^1 p\pi_1(p \mid X)\,dp \\ &= \frac{X+1}{12}\end{aligned}$$

In particular, if $X = 0$, $\widehat{p}_B(0) = 1/12$.

This argument generalizes: Suppose we chose, instead, $\pi_0(p) = Cp(1-p)$ or $Cp^5(1-p)^5$ or, in general

$$\pi_0(p) = Cp^{\alpha-1}(1-p)^{\beta-1}$$

Then the full likelihood is

$$\begin{aligned}\pi_1(p, X) &= \pi_0(p)L(p, X) \\ &= Cp^{\alpha-1}(1-p)^{\beta-1}p^X(1-p)^{10-X} \\ &= Cp^{\alpha+X-1}(1-p)^{\beta+10-X-1} \\ &\approx \text{Beta}(\alpha + X, \beta + 10 - X)\end{aligned}$$

and

$$\begin{aligned}\widehat{p}_B(X) &= E\big(\text{Beta}(\alpha + X, \beta + 10 - X)\big) \\ &= \frac{X + \alpha}{\alpha + \beta + 10}\end{aligned}$$

This always gives us an answer, but the answer depends on the parameters α, β in the prior. If X and 10 (or their analogs) are much larger than α and β, then the dependence will be small, but there will still be a dependence. A disadvantage of Bayesian methods is that there is rarely an obvious and well-motivated prior.

This interaction between the prior and the likelihood is a special case of a general situation:

Consider a general family of distributions, for example Beta(α, β) or Gam(α, β) or Norm(μ, σ^2).

Let $\pi_0(p, \alpha, \beta)$ be a prior distribution from this family.

Then, we say that this family is a *conjugate prior (family)* for the likelihood $L(p, X)$ if

$$\begin{aligned}\pi_1(p, \alpha, \beta \mid X) &= \pi_0(p, \alpha, \beta) L(p, X) \\ &= \pi_0(p, \alpha_1, \beta_1)\end{aligned}$$

is always a member of the same family, where $\alpha_1 = \alpha(X)$ and $\beta_1 = \beta(X)$. The expressions $\alpha(X)$ and $\beta(X)$ are called the *updating formulas* for α and β.

In this case, the posterior density $\pi_1(p \mid X)$ is

$$\begin{aligned} & Cp^{\alpha-1}(1-p)^{\beta-1}\, p^X(1-p)^{10-X} \\ = \; & Cp^{\alpha+X-1}(1-p)^{\beta+10-X-1} \end{aligned}$$

so that the beta density family is a *conjugate prior* for binomial sampling. The updating formulas are $\alpha_1(X) = \alpha + X$ and $\beta_1(X) = \beta + 10 - X$.

There are many important examples of conjugate priors for different types of likelihoods. What typically makes conjugate priors work is that the likelihood

$$L(p, X) = p^X(1-p)^{10-X}$$

can be viewed either as a density in p or as a density in X, although with different normalizing constants.

Beta densities as a conjugate prior for binomial sampling generalize to multinomial sampling:

Recall that a distribution on the simplex (p_1, p_2, p_3, p_4) (that is, $p_i \geq 0$ and $p_1 + p_2 + p_3 + p_4 = 1$) is a *Dirichlet distribution* $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ if

$$\pi_0(p) = C p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_4^{\alpha_4 - 1}$$

where

$$C = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)\Gamma(\alpha_4)}$$

Suppose that we do n independent multinomial trials and obtain, in some order,

$$n_i \text{ outcomes of Type } i \text{ (prob. } p_i \text{ each)}$$

for $i = 1, 2, 3, 4$ and $X = (n_1, n_2, n_3, n_4)$. Then the likelihood is

$$L(p, X) = p_1^{n_1}\, p_2^{n_2}\, p_3^{n_3}\, p_4^{n_4}$$

If we multiply $L(p, X) = p_1^{n_1} \dots p_4^{n_4}$ by the prior

$$\pi_0(p) = C p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} p_4^{\alpha_4 - 1}$$

we obtain the posterior

$$\pi_0(p \mid X) = C_X p_1^{\alpha_1 + n_1 - 1} p_2^{\alpha_2 + n_2 - 1} p_3^{\alpha_3 + n_3 - 1} p_4^{\alpha_4 + n_4 - 1}$$

This means that the family of Dirichlet distributions are a conjugate prior for *multinomial sampling* with updating formulas $\alpha_i(X) = \alpha_i + n_i$. Since

$$E(p_i) = \frac{\alpha_i}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4}$$

for a Dirichlet $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, we have

$$E(p_i \mid X) = \frac{\alpha_i + n_i}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + n}$$

for the posterior distribution. If the α_i are small, this is close to the MLE $\widehat{p_i}(X) = n_i/n$.

As another example, suppose that $X_1, X_2, \ldots, X_n$ are independent *Poisson* random variables. Then the likelihood is

$$\begin{aligned} L(\mu, X_1, \ldots, X_n) &= \prod_{i=1}^{n} \left(e^{-\mu} \frac{\mu^{X_i}}{X_i!} \right) \\ &= C_X \, e^{-n\mu} \mu^{\left(\sum_{i=1}^{n} X_i\right)} \end{aligned}$$

Note that this is a Poisson density in X, but a gamma density in μ. If we consider a gamma-density prior for μ

$$\pi_0(\mu, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \, \mu^{\alpha-1} \exp(-\beta\mu)$$

Then the posterior distribution is

$$\begin{aligned} \pi_1(\mu \mid X) &= C_X \mu^{\alpha-1} e^{-\beta\mu} e^{-n\mu} \mu^{S(x)} \\ &= C \, \mu^{\alpha+S(X)-1} e^{-(\beta+n)} \end{aligned}$$

where $S(X) = \sum_{i=1}^{n} X_i$.

Thus gamma densities are a conjugate prior for Poisson sampling with updating formulas

$$\alpha(X) = \alpha + \textstyle\sum_{i=1}^{n} X_i, \quad \beta(X) = \beta + n$$

Since the mean of $\mathsf{Gam}(\alpha, \beta)$ is α/β, the Bayes estimator of μ is

$$\widehat{p}_B(X) = E\big(\mathsf{Gam}(\alpha + S(X), \beta + n)\big) = \frac{\alpha + S(X)}{\beta + n}$$

for $S(X) = \sum_{i=1}^{n} X_i$. This is close to the MLE for μ:

$$\widehat{\mu}(X) = \frac{S(X)}{n} = \overline{X}$$

It is typical to set $\pi_0(\mu) = \mathsf{Gam}(\epsilon, \epsilon)$ for $\epsilon = 0.001$. This distribution has mean one but, due to the $\mu^{\epsilon-1}$ singularity at $\mu = 0$, has the vast majority of its mass very close to 0. For this prior, the estimated value of the Poisson mean μ for observed $X = 0$ is

$$\widehat{p}_B(0) = \frac{\epsilon}{\epsilon + n}$$

As yet another example, suppose that $X_1, X_2, \ldots, X_n$ are independent exponentially distributed random variables where r is the *rate* ($E(X_i) = 1/r$). Then

$$L(r, X_1, \ldots, X_n) = \prod_{i=1}^{n} \big(r \exp(-rX_i)\big)$$

$$= \; r^n \exp(-rS(X)), \qquad S(X) = \sum_{i=1}^{n} X_i$$

Since $L(r, X)$ is a gamma density in r, this suggests

$$\pi_0(r, \alpha, \beta) = Cr^{\alpha-1}e^{-\beta r}, \qquad r \geq 0$$

Then the posterior density is

$$\begin{aligned} & C_X \, \pi_0(r, \alpha, \beta) L(r, X) \\ &= \; C_X \, r^{\alpha-1} \exp(-\beta r) \; r^n e^{-rS(X)} \\ &= \; C_X \, r^{\alpha+n-1} \exp\big(-\big(\beta + S(X)\big)r\big) \\ &\approx \; \mathrm{Gam}\big(\alpha + n, \, \beta + S(X)\big) \end{aligned}$$

The Bayes estimator of r is then

$$\widehat{r}_B(X) = E\big(\mathsf{Gam}(\alpha + n,\, \beta + S(X))\big) = \frac{\alpha + n}{\beta + S(X)}$$

If α, β are small, this is close to the MLE $\widehat{r}(X) = 1/\overline{X} = n/S(X)$.

Thus the gamma distribution family is a conjugate prior for both Poisson and exponential sampling, but the role of n and $S(X)$ are reversed in the updating formulas:

For Poisson sampling:

$$\pi_1(\mu, \alpha, \beta \mid X) \;\approx\; \mathsf{Gam}\big(\alpha + S(X),\, \beta + n\big)$$

while for exponential sampling:

$$\pi_1(r, \alpha, \beta \mid X) \;\approx\; \mathsf{Gam}\big(\alpha + n,\, \beta + S(X)\big)$$

Now let's consider a more difficult problem. Assume that

$$X_1, X_2, \ldots, X_n$$

are independent normal $N(\mu, \sigma^2)$. Can we find a conjugate prior for one-dimensional normal sampling with two unknown parameters?

Recall that in the examples before, a likelihood with a single parameter

- p for Bernoulli sampling
- μ for Poisson sampling
- r for exponential sampling

had a conjugate prior with two parameters, Beta(α, β) or Gam(α, β). Thus we should expect a conjugate prior for (μ, σ^2) to have four parameters.

Let's start by writing down the normal likelihood:

$$L(\mu, \sigma^2, X) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(X_i - \mu)^2\Big)$$

$$= \;\Big(\frac{1}{2\pi\sigma^2}\Big)^{n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\Big)$$

As a function of $v = 1/\sigma^2$, which is called the *precision* of X_i as opposed to the variance, this is a gamma density:

$$L(\mu, v, X) \;=\; C_X\, v^{n/2} \exp\Big(-\frac{v}{2}\sum_{i=1}^{n}(X_i - \mu)^2\Big)$$

$$\approx\; \mathsf{Gam}\Big(\frac{n+2}{2},\;\; \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\Big)$$

However, we want a joint density in (μ, v):

$$
\begin{aligned}
L(\mu, v, X) &= C_X\, v^{n/2}\, \mathsf{exp}\big(-\frac{v}{2}\sum_{i=1}^{n}(X_i-\mu)^2\big) \\
&= C_X\, v^{(n-1)/2}\, \mathsf{exp}\big(-\frac{v}{2}\sum_{i=1}^{n}(X_i-\overline{X})^2\big) \\
&\qquad \times \sqrt{\frac{vn}{2\pi}}\, \mathsf{exp}\big(-\frac{vn}{2}(\overline{X}-\mu)^2\big)
\end{aligned}
$$

Note that the last factor is $N(\overline{X}, 1/vn)$ in μ, and that $\int_{-\infty}^{\infty} L(\mu, v, X)d\mu$ is

$$
\begin{aligned}
& C_X\, v^{(n-1)/2}\, \mathsf{exp}\big(-\frac{v}{2}\sum_{i=1}^{n}(X_i-\overline{X})^2\big) \\
\approx\ & \mathsf{Gam}\Big(\frac{n+1}{2},\ \frac{1}{2}\sum_{i=1}^{n}(X_i-\overline{X})^2\Big)
\end{aligned}
$$

This means that $L(\mu, v, X)$ as a density in (μ, v) describes a two-dimensional density for random variables (M, V) defined by, first,

$$V \;\approx\; \text{Gam}\Big(\frac{n+1}{2},\;\; \frac{1}{2}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big)$$

and then, conditional on V, $M \approx N(\overline{X}, \frac{1}{nV})$.

To simulate a two-dimensional random variate (μ, σ^2) from this distribution, we would

(i) first simulate V as above
(ii) set $\sigma^2 = 1/V$, and
(iii) simulate $\mu \approx N(\overline{X}, \sigma^2/n)$

For these reasons, this distribution in (μ, σ^2) (or (μ, v)) is called the *Inverse-Gamma-Normal* distribution.

Usually the inverse-gamma-normal density

$$L(\mu, v, X) \;=\; C_X\, v^{(n-1)/2} \exp\Big(-\frac{v}{2}\sum_{i=1}^{n}(X_i - \overline{X})^2\Big) \times \sqrt{\frac{v}{2\pi}} \exp\Big(-\frac{vn}{2}(\overline{X} - \mu)^2\Big)$$

is used with priors on v and μ:

$$\begin{aligned} \pi_{01}(v, \alpha_\epsilon, \beta_\epsilon) &\approx \mathsf{Gam}(\alpha_\epsilon, \beta_\epsilon) \quad \text{in } v \\ &= C v^{\alpha_\epsilon - 1} e^{-\beta_\epsilon v} \\ \pi_{02}(\mu, \mu_\epsilon, v_\epsilon) &\approx \mathsf{Norm}(\mu_\epsilon, v_\epsilon) \quad \text{in } \mu \\ &= C \exp\big(-(1/2) v_\epsilon (\mu - \mu_\epsilon)^2\big) \end{aligned}$$

for four parameters $\alpha_\epsilon, \beta_\epsilon, \mu_\epsilon, v_\epsilon$, where we ignore factors that don't depend on v or μ. Often $\alpha_\epsilon = \beta_\epsilon = 0.001$, $v_\epsilon = 10^{-6}$, and $\mu_\epsilon = 0$.

For observations $X_1, X_2, \ldots, X_n \approx N(\mu, v)$ in terms of the precision v,

$$\begin{aligned}
&\pi_1(v, \alpha_\epsilon, \beta_\epsilon \mid X) \\
&\quad = \; C v^{\alpha_\epsilon - 1} e^{-\beta_\epsilon v} v^{(n-1)/2} e^{-\frac{1}{2} v \sum_{i=1}^n (X_i - \overline{X})^2} \\
&\quad \approx \; \mathsf{Gam}\big(\alpha_\epsilon + \frac{n-1}{2},\; \beta_\epsilon + \frac{1}{2}\sum_{i=1}^n (X_i - \overline{X})^2\big)
\end{aligned}$$

$$\begin{aligned}
&\pi_1(\mu, \mu_\epsilon, v_\epsilon \mid X, v) \\
&\quad = \; C e^{-(1/2) v_\epsilon (\mu - \mu_\epsilon)^2} e^{-(1/2) v n (\mu - \overline{X})^2} \\
&\quad = \; C \exp\left(-\frac{1}{2}(v_\epsilon + nv)\left(\mu - \frac{v_\epsilon \mu_\epsilon + nv\overline{X}}{v_\epsilon + nv} \right)^2 \right) \\
&\quad \approx \; N\left(\frac{v_\epsilon \mu_\epsilon + nv\overline{X}}{v_\epsilon + nv},\; v_\epsilon + nv \right)
\end{aligned}$$

There are several different ways of setting up conjugate priors for normal sampling. This is only one way.

These formulas leads to the updating formulas for $X_1, X_2, \ldots, X_n \approx N(\mu, v)$:

$$\text{For } v: \quad \alpha_\epsilon \to \alpha_\epsilon + \frac{n-1}{2}$$

$$\beta_\epsilon \to \beta_\epsilon + \frac{1}{2}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$\text{For } \mu: \quad \mu_\epsilon \to \frac{v_\epsilon \mu_\epsilon + nv\overline{X}}{v_\epsilon + nv}$$

$$v_\epsilon \to v_\epsilon + nv$$

Recall that $E(v) = \alpha_\epsilon/\beta_\epsilon$ for $v \approx \text{Gam}(\alpha_\epsilon, \beta_\epsilon)$, so that the first two updates imply $E(v) = 1/s_X^2$ if $\alpha_\epsilon, \beta_\epsilon$ are small.

Thank you for coming.