where $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ is the loss we incur if the true value is $\boldsymbol{\theta}$ but we guess $\hat{\boldsymbol{\theta}}$.

If we use ℓ_2 loss, $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2^2$, then the optimal estimate is the posterior mean, $\overline{\boldsymbol{\theta}} = \mathbb{E}[\theta|\mathcal{D}]$, as we show in Sec. 8.1.5.1. If we use ℓ_1 loss, $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = |\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}|_1$, then the optimal estimate is the posterior median, as we show in Sec. 8.1.5.2. If we use 0-1 loss, $\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathbb{I}\left(\theta = \hat{\theta}\right)$, then the optimal estimate is the posterior mode or MAP estimate, $\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D})$, as we show in Sec. 8.1.2.1. This is the easiest point estimate to compute, since it just requires optimization, and not integration. However, 0-1 loss is a very unnatural loss function to use for parameter esimation, which are real-valued vectors.

7.1.2.2 Credible intervals

We often want a measure of confidence in our parameter estimates. A standard measure of confidence in some (scalar) quantity θ is the "width" of its posterior distribution. This can be measured using a $100(1-\alpha)\%$ **credible interval**[1] which is a (contiguous) region $C = (\ell, u)$ (standing for lower and upper) which contains $1-\alpha$ of the posterior probability mass, i.e.,

$$C_\alpha(\mathcal{D}) = (\ell, u) : P(\ell \leq \theta \leq u|\mathcal{D}) = 1 - \alpha \tag{7.3}$$

There may be many intervals that satisfy Eq. (7.3), so we usually choose one such that there is $(1-\alpha)/2$ mass in each tail; this is called a **central interval**. If the posterior has a known functional form, we can compute the posterior central interval using $\ell = F^{-1}(\alpha/2)$ and $u = F^{-1}(1-\alpha/2)$, where F is the cdf of the posterior, and F^{-1} is the inverse cdf. For example, if the posterior is Gaussian, $p(\theta|\mathcal{D}) = \mathcal{N}(0, 1)$, and $\alpha = 0.05$, then we have $\ell = \Phi(\alpha/2) = -1.96$, and $u = \Phi(1-\alpha/2) = 1.96$, where Φ denotes the cdf of the Gaussian. This is illustrated in Fig. 3.6b. This justifies the common practice of quoting a credible interval in the form of $\mu \pm 2\sigma$, where μ represents the posterior mean, σ represents the posterior standard deviation, and 2 is a good approximation to 1.96.

In general, it is often hard to compute the inverse cdf of the posterior. In this case, a simple alternative is to draw samples from the posterior, and then to use a Monte Carlo approximation to the posterior quantiles: we simply sort the S samples, and find the one that occurs at location α/S along the sorted list. As $S \to \infty$, this converges to the true quantile. See beta_credible_int_demo.py for a demo of this.

A problem with central intervals is that there might be points outside the central interval which have higher probability than points that are inside, as illustratedin Figure 7.1(a). This motivates an alternative quantity known as the **highest posterior density** or **HPD** regionm which is the set of points which have a probability above some threshold. More precisely we find the threshold p^* on the pdf such that

$$1 - \alpha = \int_{\theta: p(\theta|\mathcal{D}) > p^*} p(\theta|\mathcal{D}) d\theta \tag{7.4}$$

and then define the HPD as

$$C_\alpha(\mathcal{D}) = \{\theta : p(\theta|\mathcal{D}) \geq p^*\} \tag{7.5}$$

1. A credible interval is not the same as a confidence interval, which is a concept from frequentist statistics which we discuss in Sec. E.3.4.

Summary of Comments on pml1.pdf

Page: 174

Author: petercerno Subject: Comment on Text Date: 05.02.21, 15:03:44
Should be in bold (as a vector)

Author: petercerno Subject: Comment on Text Date: 05.02.21, 15:06:16
\Phi^{-1}

Author: petercerno Subject: Comment on Text Date: 05.02.21, 15:06:28
\Phi^{-1}

Author: petercerno Subject: Comment on Text Date: 05.02.21, 15:20:05
Nit: Only posterior median F^{-1}(0.5) is guaranteed to be inside the central interval, although for Gaussian the median coincides with the posterior mean. So in this example it does not matter.
In general, there are distributions, in which the mean is way different from the median and outside the central interval.

Author: petercerno Subject: Cross-Out Date: 05.02.21, 15:20:22

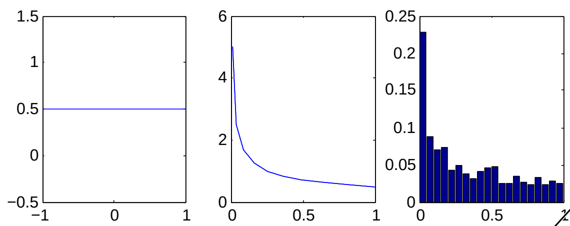


Figure 7.3: Computing the distribution of $z = y^2$, where $p(y)$ is uniform (left). The analytic result is shown in the middle, and the Monte Carlo approximation is shown on the right. Generated by change_of_vars_demo1d.py.

7.1.2.3 Posterior samples

Often we want to compute the expected value of some function of a random variable, $Z = f(Y)$. We can approximate this by drawing many samples of Y from its (posterior) distribution, and then using

$$\mathbb{E}[f(Y)] = \int f(y)p(y)dy \approx \frac{1}{S}\sum_{s=1}^{S} f(y_s) \tag{7.6}$$

This is called **Monte Carlo integration**, and has the advantage over numerical integration (which is based on evaluating the function at a fixed grid of points) that the function is only evaluated in places where there is non-negligible probability. Thus MC integration scales better to high dimensional problems.

For example, suppose $y \sim \text{Unif}(-1, 1)$ and $z = f(x) = y^2$. The exact mean is given by

$$\mathbb{E}[z] = \int_{-1}^{-1} \frac{1}{2} y^2 dy = \frac{1}{2}[\frac{y^3}{3}]_{-1}^{1} = 1/3 \tag{7.7}$$

We can approximate this by drawing many samples from $p(y)$, squaring them, and then averaging; we find $\mathbb{E}[f] = 0.34$, as shown in Fig. 7.3.

By varying the function f, we can approximate many quantities of interest, such as the mean,

Page: 176

Author: petercerno Subject: Comment on Text Date: 05.02.21, 15:32:05

This is some weird symbol

7.2.1 The beta-binomial model

Suppose we toss a coin N times, and want to infer the probability of heads. Let $y_n = 1$ denote the event that the n'th trial was heads, $y_n = 0$ represent the event that the n'th trial was tails, and let $\mathcal{D} = \{y_n : n = 1 : N\}$ be all the data. We assume $y_n \sim \text{Ber}(\theta)$, where $\theta \in [0, 1]$ is the rate parameter (probability of heads). In this section, we discuss how to compute $p(\theta|\mathcal{D})$.

7.2.1.1 Likelihood

We assume the data are **iid** or **independent and identically distributed**. Thus the likelihood has the form

$$p(\mathcal{D}|\theta) = \prod_{n=1}^{N} \theta^{y_n}(1-\theta)^{1-y_n} = \theta^{N_1}(1-\theta)^{N_0} \quad (7.11)$$

where we have defined $N_1 = \sum_{n=1}^{N} \mathbb{I}(y_n = 1)$ and $N_0 = \sum_{n=1}^{N} \mathbb{I}(y_n = 0)$, representing the number of heads and tails. These counts are called the **sufficient statistics** of the data, since this is all we need to know about $\mathcal{D}$ to infer θ. The total count, $N = N_0 + N_1$, is called the sample size.

Note that we can also consider a Binomial likelihood model, in which we perform N trials and observe the number of heads, y, rather than observing a sequence of coin tosses. Now the likelihood has the following form:

$$p(\mathcal{D}|\theta) = \text{Bin}(y|N, \theta) \binom{N}{y} \theta^y (1-\theta)^{N-y} \quad (7.12)$$

The scaling factor $\binom{N}{y}$ is independent of θ, so we can ignore it. Thus this likelihood is proportional to the Bernoulli likelihood in Eq. (7.11), so our inferences about θ will be the same for both models.

7.2.1.2 Prior

To simplify the computations, we will assume that the prior $p(\boldsymbol{\theta}) \in \mathcal{F}$ is a conjugate prior for the likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$. This means that the posterior is in the same parameterized family as the prior, i.e., $p(\boldsymbol{\theta}|\mathcal{D}) \in \mathcal{F}$.

To ensure this property when using the Bernoulli (or Binomial) likelihood, we should use a prior of the following form:

$$p(\theta) \propto \theta^{\breve{\alpha}-1}(1-\theta)^{\breve{\beta}-1} \quad (7.13)$$

We recognize this as the pdf of a beta distribution (see Sec. 3.4.4).

7.2.1.3 Posterior

If we multiply the Bernoulli likelihood in Eq. (7.11) with the beta prior in Eq. (3.66) we get a beta posterior:

$$p(\theta|\mathcal{D}) \propto \theta^{N_1}(1-\theta)^{N_0}\, \theta^{\breve{\alpha}-1}(1-\theta)^{\breve{\beta}-1} \quad (7.14)$$

$$\propto \text{Beta}(\theta|\, \breve{\alpha} + N_1, \breve{\beta} + N_0) \quad (7.15)$$

$$= \text{Beta}(\theta|\, \widehat{\alpha}, \widehat{\beta}) \quad (7.16)$$

Page: 178

Author: petercerno Subject: Comment on Text Date: 05.02.21, 15:59:18
Missing =

Author: petercerno Subject: Cross-Out Date: 05.02.21, 15:59:40

where $\widehat{N} = \widehat{\beta} + \widehat{\alpha}$ is the strength (equivalent sample size) of the posterior. We will now show that the posterior mean is a convex combination of the prior mean, $m = \breve{\alpha} / \breve{N}$ (where $\breve{N} \triangleq \breve{\alpha} + \breve{\beta}$ is the prior strength), and the MLE: $\hat{\theta}_{\text{mle}} = \frac{N_1}{N}$:

$$\mathbb{E}\left[\theta|\mathcal{D}\right] = \frac{\breve{\alpha} + N_1}{\breve{\alpha} + N_1 + \breve{\beta} + N_0} = \frac{\breve{N}\, m + N_1}{N + \breve{N}} = \frac{\breve{N}}{N + \breve{N}} m + \frac{N}{N + \breve{N}} \frac{N_1}{N} = \lambda m + (1 - \lambda)\hat{\theta}_{\text{mle}} \tag{7.19}$$

where $\lambda = \frac{\breve{N}}{\widehat{N}}$ is the ratio of the prior to posterior equivalent sample size. So the weaker the prior, the smaller is λ, and hence the closer the posterior mean is to the MLE.

To capture some notion of uncertainty in our estimate, a common approach is to compute the **standard error** of our estimate, which is just the posterior standard deviaton:

$$\text{se}(\theta) = \sqrt{\mathbb{V}\left[\theta|\mathcal{D}\right]} \tag{7.20}$$

In the case of the Bernoulli model, we showed that the posterior is a beta distribution. The variance of the beta posterior is given by

$$\mathbb{V}\left[\theta|\mathcal{D}\right] = \frac{\widehat{\alpha}\widehat{\beta}}{(\widehat{\alpha} + \widehat{\beta})^2(\widehat{\alpha} + \widehat{\beta} + 1)} \tag{7.21}$$

where $\widehat{\alpha} = \breve{\alpha} + N_1$ and $\widehat{\beta} = \breve{\beta} + N_0$. If $N \gg \breve{\alpha} + \breve{\beta}$, this simplifies to

$$\mathbb{V}\left[\theta|\mathcal{D}\right] \approx \frac{N_1 N_0}{N^3} = \frac{\hat{\theta}(1 - \hat{\theta})}{N} \tag{7.22}$$

where $\hat{\theta}$ is the MLE. Hence the standard error is given by

$$\sigma = \sqrt{\mathbb{V}\left[\theta|\mathcal{D}\right]} \approx \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{N}} \tag{7.23}$$

We see that the uncertainty goes down at a rate of $1/\sqrt{N}$. We also see that the uncertainty (variance) is maximized when $\hat{\theta} = 0.5$, and is minimized when $\hat{\theta}$ is close to 0 or 1. This makes sense, since it is easier to be sure that a coin is biased than to be sure that it is fair.

7.2.1.4 Posterior predictive

Suppose we want to predict future observations. A very common approach is to first compute an estimate of the parameters based on training data, $\hat{\boldsymbol{\theta}}(\mathcal{D})$, and then to plug that parameter back into the model and use $p(y|\hat{\boldsymbol{\theta}})$ to predict the future; this is called a **plug-in approximation**. However, this can result in overfitting. As an extreme example, suppose we have seen $N = 3$ heads in a row. The MLE is $\hat{\theta} = 3/3 = 1.0$. However, if we use this estimate, we would predict that tails are impossible.

One solution to this is to compute a MAP estimate, and plug that in, as we discussed in Sec. 4.4.1. Here we discuss a fully Bayesian solution, in which we marginalize out θ.

Page: 180

Author: petercerno Subject: Comment on Text Date: 05.02.21, 16:17:09

Side note: It is interesting that even with uninformative prior we would not get E[\theta | D] equal to MLE. The reason is that MLE is not the mean of the distribution, but rather the mode of the distribution (with an uninformative prior). (This point could be worth emphasizing).

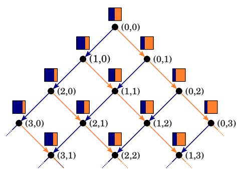


Figure 7.6: Illustration of sequential Bayesian updating for the beta-Bernoulli model. Each colored box represents the predicted distribution $p(x_t|\mathbf{h}_t)$, where $\mathbf{h}_t = (N_{1,t}, N_{0,t})$ is the sufficient statistic derived from history of observations up until time t, namely the total number of heads and tails. The probability of heads (blue bar) is given by $p(x_t = 1|\mathbf{h}_t) = (N_{t,1} + 1)/(t + 2)$, assuming we start with a uniform $\text{Beta}(\theta|1, 1)$ prior. From Figure 3 of [Ort+19]. Used with kind permission of Pedro Ortega.

Bernoulli model

For the Bernoulli model, the resulting **posterior predictive distribution** has the form

$$p(y = 1|\mathcal{D}) = \int_0^1 p(y = 1|\theta)p(\theta|\mathcal{D})d\theta \tag{7.24}$$

$$= \int_0^1 \theta \, \text{Beta}(\theta| \,\widehat{\alpha}, \widehat{\beta})d\theta = \mathbb{E}\left[\theta|\mathcal{D}\right] = \frac{\widehat{\alpha}}{\widehat{\alpha} + \widehat{\beta}} \tag{7.25}$$

In Sec. 4.4.1, we had to use the Beta(2,2) prior to recover add-one smoothing, which is a rather unnatural prior. In the Bayesian approach, we can get the same effect using a uniform prior, $p(\theta) = \text{Beta}(\theta|1, 1)$, since the predictive distribution becomes

$$p(y = 1|\mathcal{D}) = \frac{N_1 + 1}{N_1 + N_0 + 2} \tag{7.26}$$

This is known as **Laplace's rule of succession**. See Fig. 7.6 for an illustration of this in the sequential setting.

Binomial model

Now suppose we were interested in predicting the number of heads in $M > 1$ future coin tossing trials, i.e., we are using the binomial model instead of the Bernoulli model. The posterior over θ is the same as before, but the posterior predictive distribution is different:

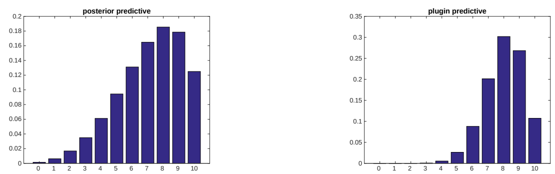
$$p(y|\mathcal{D}, M) = \int_0^1 \text{Bin}(y|M, \theta)\text{Beta}(\theta|\widehat{\alpha}, \widehat{\beta})d\theta \tag{7.27}$$

$$= \binom{M}{y} \frac{1}{B(\widehat{\alpha}, \widehat{\beta})} \int_0^1 \theta^x (1 - \theta)^{M-x} \theta^{\widehat{\alpha}-1}(1 - \theta)^{\widehat{\beta}-1} d\theta \tag{7.28}$$

Page: 181

Author: petercerno Subject: Comment on Text Date: 17.02.21, 16:18:55
y

Author: petercerno Subject: Comment on Text Date: 17.02.21, 16:18:49
y



(a) *(b)*

Figure 7.7: (a) Posterior predictive distributions for 10 future trials after seeing $N_1 = 4$ heads and $N_0 = 1$ tails. (b) Plug-in approximation based on the same data. In both cases, we use a uniform prior. Generated by beta_binom_post_pred_plot.py.

We recognize the integral as the normalization constant for a Beta($\widehat{\alpha} + y, M - y + \widehat{\beta}$) distribution. Hence

$$\int_0^1 \theta^{y+\widehat{\alpha}-1}(1-\theta)^{M-y+\widehat{\beta}-1} d\theta = B(y+\widehat{\alpha}, M-y+\widehat{\beta}) \tag{7.29}$$

Thus we find that the posterior predictive is given by the following, known as the (compound) **beta-binomial** distribution:

$$Bb(x|M, \widehat{\alpha}, \widehat{\beta}) \triangleq \binom{M}{x} \frac{B(x+\widehat{\alpha}, M-x+\widehat{\alpha})}{B(\widehat{\alpha}, \widehat{\beta})} \tag{7.30}$$

In Fig. 7.7(a), we plot the posterior predictive density for $M = 10$ after seeing $N_1 = 4$ heads and $N_0 = 1$ tails, when using a uniform Beta(1,1) prior. In Fig. 7.7(b), we plot the plug-in approximation, given by

$$p(\theta|\mathcal{D}) \approx \delta(\theta - \hat{\theta}) \tag{7.31}$$

$$p(y|\mathcal{D}, M) = \int_0^1 \text{Bin}(y|M,\theta)p(\theta|\mathcal{D})d\theta = \text{Bin}(y|M,\hat{\theta}) \tag{7.32}$$

where $\hat{\theta}$ is the MAP estimate. Looking at Fig. 7.7, we see that the Bayesian prediction has longer tails, spreading its probablity mass more widely, and is therefore less prone to overfitting and black-swan type paradoxes. (Note that we use a uniform prior in both cases, so the difference is not arising due to the use of a prior; rather, it is due to the fact that the Bayesian approach integrates out the unknown parameters when making its predictions.)

7.2.1.5 Marginal likelihood

The **marginal likelihood** or **evidence** for a model $\mathcal{M}$ is defined as

$$p(\mathcal{D}|\mathcal{M}) = \int p(\boldsymbol{\theta}|\mathcal{M})p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})d\boldsymbol{\theta} \tag{7.33}$$

Page: 182

Author: petercerno Subject: Comment on Text Date: 17.02.21, 16:19:58

\beta

Prior

The conjugate prior is known as the **inverse Wishart** distribution, which is a distribution over positive definite matrices. This has the following pdf:

$$\mathrm{IW}(\boldsymbol{\Sigma}|\,\breve{\mathbf{S}},\breve{\nu}) \propto |\boldsymbol{\Sigma}|^{-(\breve{\nu}+D+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\breve{\mathbf{S}}\,\boldsymbol{\Sigma}^{-1})\right) \tag{7.107}$$

Here $\breve{\nu} > D-1$ is the degrees of freedom (dof), and $\breve{\mathbf{S}}$ is a symmetric pd matrix. We see that $\breve{\mathbf{S}}$ plays the role of the prior scatter matrix, and $N_0 \triangleq \breve{\nu} + D + 1$ controls the strength of the prior, and hence plays a role analogous to the sample size N.

If $D = 1$, the inverse Wishart reduces to the inverse Gamma:

$$\mathrm{IW}(\sigma^2|s^{-1},\nu) = \mathrm{IG}(\sigma^2|\nu/2, s/2) \tag{7.108}$$

If $s = 1$, this reduces to the inverse chi-squared distribution.

Posterior

Multiplying the likelihood and prior we find that the posterior is also inverse Wishart:

$$\begin{aligned}
p(\boldsymbol{\Sigma}|\mathcal{D},\boldsymbol{\mu}) &\propto |\boldsymbol{\Sigma}|^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_\mu)\right)|\boldsymbol{\Sigma}|^{-(\breve{\nu}+D+1)/2} \\
&\quad \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\,\breve{\mathbf{S}})\right) && (7.109) \\
&= |\boldsymbol{\Sigma}|^{-\frac{N+(\breve{\nu}+D+1)}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{S}_\mu + \breve{\mathbf{S}})\right]\right) && (7.110) \\
&= \mathrm{IW}(\boldsymbol{\Sigma}|\,\widehat{\mathbf{S}},\widehat{\nu}) && (7.111) \\
\widehat{\nu} &= \breve{\nu} + N && (7.112) \\
\widehat{\mathbf{S}} &= \breve{\mathbf{S}} + \mathbf{S}_\mu && (7.113)
\end{aligned}$$

In words, this says that the posterior strength $\widehat{\nu}$ is the prior strength $\breve{\nu}$ plus the number of observations N, and the posterior scatter matrix $\widehat{\mathbf{S}}$ is the prior scatter matrix $\breve{\mathbf{S}}$ plus the data scatter matrix $\mathbf{S}_\mu$.

We posterior mode of the inverse Wishart can be used to derive the shrinkage estimate for $\boldsymbol{\Sigma}$ discussed in Sec. 4.4.2.

7.2.4.3 LKJ prior

The conjugate prior for a covariance matrix is the inverse Wishart.However, this distribution is hard to sample from. Consequently it is common to use the **LKJ distribution** as a prior instead (this is named after the authors of [LKJ09]). This is a prior on correlation matrices $\mathbf{C}$ of the form $p(\mathbf{C}|\eta) \propto |\mathbf{C}|^{\eta-1}$, so $\eta = 1$ leads to a uniform a distribution. We can combine this with a suitable prior for the standard deviation, such as half-Cauchy prior (Sec. 7.2.3.5), to derive a prior for $\boldsymbol{\Sigma} = \sigma\mathbf{C}$.

7.2.4.4 Posterior of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$

In this section, we compute $p(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathcal{D})$ using a conjugate prior.

Page: 195

Author: petercerno Subject: Comment on Text Date: 17.02.21, 16:58:05
The

Author: petercerno Subject: Comment on Text Date: 17.02.21, 16:58:36
Nit: Missing space.

where

$$\mathbf{M} \triangleq N(\boldsymbol{\mu} - \overline{\mathbf{y}})(\boldsymbol{\mu} - \overline{\mathbf{y}})^\top + \breve{\kappa}\, (\boldsymbol{\mu} - \breve{\mathbf{m}})(\boldsymbol{\mu} - \breve{\mathbf{m}})^\top + \mathbf{S}_{\overline{\mathbf{y}}} + \breve{\mathbf{S}} \tag{7.127}$$

$$= (\breve{\kappa} + N)\boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}(\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{y}})^\top - (\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{x}})\boldsymbol{\mu}^\top + \breve{\kappa}\breve{\mathbf{m}}\breve{\mathbf{m}}^\top + \mathbf{S}_0 + \breve{\mathbf{S}} \tag{7.128}$$

We can simplify the $\mathbf{M}$ matrix using a trick called completing the square (Sec. B.2.1.6). Applying this to the above, we have

$$(\breve{\kappa} + N)\boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}(\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{y}})^\top - (\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{x}})\boldsymbol{\mu}^\top \tag{7.129}$$

$$= (\breve{\kappa} + N)\left(\boldsymbol{\mu} - \frac{\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{y}}}{\breve{\kappa} + N}\right)\left(\boldsymbol{\mu} - \frac{\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{x}}}{\breve{\kappa} + N}\right)^\top \tag{7.130}$$

$$- \frac{(\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{x}})(\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{y}})^\top}{\breve{\kappa} + N} \tag{7.131}$$

$$= \widehat{\kappa}\, (\boldsymbol{\mu} - \widehat{\mathbf{m}})(\boldsymbol{\mu} - \widehat{\mathbf{m}})^\top - \widehat{\kappa}\widehat{\mathbf{m}}\widehat{\mathbf{m}}^\top \tag{7.132}$$

Hence we can rewrite the posterior as follows:

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) \propto |\boldsymbol{\Sigma}|^{(\widehat{\nu}+D+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\widehat{\kappa}\, (\boldsymbol{\mu} - \widehat{\mathbf{m}})(\boldsymbol{\mu} - \widehat{\mathbf{m}})^\top + \widehat{\mathbf{S}}\right)\right]\right) \tag{7.133}$$

$$= \mathrm{NIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\, \widehat{\mathbf{m}}, \widehat{\kappa}, \widehat{\nu}, \widehat{\mathbf{S}}) \tag{7.134}$$

where

$$\widehat{\mathbf{m}} = \frac{\breve{\kappa}\breve{\mathbf{m}} + N\overline{\mathbf{y}}}{\widehat{\kappa}} = \frac{\breve{\kappa}}{\breve{\kappa} + N}\, \breve{\mathbf{m}} + \frac{N}{\breve{\kappa} + N}\overline{\mathbf{y}} \tag{7.135}$$

$$\widehat{\kappa} = \breve{\kappa} + N \tag{7.136}$$

$$\widehat{\nu} = \breve{\nu} + N \tag{7.137}$$

$$\widehat{\mathbf{S}} = \breve{\mathbf{S}} + \mathbf{S}_{\overline{\mathbf{y}}} + \frac{\breve{\kappa}\, N}{\breve{\kappa} + N}(\overline{\mathbf{y}} - \breve{\mathbf{m}})(\overline{\mathbf{x}} - \breve{\mathbf{m}})^\top \tag{7.138}$$

$$= \breve{\mathbf{S}} + \mathbf{S}_0 + \breve{\kappa}\breve{\mathbf{m}}\breve{\mathbf{m}}^\top - \widehat{\kappa}\widehat{\mathbf{m}}\widehat{\mathbf{m}}^\top \tag{7.139}$$

This result is actually quite intuitive: the posterior mean $\widehat{\mathbf{m}}$ is a convex combination of the prior mean and the MLE; the posterior scatter matrix $\widehat{\mathbf{S}}$ is the prior scatter matrix $\breve{\mathbf{S}}$ plus the empirical scatter matrix $\mathbf{S}_{\overline{\mathbf{y}}}$ plus an extra term due to the uncertainty in the mean (which creates its own virtual scatter matrix); and the posterior confidence factors $\widehat{\kappa}$ and $\widehat{\nu}$ are both incremented by the size of the data we condition on.

Posterior marginals

We have computed the joint posterior

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \mathcal{D})p(\boldsymbol{\Sigma}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}|\, \widehat{\mathbf{m}}, \frac{1}{\widehat{\kappa}}\boldsymbol{\Sigma})\mathrm{IW}(\boldsymbol{\Sigma}|\, \widehat{\mathbf{S}}, \widehat{\nu}) \tag{7.140}$$

We now discuss how to compute the posterior marginals, $p(\boldsymbol{\Sigma}|\mathcal{D})$ and $p(\boldsymbol{\mu}|\mathcal{D})$.

Page: 198

Author: petercerno Subject: Comment on Text Date: 18.02.21, 13:44:59

y