MLE and MAP Examples

1 Multinomial Distribution

Given some integer $k > 1$, let Θ be the set of vectors $\theta = (\theta_1, ..., \theta_k)$ satisfying $\theta_i \geq 0$ and $\sum_{i=1}^k \theta_i = 1$. For any $\theta \in \Theta$ we define the probability mass function

$$p_\theta(x) = \begin{cases} \prod_{i=1}^k \theta_i^{I(x=i)} & \text{for} \quad x \in \{1, 2, ..., k\} \\ 0 & \text{o.w.} \end{cases}$$

Given n observations $X_1, ..., X_n \in \{1, ..., k\}$, we would like to derive the maximum likelihood estimate for the parameter θ under this model.

First, let's write down the likelihood of the data for some $\theta \in \Theta$ (recall that we have assumed $X_1, ..., X_n \in \{1, ..., k\}$, since otherwise there is no meaningful solution):

$$\begin{aligned} \mathcal{L}(\theta; X_1, ..., X_n) &= \prod_{j=1}^n p_\theta(X_j) \\ &= \prod_{j=1}^n \prod_{i=1}^k \theta_i^{I(X_j=i)} \\ &= \prod_{i=1}^k \prod_{j=1}^n \theta_i^{I(X_j=i)} \\ &= \prod_{i=1}^k \theta_i^{S_i} \end{aligned}$$

where, for brevity, we have defined $S_i = \sum_{j=1}^n I(X_j = i)$. Our goal is to maximize $\mathcal{L}$ with respect to θ, subject to the constraint that $\theta \in \Theta$. Equivalently, we can maximize the log likelihood:

$$\log \mathcal{L}(\theta; X_1, ..., X_n) = \sum_{i=1}^k S_i \log \theta_i.$$

Introducing a Lagrange multiplier for the constraint $\sum_{i=1}^k \theta_i = 1$, we have

$$\Lambda(\theta, \lambda) = \sum_{i=1}^k S_i \log \theta_i + \lambda \left(\sum_{i=1}^k \theta_i - 1 \right).$$

(we need not worry about the positivity constraints, since the solution will satisfy those regardless). Differentiating with respect to θ_i and λ, for each $i = 1, ..., k$ we have

$$\frac{\partial \Lambda}{\partial \theta_i} = \frac{S_i}{\theta_i} + \lambda$$

and $\frac{\partial \Lambda}{\partial \lambda} = \sum_{i=1}^{k} \theta_i - 1$. Setting the latter partial derivative to 0 gives back the original constraint $\sum_{i=1}^{k} \theta_i = 1$, as expected. Also for $i = 1, ..., k$,

$$\begin{aligned} \frac{\partial \Lambda}{\partial \theta_i} = 0 \quad &\Rightarrow \quad \frac{S_i}{\widehat{\theta}_i} = -\lambda \\ &\Rightarrow \quad \widehat{\theta}_i = \frac{S_i}{-\lambda}. \end{aligned} \tag{1}$$

By definition of S_i we have

$$\begin{aligned} \sum_{i=1}^{k} \frac{S_i}{-\lambda} &= \frac{\sum_{i=1}^{k} S_i}{-\lambda} \\ &= \frac{n}{-\lambda} \end{aligned}$$

which implies that in order for the summation constraint on θ_i to be satisfied, we require $\frac{n}{-\lambda} = 1$, i.e. $\lambda = -n$.

Plugging this value for λ into (1),

$$\begin{aligned} \widehat{\theta}_i &= \frac{S_i}{n} \\ &= \frac{1}{n} \sum_{j=1}^{n} X_j \end{aligned}$$

i.e. the maximum likelihood estimates for the elements of θ are simply the intuitively obvious estimators – the empirical means.

2 Multivariate Normal (unknown mean and variance)

The PDF of the multivariate normal in d dimensions is

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\}$$

where the parameters are the mean vector $\mu \in \mathbb{R}^d$, and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, which must be symmetric positive definite.

2.1 MLE

The likelihood function given $X_1, ..., X_n \in \mathbb{R}^d$ is

$$\begin{aligned} \mathcal{L}(\mu, \Sigma; X_1, ..., X_n) &= c_1 |\Sigma|^{-n/2} \prod_{i=1}^{n} \exp \left\{ -\frac{1}{2} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \right\} \\ &= c_1 |\Sigma|^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \right\} \end{aligned}$$

where $c_1 = (2\pi)^{-nd/2}$ is a constant independent of the data and parameters and can be ignored.

The log likelihood is

$$\log \mathcal{L}(\mu, \Sigma; X_1, ..., X_n) = -\frac{n}{2} \log |\Sigma| - \frac{1}{2} \left(\sum_{i=1}^{n} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \right) + c_2$$

where c_2 is another inconsequential constant.

It is easiest to first maximize this with respect to μ. The corresponding partial derivative is

$$\begin{aligned} \frac{\partial}{\partial \mu} \log \mathcal{L} &= -\sum_{i=1}^{n} \Sigma^{-1} (X_i - \mu) \\ &= -\Sigma^{-1} \sum_{i=1}^{n} (X_i - \mu) \\ &= n\Sigma^{-1} \left(\mu - \frac{1}{n} \sum_{i=1}^{n} X_i \right) \end{aligned}$$

which implies

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The partial derivative of the log likelihood with respect to Σ is

$$\begin{aligned} \frac{\partial}{\partial \Sigma} \log \mathcal{L} &= -\frac{n}{2} \frac{1}{|\Sigma|} |\Sigma| \Sigma^{-1} - \frac{1}{2} \left(\sum_{i=1}^{n} -\Sigma^{-1} (X_i - \mu)(X_i - \mu)^T \Sigma^{-1} \right) \\ &= -\frac{n}{2} \Sigma^{-1} + \frac{1}{2} \Sigma^{-1} \left(\sum_{i=1}^{n} (X_i - \mu)(X_i - \mu)^T \right) \Sigma^{-1}. \end{aligned}$$

Equating this to 0 and plugging in the estimate $\widehat{\mu}$, we see that the estimator $\widehat{\Sigma}$ must solve the equation

$$n\widehat{\Sigma}^{-1} = \widehat{\Sigma}^{-1} \left(\sum_{i=1}^{n} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T \right) \widehat{\Sigma}^{-1}.$$

It is easy to see that a solution for $\widehat{\Sigma}$ is

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \widehat{\mu})(X_i - \widehat{\mu})^T,$$

which is known as the sample covariance matrix.

2.2 MAP under the conjugate prior

The conjugate prior for the mean and covariance of a multivariate normal is sometimes called the Normal-inverse-Wishart distribution and has the density

$$f(\mu, \Sigma | \mu_0, \beta, \Psi, \nu) = p(\mu | \mu_0, \beta\Sigma) w(\Sigma | \Psi, \nu)$$

where p is the density of the multivariate normal distribution, and w is the density of the inverse-Wishart distribution given by

$$w(\Sigma|\Psi,\nu) = \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d\left(\frac{\nu}{2}\right)}|\Sigma|^{-(\nu+d+1)/2}\exp\left\{-\frac{1}{2}\operatorname{Tr}(\Psi\Sigma^{-1})\right\}$$

where Γ_d is the multivariate Gamma function, and $\operatorname{Tr}(\cdot)$ denotes the trace of a matrix. The parameters of the Normal-inverse-Wishart are $\mu_0 \in \mathbb{R}^d$, $\beta \in \mathbb{R}^+$, $\Psi \in \mathbb{R}^{d\times d}$ positive definite, and $\nu \in \mathbb{R}$ with $\nu > d-1$. We need to

1. calculate the posterior distribution of μ and Σ assuming this prior and n observations $X_1, ..., X_n \in \mathbb{R}^d$;
2. convince ourselves that the posterior is indeed a Normal-inverse-Wishart distribution and find its parameters;
3. and finally find the values for μ and Σ that maximize the posterior distribution.

2.2.1 Calculating the posterior distribution

We have

$$f(\mu,\Sigma|\mu_0,\beta,\Psi,\nu,X_1,...,X_n) \propto \left(\prod_{i=1}^n p(X_i|\mu,\Sigma)\right) f(\mu,\Sigma|\mu_0,\beta,\Psi,\nu)$$

where we omit the term in the denominator which is a finite, non-zero constant that doesn't depend on μ or Σ, since any such term does not affect the shape of the posterior distribution and only factors in the normalizing constant that ensures the posterior integrates to 1. Continuing from above,

$$\begin{aligned}
f(\mu,\Sigma|\mu_0,\beta,\Psi,\nu,X_1,...,X_n) \propto & \left(\prod_{i=1}^n \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right\}\right)\times & (2)\\
& \frac{1}{(2\pi)^{d/2}|\beta\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mu-\mu_0)^T(\beta\Sigma)^{-1}(\mu-\mu_0)\right\}\times \\
& \frac{|\Psi|^{\nu/2}}{2^{\nu d/2}\Gamma_d\left(\frac{\nu}{2}\right)}|\Sigma|^{-(\nu+d+1)/2}\exp\left\{-\frac{1}{2}\operatorname{Tr}(\Psi\Sigma^{-1})\right\} & (3)\\
\propto & \exp\left\{-\frac{1}{2}\sum_{i=1}^n(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\right\}\times \\
& \exp\left\{-\frac{1}{2}(\mu-\mu_0)^T(\beta\Sigma)^{-1}(\mu-\mu_0)\right\}\times \\
& |\Sigma|^{-(\nu+n+d+2)/2}\exp\left\{-\frac{1}{2}\operatorname{Tr}(\Psi\Sigma^{-1})\right\}. & (4)
\end{aligned}$$

While messy, this fully defines the posterior distribution.

2.2.2 The posterior is a Normal-inverse-Wishart distribution

Our goal now is to find $\mu_0', \beta', \Psi', \nu'$ such that (4) looks like a Normal-inverse-Wishart density with those parameters. First we write out explicitly the form of the Normal-inverse-Wishart density, up to constants, for these as of yet unknown parameters:

$$f(\mu, \Sigma | \mu_0', \beta', \Psi', \nu') \propto \exp\left\{-\frac{1}{2}(\mu - \mu_0')^T(\beta'\Sigma)^{-1}(\mu - \mu_0')\right\} \times \\ |\Sigma|^{-(\nu'+d+2)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}(\Psi'\Sigma^{-1})\right\}. \quad (5)$$

It is immediately clear that the only value ν' can take to satisfy $f(\mu, \Sigma | \mu_0, \beta, \Psi, \nu, X_1, ..., X_n) = f(\mu, \Sigma | \mu_0', \beta', \Psi', \nu')$ is

$$\nu' = \nu + d.$$

Furthermore we see that for this value of ν' the term involving $|\Sigma|$ in (4) is accounted for in (5). So now we only need find μ_0', β', Ψ' such that

$$\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^T\Sigma^{-1}(X_i - \mu)\right\} \exp\left\{-\frac{1}{2}(\mu - \mu_0)^T(\beta\Sigma)^{-1}(\mu - \mu_0)\right\} \times \\ \times \exp\left\{-\frac{1}{2}\operatorname{Tr}(\Psi\Sigma^{-1})\right\} \quad (6)$$

is equal to

$$\exp\left\{-\frac{1}{2}(\mu - \mu_0')^T(\beta'\Sigma)^{-1}(\mu - \mu_0')\right\} \exp\left\{-\frac{1}{2}\operatorname{Tr}(\Psi'\Sigma^{-1})\right\} \quad (7)$$

up to a multiplicative constant independent of μ and Σ (note that the terms involving $|\Sigma|$ are already equalized and needn't be considered any more).

By taking the log of each quantity and multiplying by -2, we see that the above problem is equivalent to finding μ_0', β', Ψ' such that

$$\sum_{i=1}^{n}(X_i - \mu)^T\Sigma^{-1}(X_i - \mu) + (\mu - \mu_0)^T(\beta\Sigma)^{-1}(\mu - \mu_0) + \operatorname{Tr}(\Psi\Sigma^{-1}) \quad (8)$$

is equal to

$$(\mu - \mu_0')^T(\beta'\Sigma)^{-1}(\mu - \mu_0') + \operatorname{Tr}(\Psi'\Sigma^{-1}) \quad (9)$$

up to an *additive* constant independent of μ and Σ.

Defining $S = \sum_{i=1}^{n} X_i$, we can rewrite (8) as

$$\begin{aligned}
&\sum_{i=1}^{n}(X_i - \mu)^T\Sigma^{-1}(X_i - \mu) + (\mu - \mu_0)^T(\beta\Sigma)^{-1}(\mu - \mu_0) + \operatorname{Tr}(\Psi\Sigma^{-1}) \\
&= \frac{1}{\beta}\mu^T\Sigma^{-1}\mu - 2\frac{1}{\beta}\mu^T\Sigma^{-1}\mu_0 + \frac{1}{\beta}\mu_0^T\Sigma^{-1}\mu_0 + \\
&+ \sum_{i=1}^{n} X_i^T\Sigma^{-1}X_i - 2\mu^T\Sigma^{-1}S + n\mu^T\Sigma^{-1}\mu + \operatorname{Tr}(\Psi\Sigma^{-1}) \\
&= \left(\frac{1}{\beta} + n\right)\mu^T\Sigma^{-1}\mu - 2\mu^T\Sigma^{-1}\left(\frac{1}{\beta}\mu_0 + S\right) + \quad (10) \\
&+ \frac{1}{\beta}\mu_0^T\Sigma^{-1}\mu_0 + \sum_{i=1}^{n} X_i^T\Sigma^{-1}X_i + \operatorname{Tr}(\Psi\Sigma^{-1}). \quad (11)
\end{aligned}$$

Notice each appearance of μ is now in the two terms on line (10), which look like the first two terms of the expansion of a quadratic form similar to the first term in (9). In order to "complete the square", we must set

$$\beta' = \frac{1}{\frac{1}{\beta} + n}$$

so that the first term on line (10) is equal to the quadratic term (in μ) from the expansion of $(\mu - \mu_0')^T(\beta'\Sigma)^{-1}(\mu - \mu_0')$. The linear term in μ on line (10) now implies that we must set

$$\mu_0' = \beta' \left(\frac{1}{\beta}\mu_0 + S \right) = \frac{\frac{1}{\beta}\mu_0 + S}{\frac{1}{\beta} + n}.$$

Continuing from line (11) we have

$$\begin{aligned}
&\left(\frac{1}{\beta} + n\right) \mu^T\Sigma^{-1}\mu - 2\mu^T\Sigma^{-1}\left(\frac{1}{\beta}\mu_0 + S\right) + \frac{1}{\beta}\mu_0^T\Sigma^{-1}\mu_0 + \sum_{i=1}^n X_i^T\Sigma^{-1}X_i + \mathrm{Tr}(\Psi\Sigma^{-1}) = \\
&= \mu^T(\beta'\Sigma)^{-1}\mu - 2\mu^T(\beta'\Sigma)^{-1}\mu_0' + \mu_0'^T(\beta'\Sigma)^{-1}\mu_0' - \mu_0'^T(\beta'\Sigma)^{-1}\mu_0' + \\
&\quad + \frac{1}{\beta}\mu_0^T\Sigma^{-1}\mu_0 + \sum_{i=1}^n X_i^T\Sigma^{-1}X_i + \mathrm{Tr}(\Psi\Sigma^{-1}) = \\
&= (\mu - \mu_0')^T(\beta'\Sigma)^{-1}(\mu - \mu_0') - \mu_0'^T(\beta'\Sigma)^{-1}\mu_0' + \frac{1}{\beta}\mu_0^T\Sigma^{-1}\mu_0 + \sum_{i=1}^n X_i^T\Sigma^{-1}X_i + \mathrm{Tr}(\Psi\Sigma^{-1})
\end{aligned}$$

and since the $(\mu - \mu_0')^T(\beta'\Sigma)^{-1}(\mu - \mu_0')$ term is completed we can drop it from the last line and from (9). I.e. we now only need find Ψ' such that

$$-\mu_0'^T(\beta'\Sigma)^{-1}\mu_0' + \frac{1}{\beta}\mu_0^T\Sigma^{-1}\mu_0 + \sum_{i=1}^n X_i^T\Sigma^{-1}X_i + \mathrm{Tr}(\Psi\Sigma^{-1}) \tag{12}$$

equals

$$\mathrm{Tr}(\Psi'\Sigma^{-1})$$

up to constants.

Continuing from (12):

$$\begin{aligned}
&-\mathrm{Tr}(\mu_0'^T(\beta'\Sigma)^{-1}\mu_0') + \mathrm{Tr}\left(\frac{1}{\beta}\mu_0^T\Sigma^{-1}\mu_0\right) + \sum_{i=1}^n \mathrm{Tr}\left(X_i^T\Sigma^{-1}X_i\right) + \mathrm{Tr}(\Psi\Sigma^{-1}) = \\
&= -\mathrm{Tr}\left(\frac{1}{\beta'}\mu_0'\mu_0'^T\Sigma^{-1}\right) + \mathrm{Tr}\left(\frac{1}{\beta}\mu_0\mu_0^T\Sigma^{-1}\right) + \sum_{i=1}^n \mathrm{Tr}\left(X_iX_i^T\Sigma^{-1}\right) + \mathrm{Tr}(\Psi\Sigma^{-1}) = \\
&= \mathrm{Tr}\left(\left[\Psi + \frac{1}{\beta}\mu_0\mu_0^T + \sum_{i=1}^n X_iX_i^T - \frac{1}{\beta'}\mu_0'\mu_0'^T\right]\Sigma^{-1}\right)
\end{aligned} \tag{13}$$

so we set

$$\Psi' = \Psi + \frac{1}{\beta}\mu_0\mu_0^T + \sum_{i=1}^n X_iX_i^T - \frac{1}{\beta'}\mu_0'\mu_0'^T.$$

After simplifying this a bit, you should be able to recognize the empirical means and covariance matrix:

$$\Psi' = \Psi + \sum_{i=1}^n \left(X_i - \frac{1}{n}S\right)\left(X_i - \frac{1}{n}S\right)^T + \frac{\frac{n}{\beta}}{\frac{1}{\beta} + n}\left(\frac{1}{n}S - \mu_0\right)\left(\frac{1}{n}S - \mu_0\right)^T.$$

2.2.3 Maximizing the posterior

In the last section we showed that the posterior distribution of μ and Σ after n observations under a Normal-inverse-Wishart distribution is again a Normal-inverse-Wishart distribution with parameters

$$\mu_0' = \frac{\frac{1}{\beta}\mu_0 + S}{\frac{1}{\beta} + n},$$

$$\beta' = \frac{1}{\frac{1}{\beta} + n},$$

$$\Psi' = \Psi + \sum_{i=1}^{n} \left(X_i - \frac{1}{n} S \right) \left(X_i - \frac{1}{n} S \right)^T + \frac{\frac{n}{\beta}}{\frac{1}{\beta} + n} \left(\frac{1}{n} S - \mu_0 \right) \left(\frac{1}{n} S - \mu_0 \right)^T,$$

and

$$\nu' = \nu + d,$$

where $S = \sum_{i=1}^{n} X_i$. Recall that the density of this distribution is

$$f(\mu, \Sigma | \mu_0', \beta', \Psi', \nu') = p(\mu | \mu_0', \beta' \Sigma) w(\Sigma | \Psi', \nu') \tag{14}$$

where p is the density of the multivariate normal distribution, and w is the density of the inverse-Wishart distribution given by

$$w(\Sigma | \Psi', \nu') = \frac{|\Psi'|^{\nu'/2}}{2^{\nu' d/2} \Gamma_d \left(\frac{\nu'}{2} \right)} |\Sigma|^{-(\nu' + d + 1)/2} \exp \left\{ -\frac{1}{2} \operatorname{Tr}(\Psi' \Sigma^{-1}) \right\}.$$

Since μ only appears in the first term on the right hand side of 14, it is obvious that the value of μ that maximizes the posterior is also the value that maximizes $p(\mu | \mu_0', \beta' \Sigma)$, which, of course, is μ_0':

$$\widehat{\mu} = \frac{\frac{1}{\beta}\mu_0 + S}{\frac{1}{\beta} + n}.$$

We can think of this quantity as the maximum likelihood estimator for the mean of the normal distribution if, along with $X_1, ..., X_n$, we had also observed $1/\beta$-many samples, all equal to μ_0.

Maximizing the posterior with respect to Σ is equivalent to minimizing

$$J(\Sigma) = (\nu' + d + 2) \log |\Sigma| + (\widehat{\mu} - \mu_0')^T (\beta' \Sigma)^{-1} (\widehat{\mu} - \mu_0') + \operatorname{Tr}(\Psi' \Sigma^{-1}) \tag{15}$$

where we have plugged in the MAP value for μ;

$$\frac{\partial J}{\partial \Sigma} = (\nu' + d + 2) \Sigma^{-1} - \frac{1}{\beta'} \Sigma^{-1} (\widehat{\mu} - \mu_0')(\widehat{\mu} - \mu_0')^T \Sigma^{-1} - \Sigma^{-1} \Psi' \Sigma^{-1} \tag{16}$$

and the maximizer is

$$\widehat{\Sigma} = \frac{1}{\beta'(\nu' + d + 2)} (\widehat{\mu} - \mu_0')(\widehat{\mu} - \mu_0')^T + \frac{1}{\nu' + d + 2} \Psi'. \tag{17}$$