Lecture Notes Summary

Complete derivation of the 0-Entropic prior for the k-dim Gaussians.

Topics Covered

The Manifold of k-dim Gaussians

Let $x, \mu \in R^k$ and $\Sigma \in R^{k\times k}$ symmetric and positive definite. We write $\Sigma > 0$ to indicate that Σ is symmetric and positive definite. We say that $X \sim \mathrm{N}(\mu, \Sigma)$, i.e. it is $k-$dim gaussian with mean vector μ and variance matrix Σ when the pdf (w.r.t. Lebesgue measure dx in R^k) is,

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

Information Separation

Let $\theta_0 = (\mu_0, \Sigma_0)$ and $\theta = (\mu, \Sigma)$. Then the Kullback number between the $P_0 = N(\mu_0, \Sigma_0)$ and $P = N(\mu, \Sigma)$ is denoted by $\mathrm{I}(\theta_0 : \theta)$. By defininiton,

$$\mathrm{I}(\theta_0 : \theta) = \mathrm{E}_0\left(\log\frac{P_0(dx)}{P(dx)}\right)$$

and for the k-gaussians we have,

$$\mathrm{I}(\theta_0 : \theta) = \frac{1}{2}\log|\Sigma_0^{-1}\Sigma| + \frac{1}{2}\mathrm{E}_0\left((x-\mu)'\Sigma^{-1}(x-\mu) - (x-\mu_0)'\Sigma_0^{-1}(x-\mu_0)\right)$$

The Trace Trick:

The trace of a scalar is just the scalar itself and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. This implies the very useful "trace trick":

$$\mathrm{E}_0\left((x-\mu_0)'\Sigma_0^{-1}(x-\mu_0)\right) = \mathrm{E}_0\ \mathrm{tr}\left((x-\mu_0)(x-\mu_0)'\Sigma_0^{-1}\right) = \mathrm{tr}\left(\Sigma_0\Sigma_0^{-1}\right) = k$$

and by adding and substracting μ_0 and using the trace trick again,

$$\mathrm{E}_0\left((x-\mu)^{'}\Sigma^{-1}(x-\mu)\right) = (\mu-\mu_0)^{'}\Sigma^{-1}(\mu-\mu_0) + \mathrm{tr}\left(\Sigma_0\Sigma^{-1}\right).$$

Thus,

$$\mathrm{I}(\theta_0:\theta) = \frac{1}{2}(\mu-\mu_0)^{'}\Sigma^{-1}(\mu-\mu_0) + \frac{1}{2}\log|\Sigma\Sigma_0^{-1}| + \frac{1}{2}\mathrm{tr}\left(\Sigma_0\Sigma^{-1}\right) - \frac{k}{2}$$

Notice that the dimension of the manifold of k-dim gaussians is the number of independent parameters in $\theta = (\mu, \Sigma)$ i.e. $k + k(k+1)/2$, since $\mu \in R^k$ and σ is symmetric. When $k = 1$ the manifold is of dimension two, and when $k = 2$ it is of dimension 5. Remember the 5 number summary for simple linear regression: The two means, the two SDs and the correlation coefficient.

The Information Metric

The fastest way to get all the entries of the Fisher information matrix $G(\theta) = (g_{ij}(\theta))$, (i.e. the information metric) is by expanding $J(t) = \mathrm{I}(\theta : \theta + tv)$ in a taylor series about $t = 0$ and using the fact that,

$$J(t) = \frac{t^2}{2}v^{'}G(\theta)v + o(t^2)$$

and thus,

$$J^{''}(0) = v^{'}G(\theta)v$$

gives all the entries g_{ij} at once by simply collecting the coefficients of the terms of the taylor series that are quadratic in t. Clearly only the first term of $\mathrm{I}(\theta_0 : \theta)$ contains μ and it is already quadratic in μ. The other terms contain Σ but not μ and therefore the information matrix must be block diagonal,

$$G(\theta) = \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & Q \end{bmatrix}$$

where $Q > 0$ is of dimension $k(k+1)/2$. Let us separate the components of the velocity vector as (v, V) as in $\theta = (\mu, \Sigma)$. We have,

$$J(t) = \frac{t^2}{2}v^{'}\Sigma^{-1}v + \frac{1}{2}\log\frac{|\Sigma + tV|}{|\Sigma|} + \frac{1}{2}\mathrm{tr}\left(\Sigma(\Sigma+tV)^{-1}\right) - \frac{k}{2}$$

that simplifies to,

$$J(t) = \frac{t^2}{2}v^{'}\Sigma^{-1}v + \frac{1}{2}\log|I + t\Sigma^{-1}V| + \frac{1}{2}\mathrm{tr}\left((I+t\Sigma^{-1}V)^{-1}\right) - \frac{k}{2}$$

We notice that for any square matrix A, we have,

$$|I + tA| = 1 + \mathrm{tr}(A)t + \mathrm{tr}_2(A)t^2 + o(t^2)$$

where tr_2 is the familiar second trace from the standard expansion of the characteristic polynomial of a square matrix. Recall that the second trace is the sum of all the 2 by 2 determinants obtained from A by discarding all but 2 rows and 2 cols of A. It is not difficult to check this formula by induction on the dimension of A. It is also straight forward (again by using induction on the size of A) to check that,

$$\mathrm{tr}_2(A) = \frac{1}{2}\left(\mathrm{tr}^2(A) - \mathrm{tr}(A^2)\right).$$

It is also clear that,

$$(I+tA)^{-1} = I - tA + t^2A^2 + o(t^2)$$

and hence,

$$\frac{1}{2}\mathrm{tr}(I+tV)^{-1} = \frac{k}{2} - \frac{t}{2}\mathrm{tr}(A) + \frac{t^2}{2}\mathrm{tr}(A^2) + o(t^2)$$

and from the previous expression for the second trace we obtain,

$$\begin{aligned}\frac{1}{2}\log|I+tV| &= \frac{1}{2}\mathrm{tr}(A)t + \frac{1}{2}(2\mathrm{tr}_2(A) - \mathrm{tr}^2(A))\frac{t^2}{2} + o(t^2)\\ &= \frac{1}{2}\mathrm{tr}(A) - \frac{1}{2}\mathrm{tr}(A^2)\frac{t^2}{2} + o(t^2).\end{aligned}$$

Collecting all the coefficients of t^2 we finally obtain,

$$J^{''}(0) = \frac{t^2}{2}\left(v^{'}\Sigma^{-1}v + \frac{1}{2}\mathrm{tr}(A^2)\right) + o(t^2)$$

where $A = V\Sigma^{-1}$ or $A = \Sigma^{-1}V$ both give the same answer. This allows to compute the entries of $Q = (q_{ij})$. Let us write the lower triangular elements of V in a vector v (do not confuse with the previous v that it won't be needed any more) of dimension $k(k+1)/2$,

$$v^{'} = (v_{11}, v_{21}, v_{22}, v_{31}, v_{32}, v_{33}, \ldots, v_{kk})$$

so that,

$$v^{'}Qv = \frac{1}{2}\mathrm{tr}(A^2) = \frac{1}{2}\mathrm{tr}(V\Sigma^{-1}V\Sigma^{-1}).$$

Now, denote $\Sigma^{-1} = (\sigma^{ij})$ and by e_j the j-th cannonical vector (1 in position j and 0 everywhere else) and by E_{ij} the matrix with all zeroes except for a 1 in the i-th row j-th column. We have,

$$\begin{aligned}q_{ij} = e_i^{'}Qe_j &= \frac{1}{2}\mathrm{tr}(V_i\Sigma^{-1}V_j\Sigma^{-1})\\ &= \frac{1}{2}\mathrm{tr}\left([E_{i_1i_2} + (1-\delta_{i_1i_2})E_{i_1i_2}]\Sigma^{-1}[E_{j_1j_2} + (1-\delta_{j_1j_2})E_{j_2j_1}]\Sigma^{-1}\right)\end{aligned}$$

where we have used the fact that when the i-th entry in v is $v_{i_1i_2}$ then the corresponding V is,

$$V_i = E_{i_1i_2} + (1-\delta_{i_1i_2})E_{i_1i_2}.$$

expanding the products we obtain,

$$\begin{aligned}q_{ij} = \frac{1}{2}[&\sigma^{i_1j_1}\sigma^{i_2j_2} + (1-\delta_{j_1j_2})\sigma^{i_1j_2}\sigma^{i_2j_1}\\ &+ (1-\delta_{i_1i_2})\sigma^{i_2j_1}\sigma^{i_1j_2}\\ &+ (1-\delta_{i_1i_2})(1-\delta_{j_1j_2})\sigma^{i_1j_1}\sigma^{i_2j_2}]\end{aligned}$$

Example 1

When $k = 2$. we have,

$$Q = \begin{bmatrix} \frac{1}{2}(\sigma^{11})^2 & \sigma^{11}\sigma^{21} & (\sigma^{12})^2/2 \\ \sigma^{11}\sigma^{21} & \sigma^{22}\sigma^{11} + (\sigma^{11})^2 & \sigma^{22}\sigma^{21} \\ (\sigma^{12})^2/2 & \sigma^{22}\sigma^{21} & \frac{1}{2}(\sigma^{11})^2 \end{bmatrix}$$

#1

In this case we can check directly that,

$$\det Q = |\Sigma^{-1}|^3/4$$

In general, it is possible to show that

$$\det Q = 2^{-k}|\Sigma^{-1}|^{k+1}$$

Proof: A tedious but straight forward strategy may be as follows: Using the explicit formulas for q_{ij}, just prove it first for the case when Σ^{-1} is an elementary matrix, i.e., one of the three types obtainable by performing one of the three elementary row operations to the identity matrix. Recall that the elementary row operations are: swapping two rows, multiplying a row by a non-zero number, and, adding a multiple of one row to another row. Then, one needs to check that $Q(EA) = Q(E)Q(A)$ and finally use the standard fact that every non-singular matrix is the product of elementary matrices.

The Entropic Priors

The 0-Entropic Prior for a regular model M is given by a family of scalar probability density functions on M. For an initial prior guess P_0 and prior number of observations $\alpha > 0$, sufficiently large so that the prior is proper, the scalar density is:

$$\pi(P|P_0, \alpha) = \frac{1}{Z}\exp(-\alpha\ \mathrm{I}(P_0 : P))$$

where α needs to be large enough so that,

$$Z = \int_M e^{-\alpha\ I(P_0:P)}\, dM\ <\ \infty$$

where the integration is with respect to the volume form in the Riemannian manifold (M, g) with g the information metric. i.e., in a given θ-parametrization, $M = \{P_\theta : \theta \in \Theta\}$,

$$dM = \sqrt{\det g(\theta)}\ d\theta$$

and, the density becomes a function of $\theta \in \Theta$,

$$\pi(\theta|\theta_0, \alpha) = \frac{1}{Z}e^{-\alpha\, I(\theta_0:\theta)}$$

Notice that the choice of $P_0 = P_{\theta_0}$ is just a simple special case.

The 1-Entropic Priors are just like the above but with $\mathrm{I}(\theta : \theta_0)$ in the exponent. It is possible to show that the 0-Entropy, $\mathrm{I}_0 = \mathrm{I}(P_0 : P)$, and the 1-Entropy, $\mathrm{I}_1 = \mathrm{I}(P : P_0)$ are two extremes of a continium of entropies I_δ for $0 < \delta < 1$ that produce δ-Entropic Priors with tails that are not exponential but polynomial that are multivariate generalizations of the student-t distributions.

The Entropic Priors continued on next page...

Entropic Priors for the Multivariate Gaussians

When the model M is the special case of the manifold of k-dimensional Gaussian distributions, the 0-Entropic prior is obtained by applying the general formulas for entropic priors to the specific formulas for the entropy, information metric, and, volume form, previously obtained for the k-Gaussians. The scalar density for the 0-Entropic prior is,

$$\pi(\mu,\Sigma|\mu_0,\Sigma_0,\alpha) = \frac{1}{Z}\cdot\exp(-\frac{1}{2}(\mu-\mu_0)^{'}(\frac{1}{\alpha}\Sigma)^{-1}(\mu-\mu_0)) \cdot|\Sigma^{-1}|^{\alpha/2}\exp(-\frac{\alpha}{2}\mathrm{tr}(\Sigma_0\Sigma^{-1}))$$

where the normalizing constant is obtained by integrating the above with respect to the Riemannian volume element,

$$dM = \frac{d\mu\wedge d\Sigma}{2^{k/2}|\Sigma|^{(k+2)/2}}$$

and after integrating over $\mu\in R^k$ we have,

$$Z = (\frac{\pi}{\alpha})^{k/2}\int_{\Sigma>0}|\Sigma^{-1}|^{\alpha/2}\exp(-\frac{\alpha}{2}\mathrm{tr}(\Sigma_0\Sigma^{-1}))\,\frac{d\Sigma}{|\Sigma|^{(k+1)/2}}$$

Let's do the substitution, $S=(\alpha/2)\Sigma_0\Sigma^{-1}$ in the integral,

$$\begin{aligned} Z &= \left|\frac{2}{\alpha}\Sigma_0^{-1}\right|^{\alpha/2}\int_{S>0}|S|^{\alpha/2}\exp(-\mathrm{tr}(S))\,\frac{dS}{|S|^{(k+1)/2}} \\ &= \left|\frac{2}{\alpha}\Sigma_0^{-1}\right|^{\alpha/2}\Gamma_k(\frac{\alpha}{2}) \end{aligned}$$

where we have introduced the k-dimensional Gamma function defined for $2a>k-1$ by,

$$\Gamma_k(a) = \int_{S>0}|S|^a\exp(-\mathrm{tr}(S))\,\frac{dS}{|S|^{(k+1)/2}}$$

where the integral, as before, is over all the k by k symmetric positive definite matrices.

Problem 1

Show that by writting $S=TT^{'}$ with $T=(t_{ij})$ a lower triangular matrix with positive diagonal (i.e., Cholesky decomposition) and using the fact (show it!) that,

$$dS = \prod_{i=1}^{k}(2\,t_{ii}^{k-i+1})\,dT$$

the k-dimensional Gamma function can be written as a product of ordinary (1-dimensional) gamma functions,

$$\Gamma_k(a) = (\sqrt{\pi})^{k(k-1)/2}\prod_{i=1}^{k}\Gamma(a-(i-1)/2)$$

#1

Hence, the 0-Entropic prior for the multivariate Gaussian model with prior parameters μ_0,Σ_0,α is integrable provided, that $\alpha>k-1$ and given by,

$$(\mu|\mu_0, \Sigma, \alpha) \sim N\left(\mu_0, \frac{1}{\alpha}\Sigma\right)$$
$$(\Sigma|\Sigma_0, \alpha) \sim \mathcal{W}_k^{-1}(\alpha, \alpha\Sigma_0)$$

notice the exact correspondance, of the form of this distribution, with the formulas for the univariate (i.e., $k = 1$) case. The generalization of the inverse Chi-square to the multivariate case is the inverse Wishart $\mathcal{W}_k^{-1}(\alpha, \alpha\Sigma_0)$ of dimension k, with α degrees of freedom and scale matrix $\alpha\Sigma_0$. The p.d.f., $p(\Sigma)$, with respect to the usual Lebesgue measure "$d\Sigma$" in dimension $k(k+1)/2$ (recall that covariance matrices are symmetric) is non-singular for $\alpha > k - 1$ and given by,

$$p(\Sigma)\, d\Sigma = \left(2^{\alpha k/2}\Gamma_k(\alpha/2)\right)^{-1} |\alpha\Sigma_0\Sigma^{-1}|^{\alpha/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\alpha\Sigma_0\Sigma^{-1})\right) \frac{d\Sigma}{|\Sigma|^{(k+1)/2}}$$

Problem 2

Show that, like in the univariate case, $S = \alpha\Sigma_0\Sigma^{-1}$ follows a Wishart with α degrees of freedom, i.e. $S \sim \mathcal{W}_k(\alpha)$ with density,

$$p(S)\, dS = \left(2^{\alpha k/2}\Gamma_k(\alpha/2)\right)^{-1} |S|^{\alpha/2} \exp\left(-\frac{1}{2}\mathrm{tr}(S)\right) \frac{dS}{|S|^{(k+1)/2}}$$

#2

The Posterior

Finally, we can check, directly applying Bayes's Theorem, or by noticing that the multivariate Gaussians are an exponential family, that the 0-Entropic family of priors is conjugate, i.e., when the prior parameters are μ_0, Σ_0, α, the posterior distribution is of the same form as the prior but with parameters, $\mu_1, \Sigma_1, (\alpha + n)$. The posterior parameters are obtained from the prior parameters and the sufficient statistics of the data. The easiest way to obtain the formulas for the posterior parameters is to think of the prior as providing α extra virtual observations. The posterior parameters are simply the sufficient statistics based on all the $(n + \alpha)$ observations. Thus, the posterior mean μ_1 is such that,

$$(n + \alpha)\, \mu_1 = \sum_{i=1}^{n+\alpha} y_i = \sum_{i=1}^{n} y_i + \sum_{j=1}^{\alpha} y_{n+j} = n\, \bar{y} + \alpha\, \mu_0$$

from where we obtain,

$$\mu_1 = \frac{1}{\alpha + n}(\alpha\, \mu_0 + n\, \bar{y})$$

and simmilarly for the posterior variance matrix, Σ_1,

$$\begin{aligned}(n + \alpha)\, \Sigma_1 &= \sum_{i=1}^{n}(y_i - \mu_1)^2 + \sum_{j=1}^{\alpha}(y_{n+j} - \mu_1)^2 \\ &= \sum_{i=1}^{n}(y_i - \mu_1)^2 + \alpha\, \Sigma_0 + \alpha\, (\mu_0 - \mu_1)^2\end{aligned}$$

where the last two terms where obtained by the standard trick of adding and subtracting the prior mean μ_0 inside the second summation. One could further simplify the last line by adding and subtracting the mean $\bar{y}$

of the n observations obtaining other equivalent expressions for Σ_1 but the one shown here is already directly computable from the data and the prior parameters. We have,

$$\Sigma_1 = \frac{1}{\alpha + n}\left[\alpha\left(\Sigma_0 + \alpha\left(\mu_0 - \mu_1\right)^2\right) + \sum_{i=1}^{n}(y_i - \mu_1)^2\right]$$

General Comments

The student should be aware of the fact that the 0-Entropic priors are not the typical priors. Eventhough they are similar to the standard recommendations for conjugate priors found in most of the current literature and in statistical textbooks, they are not exactly equal to them. The standard recipe is not invariant under reparametrizations and therefore it is geometrically meaningless. The general recipie of conjugate priors, for the exponential family of distributions, has been very popular since it was first suggested in 1961. Its popularity is due to the fact that with conjugate priors the computation of the posterior is very easy. The integration is done once (if at all) to get the updating formulas for the posterior parameters and after that all it needs to be done is to substitute the value of the observed sufficient statistics into the formulas. It is also common to encounter variations of the recipie involving extra parameters that add "flexibility" to the choice of the prior. In low dimensions (mostly when dim $M = 1$) the recipie has been extremely useful allowing fast automatic inference for many important problems. The consensus among bayesian statisticians seems to be that for multidimensional parameter spaces the inferences end up better if instead of applying the recipie for the whole vector of thetas one assumes independent conjugate priors for each of the coordinates of the vector of thetas. The geometric view, exemplified in this lecture, explains why. The usual conjugate prior recipie is given by the formula of its density with respect to the standard Lebesgue measure, disregarding the volume element in M. This wrong, geometrically meaningless prior, becomes worst and plainly undesirable as soon as the dimension of M increases. As the dimensionality of the parameter space increases, the wrong conjugate prior becomes informative about *the wrong thing* for which of course there is no other justification besides the convenience of having a close formula for the posterior. No wonder, generations of data analysts have recommended against the recipie in multidimensional parameter spaces.

The objective shape of M, which is the support of all priors on M, does supply objective prior information to the problem. In fact it is, I claim, the most important source of prior information available in the inference problem. The amount of prior information actually *increases* with the dimensionality of M. By not using the objective Riemannian geometry of M one is wasting the main source of objective prior information.