EM for the gestalt model

1 Generative model

A gestalt, a perceptual object is characterised by a covariance component for the joint distribution of visual neural activity.

$$p(v \mid g) = \mathcal{N}(v; 0, C_v) \tag{1}$$

$$C_v = \sum_{k=1}^{K} g_k C_k \tag{2}$$

where K is the fixed number of possible gestalts in the visual scene and g_k is the strength of the gestalt number k, coming from a K-dimensional symmetric Dirichlet prior distribution with concentration parameter α controlling the sparsity of the prior.

$$p(g) = \mathrm{Dir}(g; \alpha) \tag{3}$$

The pixel intensities are generated from the neural activity through a set of linear projective field models, possibly Gabor filters, A, adding some independent observational noise.

$$p(x \mid v) = \mathcal{N}(x; Av, C_x) \tag{4}$$

$$C_x = \sigma_x I; \tag{5}$$

We might assume that a single composition of gestalts, characterised by the g vector, generates a batch of B images, described by cellular activities $V = \{v_1 \dots v_B\}$ and observations $X = \{x_1 \dots x_B\}$.

2 E-step

The joint posterior over hidden variables g and v is the following

$$p(v, g \mid x) = p(x \mid v, g)\frac{p(v, g)}{p(x)} = p(x \mid v)p(v \mid g)p(g)\frac{1}{p(x)} \tag{6}$$

For the purpose of sampling, we can discard the normalisation factor $p(x)$ and the normalisation constants of the Gaussians. The logarithm of the Dirichlet prior over g looks like the following.

$$\log p(g) = \log(\Gamma(\alpha K)) - \log(\Gamma(\alpha)^K) + (\alpha - 1)\sum_{k=1}^{K} \log(g_k) \tag{7}$$

We can discard the terms not depending on g. So taking the logarithm of the unnormalised posterior, the sampling target will look like this

$$\log p(v, g \mid x) \sim$$
$$-\frac{1}{2}\left[(x - Av)^T C_x^{-1}(x - Av) + \log(\det(C_v)) + v^T C_v^{-1} v\right] + (\alpha - 1)\sum_{k=1}^{K} \log(g_k) \tag{8}$$

where $g \in (0, 1]$ and $\sum_{k=1}^{K} g_k = 1$, and $-\infty$ everywhere else.

2.1 Hamiltonian Monte Carlo Sampling

The negative log-posterior can be regarded as an energy function for a dynamical system updating invariantly to the posterior distribution. To use Hamiltonian MC sampling, the gradient of this energy has to be constructed as follows

$$E(g, v) =$$
$$\frac{1}{2}\left[(x - Av)^T C_x^{-1}(x - Av) + \log(\det(C_v)) + v^T C_v^{-1} v\right] + (\alpha - 1)\sum_{k=1}^{K} \log(g_k) \tag{9}$$

$$\frac{\partial E(g, v)}{\partial g_j} =$$
$$\frac{1}{2}\sum_{a=1}^{D_v}\sum_{b=1}^{D_v}\left[\left(\sum_{k=1}^{K} g_k C_k\right)^{-1}\left[I - vv^T\left(\sum_{k=1}^{K} g_k C_k\right)^{-1}\right]\right]_{a,b}[C_j]_{a,b} + \frac{(\alpha - 1)}{g_j} \tag{10}$$

$$\frac{\partial E(g, v)}{\partial v} = A^T C_x^{-1}(Av - x) + \left(\sum_{k=1}^{K} g_k C_k\right)^{-1} v \tag{11}$$

denoting the dimension of v as D_v and using differentiating rules 12, 13 and 14

$$\frac{\partial}{\partial M}\log(\det(M)) = M^{-T} \tag{12}$$

$$\frac{\partial}{\partial M} a^T M^{-1} a = -M^{-T} a a^T M^{-T} \tag{13}$$

$$f : \mathbb{R}^{D\times D} \to \mathbb{R} \quad h : \mathbb{R} \to \mathbb{R}^{D\times D}$$
$$\frac{\partial}{\partial x} f(h(x)) = \sum_{a=1}^{D}\sum_{b=1}^{D}\left[\frac{\partial}{\partial h(x)} f(h(x))\right]_{a,b}\left[\frac{\partial}{\partial x} h(x)\right]_{a,b} \tag{14}$$

However, given that the prior and thus the posterior over g does not have relevant gradient information outside of the unit interval, it is a better idea to use a combined sampler that proposes from Hamiltonian dynamics only for the dimensions of v.

2.2 Gibbs sampling

A more efficient way to collect samples from the joint posterior over all hidden variables is to employ a Gibbs sampling scheme, where we sample from the conditional posteriors. The first is over v, and can be defined as follows

$$p(v \mid x, g) = \frac{p(x \mid v, g)p(v \mid g)}{p(x \mid g)} = \frac{\mathcal{N}(x; Av, \sigma_x I)\mathcal{N}(v; 0, C_v)}{\int_{-\infty}^{\infty} \mathcal{N}(x; Av, \sigma_x I)\mathcal{N}(v; 0, C_v)\mathrm{d}v} \tag{15}$$

The Gaussian over x can be rewritten to a Gaussian over v times a constant c_1 in the following way

$$\mathcal{N}(x; Av, \sigma_x I) = c_1 \mathcal{N}(v; -2(A^T A)^{-1} A^T x, \sigma_x (A^T A)^{-1}) \tag{16}$$

Consequently, the product of two Gaussians in the numerator of Eq. 15 can also be written as a Gaussian over v introducing a new constant

$$\mathcal{N}(x; Av, \sigma_x I)\mathcal{N}(v; 0, C_v) = c_1 c_2 \mathcal{N}(v; \mu_{post}, C_{post}) \tag{17}$$

The denominator of Eq. 15 is the integral of this formula, which evaluates to $c_1 c_2$, as the Gaussian integrates to one. This cancels the constants in the numerator, making the conditional posterior equal to the combined Gaussian over v, which, after expanding μ_{post} and C_{post}, is

$$p(v \mid x, g) = \mathcal{N}\left(v; -\frac{2}{\sigma_x}\left(\frac{1}{\sigma_x}A^T A + C_v^{-1}\right)^{-1} A^T x, \left(\frac{1}{\sigma_x}A^T A + C_v^{-1}\right)^{-1}\right) \tag{18}$$

which can be sampled directly. The conditional posterior over g is defined as follows

$$p(g \mid x, v) = \frac{p(x \mid g, v)p(g \mid v)}{p(x \mid v)} = \frac{p(v \mid g)p(g)}{p(v)} \tag{19}$$

which can be sampled by a Metropolis-Hastings scheme with the following target

$$\log p(g \mid x, v) \sim -\frac{1}{2}\left[\log(\det(C_v)) + v^T C_v^{-1} v\right] + (\alpha - 1)\sum_{k=1}^{K} \log(g_k) \tag{20}$$

3 M-step

The complete-data likelihood with respect to a set of observations, $X = \{x_1 \dots x_N\}$ is the following

$$p(v, g, X \mid C_{1..K}) = \prod_{n=1}^{N} p(x_n \mid v)p(v \mid g)p(g) \tag{21}$$

Its logarithm ($\mathcal{L} = \log p(v, g, x \mid C_{1..K})$) is similar to 8. We can approximate the integral of this logarithm over the joint posterior by averaging over L samples from it, separately for each observation x_n. As we will seek the values of the covariance components $C_{1...K}$ that maximise this integral, we can discard each term not depending on these parameters. This way we arrive to the following expression (substituting for C_v according to 2)

$$\mathcal{L} \sim \sum_{n=1}^{N} \frac{1}{L} \sum_{l=1}^{L} -\frac{1}{2} \left[\log \left(\det \left(\sum_{k=1}^{K} g_k^{l,n} C_k \right) \right) + v^{(l,n)T} \left(\sum_{k=1}^{K} g_k^{l,n} C_k \right)^{-1} v^{l,n} \right] \tag{22}$$

The double summation over L samples over all N observations always happens on the same terms, so we can substitute it with a single sum that iterates over the full sample set. So taking the derivative with respect to one of the covariance components $j \in [1, k]$, we get

$$\frac{\partial \mathcal{L}}{\partial C_j} = -\frac{1}{2L} \sum_{m=1}^{LN} g_j^m \left[\left(\sum_{k=1}^{K} g_k^m C_k \right)^{-1} - \left(\sum_{k=1}^{K} g_k^m C_k \right)^{-1} v^m v^{mT} \left(\sum_{k=1}^{K} g_k^m C_k \right)^{-1} \right] \tag{23}$$

using $\frac{\partial}{\partial C_j} \sum_{k=1}^{K} g_k^m C_k = g_j^m$ and the differentiating rules 12 and 13 (all C_j matrices are symmetric and regular, so transposes are identical and determinants are nonzero).

Setting 23 to zero and multiplying it from the left by $\sum_{k=1}^{K} g_k^m C_k$ gives

$$0 = -\frac{1}{2L} \sum_{m=1}^{LN} g_j^m \left[I - v^m v^{mT} \left(\sum_{k=1}^{K} g_k^m C_k \right)^{-1} \right] \tag{24}$$

multiplying by $-2L$ and by $\sum_{k=1}^{K} g_k^m C_k$, but this time from the right, gives

$$0 = \sum_{m=1}^{LN} g_j^m \left[\sum_{k=1}^{K} g_k^m C_k - v^m v^{mT} \right] \tag{25}$$

rearranging the sums yields

$$0 = \sum_{k=1}^{K} C_k \sum_{m=1}^{LN} g_j^m g_k^m - \sum_{m=1}^{LN} g_j^m v^m v^{mT} \tag{26}$$

From this we can express C_j to get

$$C_j^{new} = \frac{1}{\sum_{m=1}^{LN} g_j^{m2}} \left[\sum_{m=1}^{LN} g_j^m v^m v^{mT} - \sum_{k \neq j}^{K} C_k \sum_{m=1}^{LN} g_j^m g_k^m \right] \tag{27}$$

that may be rearranged to

$$C_j^{new} = \frac{1}{\sum_{m=1}^{LN} g_j^{m2}} \sum_{m=1}^{LN} g_j^m \left[v^m v^{mT} - \sum_{k \neq j}^{K} g_k^m C_k \right] \tag{28}$$