Vanilla Re-implementation of Gaussian Process Factor Analysis (GPFA)

In this notebook, we detail the mathematics of a commonly used Bayesian inference technique called GPFA [1 (https://journals.physiology.org/doi/full/10.1152/jn.90941.2008)].

One way to study the brain is to directly measure the neuronal spiking activities. Spikes or also known as action potentials are binary sigals that neurons use to communicate with one another. Using a few electrodes, we can record hundreds or thousands of neurons spiking activities over seconds or minutes when the subject completes a task (for example, a decision making task or arm reach movements). If we put these data into a matrix of Y, of size number of neurons q by time T (T time bins, each 2ms long for example), we see that we would end up with a large matrix with very fine time resolution. Neuroscientists usually attempt to understand this by smoothing the data and by dimensionality reduction techniques such as PCA.

Instead of seperating data smoothing and data dimensionality reduction into 2 steps, GPFA proposes a hierarchical model where the state variable is smooth - modeled by Gaussian process - and is related to observed noisy spikes through a linear regression. Every parameter is learned with Expectation Maximization method.

1. Variables and the model

- Spiking data (**the observations**).

  $Y \in \mathbb{R}^{q \times T}$, where q is the number of neurons and T is the number of time bins. $y_{:,t}$ is the t-th column of Y.
- Dimension reduced neuron states or neural trajectory (**The nuissance parameter of the model**).

  $Z \in \mathbb{R}^{p \times T}$, where p is the number of dimensions of and T is the number of time bins. We use Z here instead of X as in the source paper because with Z the notation of EM will then match Bishop and other Bayesian textbooks. $z_{:,t}$ is the t-th column of Z.
- At each time t, observations and neural states are related by linear-Gaussian:

  $$y_{:,t}|z_{:,t} \sim \mathrm{N}(Cz_{:,t} + d, R)$$

  where $C \in \mathbb{R}^{q \times p}, d \in \mathbb{R}^{q \times 1}, R \in \mathbb{R}^{q \times q}$. GPFA further constrains the covariance R to be diagonal, modeling the independent noise level for each neuron. In essence, the model, like factor analysis, tries to represent the independent variance associated with each coordinate in the matrix R and capturing the covariance between neurons in the matrix C.
- For each row of neural states, $z_{i,:}$ across time are related through Gaussian process (**how how the model structure shares information between observations**). $z_{i,:} \sim (\mathbf{0}, K_i)$ where $K_i(t_1, t_2) = \sigma_{f,i}^2 \exp(-\frac{(t_1-t_2)^2}{2\tau_i^2}) + \delta_{t_1,t_2}\sigma_{n,i}$, where the Kronecker delta is 1 if $t_1 = t_2$ and 0 otherwise. We set $\sigma_{n,i}$ to be 10^{-3}. The $\sigma_{n,i}$ helps stablizes matrix K.

2. Model priors and parameter estimation

The prior for $x_{:,t}$ is $\mathrm{N}(\mathbf{0}, I)$. In other words, each neuronal state is uncorrelated with each other. This necessitates that we set $K_i(t,t) = 1$, which is $\sigma_{f,i}^2 = 1 - \sigma_{n,i}^2$
**We note that across time neuronal states are correlated through the Gaussian process kernel. This is how how the model structure shares information between spike count observations over time. **

Parameters are estimated via EM algorithm. In our case, we observe $Y = y$, $Y \sim P_{Y|\Theta}$, $\Theta \sim P_\Theta$, where $\Theta = [X, Z], X = [C, d, R, \tau]$. We want MAP estimator for X marginalized over the nuisance parameter neural trajectory Z.

Let x^t be the current estimate of parameters.

E-step: calculate $Q(x|x^{(t)}) = \mathbb{E}_{z\sim p(z|x^{(t)},y)}[\log p(x|z,y)]$

In this step, the main workload is in calculating the conditional probability of neural trajectory given current estimate of parameters and observed data. Due to the use of Gaussian process and linear-normal distribution,

$z_{i,:} \sim \mathrm{N}(\mathbf{0}, K_i)$

$y_{:,t}|z_{:,t} \sim \mathrm{N}(Cz_{:,t} + d, R)$

Concatenating all columns of Z into one long column and doing the same for Y, we see that $\bar{Z}$ and $\bar{Y}$ are jointly Gaussian. Details can be found in Equation A1-A4 in the source paper.

Using the basic result of conditioning for jointly Gaussian random variables, we have

$\bar{Z}|\bar{Y} \sim \mathrm{N}(\bar{K}\bar{C}'(\bar{C}\bar{K}\bar{C}' + \bar{R})^{-1}(\bar{y} - \bar{d}), \bar{K} - \bar{K}\bar{C}'(\bar{C}\bar{K}\bar{C}' + \bar{R})^{-1}\bar{C}\bar{K})$

where we implicitly condition on the current estimate of parameters.

We then use the joint distribution to calculate Q.

$Q(x|x^{(t)}) \propto \mathbb{E}_{\bar{Z}\sim p(\bar{Z}|x^{(t)},\bar{Y})}[\log p(X, \bar{Y}, \bar{Z})]$

$$= \begin{bmatrix} \mathbb{E}(\bar{Z}) \\ \bar{y} - \bar{d} \end{bmatrix}^T M^{-1} \begin{bmatrix} \mathbb{E}(\bar{Z}) \\ \bar{y} - \bar{d} \end{bmatrix} - \tfrac{1}{2}\log(\det(K)) - \tfrac{1}{2}\log(\det(R)) = \square$$

where $M = \begin{bmatrix} \bar{K} & \bar{K}\bar{C}' \\ \bar{C}\bar{K} & \bar{C}\bar{K}\bar{C}' + \bar{R} \end{bmatrix}$.

To simplify, we first use the following identity to expand out M inverse.

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}$$

$$M^{-1} = \begin{bmatrix} \bar{K}^{-1} + \bar{C}'\bar{R}^{-1}\bar{C} & -\bar{C}'\bar{R}^{-1} \\ -\bar{R}^{-1}\bar{C}' & \bar{R}^{-1} \end{bmatrix}$$

Therefore, $Q(x|x^{(t)}) = \mathbb{E}(\bar{Z})^T\bar{K}^{-1}\mathbb{E}(\bar{Z}) + \mathbb{E}(\bar{Z})^T\bar{C}'\bar{R}^{-1}\bar{C}\mathbb{E}(\bar{Z}) - 2\mathbb{E}(\bar{Z})^T\bar{C}'\bar{R}^{-1}\bar{Y} + 2\mathbb{E}(\bar{Z})^T\bar{C}'\bar{R}^{-1}\bar{d} + \bar{Y}^T\bar{R}^{-1}\bar{Y} - 2\bar{Y}^T\bar{R}^{-1}\bar{d} + \bar{d}^T\bar{R}^{-1}\bar{d} - \tfrac{1}{2}\log(\det(K)) - \tfrac{1}{2}\log(\det(R))$

M-step:

Maximizing Q with respect to C and d: ¶

Restricting to the terms in Q with C and d,

$$Q(x|x^{(t)}) = \mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{C} \mathbb{E}(\bar{Z}) - 2\mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{Y} + 2\mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{d} - 2\bar{Y}^T \bar{R}^{-1} \bar{d} + \bar{d}^T \bar{R}^{-1} \bar{d}$$

$$\begin{bmatrix} \mathbb{E}\bar{Z}^T & 1 \end{bmatrix} \begin{bmatrix} \bar{c}^T \\ \bar{d}^T \end{bmatrix} \bar{R}^{-1} \begin{bmatrix} \bar{c} & \bar{d} \end{bmatrix} \begin{bmatrix} \mathbb{E}\bar{Z} \\ 1 \end{bmatrix} - \begin{bmatrix} \mathbb{E}\bar{Z}^T & 1 \end{bmatrix} \begin{bmatrix} \bar{c}^T \\ \bar{d}^T \end{bmatrix} \begin{bmatrix} \bar{R}^{-1}\bar{Y} & \bar{R}^{-1}\bar{Y} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Let $[\bar{C}\bar{d}] = A$, Taking the derivative of the first term with respect to A:

$$\frac{d\mathrm{tr}(\begin{bmatrix} \mathbb{E}\bar{Z}^T & 1 \end{bmatrix} A^T R^{-1} A \begin{bmatrix} \mathbb{E}\bar{Z} \\ 1 \end{bmatrix})}{dA} = \frac{d\mathrm{tr}(A^T R^{-1} A \begin{bmatrix} \mathbb{E}\bar{Z}\mathbb{E}\bar{Z}^T & \mathbb{E}\bar{Z} \\ \mathbb{E}\bar{Z}^T & 1 \end{bmatrix})}{dA}$$

Taking the derivative of the second term with respect to A, we get $\frac{d\mathrm{tr}(A^T \begin{bmatrix} \bar{R}^{-1}\bar{Y} & \bar{R}^{-1}\bar{Y} \end{bmatrix} \begin{bmatrix} \mathbb{E}\bar{Z}^T & 1 \\ \mathbb{E}\bar{Z}^T & 1 \end{bmatrix})}{dA}$

Using the trace property (https://web.stanford.edu/~jduchi/projects/matrix_prop.pdf):

$$\nabla_A \mathrm{tr} ABA^T C = CAB + C^T AB^T$$

$$\frac{dQ}{dA^T} = 2 \begin{bmatrix} \mathbb{E}\bar{Z}\mathbb{E}\bar{Z}^T & \mathbb{E}\bar{Z} \\ \mathbb{E}\bar{Z}^T & 1 \end{bmatrix} A^T R^{-1} - 2 \begin{bmatrix} \mathbb{E}\bar{Z}^T & 1 \end{bmatrix} R^{-1} Y$$

Setting the derivative to zero yields: $A \begin{bmatrix} \mathbb{E}\bar{Z}\mathbb{E}\bar{Z}^T & \mathbb{E}\bar{Z} \\ \mathbb{E}\bar{Z}^T & 1 \end{bmatrix} = \bar{Y} \begin{bmatrix} \mathbb{E}\bar{Z}^T & 1 \end{bmatrix}$

$$A = \bar{Y} \begin{bmatrix} \mathbb{E}\bar{Z}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbb{E}\bar{Z}\mathbb{E}\bar{Z}^T & \mathbb{E}\bar{Z} \\ \mathbb{E}\bar{Z}^T & 1 \end{bmatrix}^{-1}$$

which can then be derived in nonconcatenated form as in Equation A8 in the source paper.

Maximizing Q with respect to R:

$$Q(x|x^{(t)}) = -\tfrac{1}{2}(\mathbb{E}(\bar{Z})^T \bar{K}^{-1} \mathbb{E}(\bar{Z}) + \mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{C} \mathbb{E}(\bar{Z}) - 2\mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{Y} + 2\mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{d}$$
$$+ \bar{Y}^T \bar{R}^{-1} \bar{Y} - 2\bar{Y}^T \bar{R}^{-1} \bar{d} + \bar{d}^T \bar{R}^{-1} \bar{d}) - \tfrac{1}{2}\log(\det(K)) - \tfrac{1}{2}\log(\det(R)) = \square$$

Collecting the terms in Q with $\bar{R}$,

$$Q(x|x^{(t)}) = -\tfrac{1}{2}(\mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{C} \mathbb{E}(\bar{Z}) - 2\mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{Y} + 2\mathbb{E}(\bar{Z})^T \bar{C}' \bar{R}^{-1} \bar{d} + \bar{Y}^T \bar{R}^{-1} \bar{Y} - 2\bar{Y}^T \bar{R}^{-1} \bar{d}$$
$$+ \bar{d}^T \bar{R}^{-1} \bar{d}) - \tfrac{1}{2}\log(\det(R))$$

Adding trace() operator on each term and moving subterms, we get

$$\frac{d\square}{dR^{-1}} = -\tfrac{1}{2}(\bar{C}\mathbb{E}(\bar{Z})\mathbb{E}(\bar{Z})^T \bar{C}' + (\bar{Y} - \bar{d})(\bar{Y} - \bar{d})^T - 2(\bar{Y} - \bar{d})\mathbb{E}(\bar{Z})^T \bar{C}^T) + \tfrac{1}{2}\bar{R}$$

where we use the fact that $\frac{\log(det(X))}{X} = (X^{-1})^T$

Because $\bar{C}\bar{Z} + \bar{d} = \mathbb{E}(\bar{Y}|\bar{Z})$, with another expectation over Z, we have:

$\bar{C}\mathbb{E}(\bar{Z}) = \mathbb{E}(\bar{Y}) - \bar{d} = \bar{Y} - \bar{d}$

Therefore, $\frac{d\square}{dR^{-1}} = -\frac{1}{2}((\bar{Y} - \bar{d})(\bar{Y} - \bar{d})^T - (\bar{Y} - \bar{d})\mathbb{E}(\bar{Z})^T\bar{C}^T) + \frac{1}{2}\bar{R}$

This gives the update in Equation (A9) for R.

Maximizing Q with respect to $\tau's$:

Due to the interleaving of τ inside and outside of exponential operator, there is no close form solution. But the gradient is computatable and can be used with any gradient optimization technique.

Collecting the terms in Q with $\bar{K}$, $Q(x|x^{(t)}) = -\text{tr}(\mathbb{E}(\bar{Z})^T\bar{K}^{-1}\mathbb{E}(\bar{Z})) - \frac{1}{2}\log(\det(K))$

By chain rule,

$\frac{dQ}{d\tau_i} = \text{tr}([\frac{dQ}{dK_i}]'\frac{dK_i}{d\tau_i})$

where $\frac{dQ}{dK} = \frac{1}{2}(-K^{-1} + K^{-1}\mathbb{E}(\bar{Z})\mathbb{E}(\bar{Z})^TK^{-1})$

(We use the fact of 2.5 in Matrix cookbook (http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/proof002.html#dYinv_dx_p) $\partial(\text{tr}(X^{-1}) = \text{tr}(\partial(X^{-1})) = \text{tr}(-X^{-1}(\partial X)X^{-1}) = -\text{tr}(X^{-1}(\partial X)X^{-1})$**)**

Restricting to each K_i, we have

$\frac{dQ}{dK_i} = \frac{1}{2}(-K_i^{-1} + K_i^{-1}\mathbb{E}(Z_{i,:})^T\mathbb{E}(Z_{i,:})K_i^{-1})$ **(Note that subindices of Z and direction changes accordingly.)**

$\frac{dK_i(t_1,t_2)}{d\tau_i} = \sigma_{f,i}^2\frac{(t_1-t_2)^2}{\tau_i^3}\exp(-\frac{(t_1-t_2)^2}{2\tau_i^2})$

Calculating the expectations $\mathbb{E}(Z_{i,:}), \mathbb{E}(Z_{i,:})(Z_{i,:})^T$

Because these expectations show up in many M-step updates, we simplify these terms from the forms in Equation A5, reproduced above and copied again below.

$\bar{Z}|\bar{Y} \sim \text{N}(\bar{K}\bar{C}'(\bar{C}\bar{K}\bar{C}' + \bar{R})^{-1}(\bar{y} - \bar{d}), \bar{K} - \bar{K}\bar{C}'(\bar{C}\bar{K}\bar{C}' + \bar{R})^{-1}\bar{C}\bar{K})$

We reduce computational complexity of $(CKC' + R)^{-1}$ via Woodbury matrix identity. (Eqn 157 in Matrix Cookbook (https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)).

$(CKC' + R)^{-1} = R^{-1} - R^{-1}C(K^{-1} + C'R^{-1}C)^{-1}CR^{-1}$

We can go from an O(#(dimension of observation)^3) operation to an O(#(dimension of state)^3) operation. More implementational details are in the Appendix.

Denote $(\bar{K}^{-1} + \bar{C}'\bar{R}^{-1}\bar{C})^{-1}$ **by** M^{-1}

$$\mathbb{E}(Z_{i,:}) = \bar{K}\bar{C}'(\bar{C}\bar{K}\bar{C}' + \bar{R})^{-1}(\bar{y} - \bar{d}) = \bar{K}(I - \bar{C}'\bar{R}^{-1}\bar{C}M^{-1})\bar{C}\bar{R}^{-1}(\bar{y} - \bar{d})$$

For the covariance term:

We note that applying the Woodbury matrix identity to $M^{-1} = (\bar{K}^{-1} + \bar{C}'\bar{R}^{-1}\bar{C})^{-1}$ **again yields**
$M^{-1} = (\bar{K}^{-1} + \bar{C}'\bar{R}^{-1}\bar{C})^{-1} = \bar{K} - \bar{K}\bar{C}'(\bar{R} + \bar{C}\bar{K}\bar{C}')\bar{C}'\bar{K}$

So the covariance is: $\bar{K} - \bar{K}\bar{C}'(\bar{C}\bar{K}\bar{C}' + \bar{R})^{-1}\bar{C}\bar{K} = M^{-1}$

$$\mathbb{E}(Z_{i,:})(Z_{i,:})^T = M^{-1} + \mathbb{E}(Z_{i,:})\mathbb{E}(Z_{i,:})^T$$

where we use the fact that for a random variable $Y, Cov(Y) = \mathbb{E}YY^T - \mathbb{E}Y\mathbb{E}Y^T$

In [ ]: