1 Development of the point-process filtering and SS-GLM

1.1 The CIF (rate function)

We examine the spike trains that a neuron elicits in all the trials in the experiment and built a raster plot. In a steady state, we could simply average in each time bin and get a PSTH. When learning occurs this is impossible. Instead we define a rate function (single trial "PSTH") for trial #k and time bin #l of sizeΔ:

$$\lambda_k\left(l|\theta_k,\gamma,H_k\right)=\exp\left\{\sum_{r=1}^{R}\theta_{k,r}g_{,r}\left(l\right)\right\}\exp\left\{\sum_{j=1}^{l}\gamma_j n_{k,l-j}\right\} \tag{1.1}$$

Where:

θ_k is a vector of the "PSTH" in trial #k

$g_r\left(l\right)=\begin{cases}1 & (r-1)\cdot\frac{T}{R}<l\leq r\cdot\frac{T}{R}\\ 0 & otherwise\end{cases}$, T being the number of bins in each trial (each bin of duration Δ typically 1mSec) and R being the number of PSTH bins.

γ is a vector of the self - history dependence

H_k is the history (of spiking) and in our case it is the binary vector n_k.

1.2 The log-likelihood function

The probability of observing n_k spikes in trial #k and time bin #l is simply $p\left(N_k\left(l\right)=n_k\left(l\right)\right)=\left(\lambda_k\left(l\right)\cdot\Delta\right)^{n_k(l)}\cdot\left(1-\lambda_k\left(l\right)\cdot\Delta\right)^{1-n_k(l)}\approx\frac{\left(\lambda_k(l)\Delta\right)^{n_k(l)}\cdot\exp\left(-\lambda_k(l)\Delta\right)}{n_k(l)!}$. The Poisson approximation is valid for $n_k=0,1$ and $\lambda\Delta\ll1$. Thus, the log-likelihood of a single time bin is:

$$\log p\left(N_k\left(l\right)=n_k\left(l\right)\right)=-\lambda_k\left(l\right)\Delta+n_k\left(l\right)\cdot\log\left(\lambda_k\left(l\right)\Delta\right)$$

We assume that the trial by trial evolution of the PSTH parameters follow a gaussian distribution. Namely,

$$p\left(\theta_{k+1}|\theta_k\right)\sim\mathbb{N}\left(0,\Sigma\right) \tag{1.2}$$

At this point the mean is zero and we still didn't take stimulus features into account. A non-zero mean will be added in section 7 as a result of a fitted learning algorithm. Stimulus features are added in section 6.

If we assign the symbol θ_0 to the initial value of θ we can get the log-likelihood function of the observed spikes ($\left\{N_k\right\}_{k=1}^{K}$) and the hidden process ($\left\{\theta_k\right\}_{k=1}^{K}$) as a sum over all time bins in all trials:

$$L=\log p\left(\left\{N_k\right\}_{k=1}^{K},\left\{\theta_k\right\}_{k=1}^{K}|\psi\right)=\log\sum_{k=1}^{K}p\left(\mathbf{n_k}|\theta_k,H_k\right)\cdot p\left(\theta_k|\theta_{k-1}\right)=$$

$$=\sum_{k=1}^{K}\sum_{l=1}^{T}\left[-\lambda_k\left(l\right)\Delta+n_k\left(l\right)\cdot\log\left(\lambda_k\left(l\right)\Delta\right)\right]+K\cdot\log\left(\left(2\pi\right)^{-\frac{R}{2}}\cdot\left|\Sigma\right|^{-\frac{1}{2}}\right)-\frac{1}{2}\left(\theta_k-\theta_{k-1}\right)^T\Sigma^{-1}\left(\theta_k-\theta_{k-1}\right) \tag{1.3}$$

Our log-likelihood function is $F\left(\psi\right)=\log\int d^K\theta e^L$.

The parameters of the likelihood function are $\psi=\left(\gamma,\theta_0,\Sigma\right)$ and maximizing the likelihood can be done in the gradient ascent / simulated annealing methods.

In our optimizations we will assume Σ to be diagonal (block diagonal after introducing the features) and next we introduce the EM algorithm for likelihood maximization.

1.3 The EM algorithm (introduction)

1.3.1 Formulation

We're looking for the set of parameters, ψ^* that maximizes $F(\psi) = \log \int p\left(\{N_k\}_{k=1}^K, \{\theta_k\}_{k=1}^K | \psi\right) d^K\theta$. Assume the existance of an auxillary function $Q(\psi, \psi')$ that satisfies the following:

- $Q(\psi, \psi') \leq F(\psi) \forall \psi'$
- $F(\psi) = Q(\psi, \psi)$

This means that we can define an iterative process that maximizes the log-likelihood (locally). At step 'i' we define $\psi^{(i+1)} = \arg\max_\psi Q\left(\psi, \psi^{(i)}\right)$. This choice means that

$$F\left(\psi^{(i+1)}\right) = Q\left(\psi^{(i+1)}, \psi^{(i+1)}\right) \geq Q\left(\psi^{(i+1)}, \psi^{(i)}\right) \geq Q\left(\psi^{(i)}, \psi^{(i)}\right) = F\left(\psi^{(i)}\right)$$

This form suggest a two stage iterative algorithm. It is named EM-algorithm (the 'expectation' nature of the first step will become clear later in this section) and at step 'i' it goes as:

- E-Step: Calculate $Q\left(\psi, \psi^{(i)}\right)$.
- M-Step: find $\psi^{(i+1)}$ by maximizing $\psi^{(i+1)} = \arg\max_\psi Q\left(\psi, \psi^{(i)}\right)$.

Using the notation of N, θ to describe the observable and latent variables we find Q in the following way:

$$F(\psi) \overset{1}{=} F(\psi') + \log \frac{p(N|\psi)}{p(N|\psi')} \overset{2}{=} F(\psi') + \log \int p(\theta|N, \psi') \frac{p(\theta|N, \psi)}{p(\theta|N, \psi')} \cdot \frac{p(N|\psi)}{p(N|\psi')} d^K\theta$$

$$\overset{3}{=} F(\psi') + \log \int p(\theta|N, \psi') \frac{p(N, \theta|\psi)}{p(N, \theta|\psi')} d^K\theta \overset{4}{\geq} F(\psi') + \int p(\theta|N, \psi') \log \frac{p(N, \theta|\psi)}{p(N, \theta|\psi')} d^K\theta$$

$$\equiv Q(\psi, \psi')$$

Where:

1. From the definition of $F(\psi) = \log p\left(\{N_k\}_{k=1}^K | \psi\right)$
2. Because $\int p(\theta|N, \psi') \frac{p(\theta|N,\psi)}{p(\theta|N,\psi')} d^K\theta = 1$
3. From Bayes' law $p(\theta|N, \psi) \cdot p(N|\psi) = p(N, \theta|\psi)$
4. From Jensen's inequality: $\log \int \geq \int \log$ in concave functions (log of the mean vs. mean of log)

Note that $Q(\psi, \psi') = F(\psi') - \frac{1}{p(N|\psi')} D_{KL}\left[p(N, \theta|\psi') || p(N, \theta|\psi)\right]$ which means that the requirements above are met. In applying the E-step it is enough to reduce the auxillary function to

$$Q(\psi, \psi') = \int p(\theta|N, \psi') \log p(N, \theta|\psi) d^K\theta = E_{p(\theta|N,\psi')} L \tag{1.4}$$

Because the M-step requires optimization over ψ and not ψ'. Now this is in the form of an expected value ...Hence the 'E-Step'.

1.3.2 Remarks

Note two important points:

1. We never actually compute the log-likelihood $F(\psi)$.
2. We get the divergence 'for free' because: $\nabla_\psi \log p(N|\psi)|_{\psi=\psi'} = \nabla_\psi \log \int p(N,\theta|\psi)\, d^K\theta|_{\psi=\psi'} = \frac{1}{p(N|\psi')} \cdot \nabla_\psi \int p(N,\theta|\psi)\, d^K\theta|_{\psi=\psi'} =$ $\int \frac{p(N,\theta|\psi')}{p(N|\psi')} \cdot \nabla_\psi \log p(N,\theta|\psi)\, d^K\theta|_{\psi=\psi'} = \nabla_\psi Q(\psi,\psi')|_{\psi=\psi'}$

1.4 E-Step - i'th iteration

Here we need to compute the expectation of L (eqn 1.4) over the distribution $p(\theta|N,\psi')$. This calculation boils down to computing the following constituents:

1. $\theta_{k|K} \equiv \int p(\theta|N,\psi') \cdot \theta_k d^K\theta$
2. $W_{k,k+1|K} \equiv \int p(\theta|N,\psi') \cdot \left(\theta_k - \theta_{k|K}\right) \cdot \left(\theta_{k+1} - \theta_{k+1|K}\right) d^K\theta$ which is the covariance of θ_k, θ_{k+1}
3. $\int p(\theta|N,\psi') \cdot \theta_k^2 d^K\theta$
4. $\int p(\theta|N,\psi') \cdot \exp(\theta_k)\, d^K\theta$

The whole point here is going to be that we assume the distribution of θ_0 to be Gaussian (or a δ function) and that all posterior distributions ($p(\theta_k|...)$) are also Gaussian. This will allow developing a filtering algorithm.

This is done by several algorithms:

1.4.1 Forward filter algorithm (Kalman++ Eden et al 2004)

Define the following mean values and covariance matrices:

1. $\theta_{k|k-1} = E\left[\theta_k|N_{1:k-1},\psi^{(i)}\right]$
2. $\theta_{k|k} = E\left[\theta_k|N_{1:k},\psi^{(i)}\right]$ includes the spiking activity in the k'th trial
3. $W_{k|k-1} = Var\left[\theta_k|N_{1:k-1},\psi^{(i)}\right]$
4. $W_{k|k} = Var\left[\theta_k|N_{1:k},\psi^{(i)}\right]$

This algorithm starts from the initial values $\theta_{1|0} = \theta_0$ and $W_{0|0} = 0$ and iterates forward the following steps:

One step prediction: From the identity $p\left(\theta_k|N_{1:k-1},\psi^{(i)}\right) = \int p\left(\theta_k|\theta_{k-1},\psi^{(i)}\right) p\left(\theta_{k-1}|N_{1:k-1},\psi^{(i)}\right) d\theta_{k-1}$ we assume all densities to be Gaussian; $p\left(\theta_k|N_{1:k-1},\psi^{(i)}\right) \sim N\left(\theta_{k|k-1},W_{k|k-1}\right)$, $p\left(\theta_{k-1}|N_{1:k-1},\psi^{(i)}\right) \sim N\left(\theta_{k-1|k-1},W_{k-1|k-1}\right)$ and $p\left(\theta_k|\theta_{k-1},\psi^{(i)}\right) \sim N\left(\theta_{k-1},\Sigma\right)$. The convolution of two gaussians is also a gaussian (Appendix) so we immediately get the relation:

$$\theta_{k|k-1} = \theta_{k-1|k-1}, W_{k|k-1} = W_{k-1|k-1} + \Sigma \qquad (1.5)$$

Next, we incorporate the observed spikes in the k'th trial.

The posterior distribution: We use Bayes law to incorporate the observed spikes in the k'th trial: $p\left(\theta_k|N_{1:k},\psi^{(i)}\right) \overset{1}{=} \frac{p\left(N_{1:k}|\theta_k,\psi^{(i)}\right)\cdot p\left(\theta_k|\psi^{(i)}\right)}{p(N_{1:k})} \overset{2}{=} \frac{p\left(N_k|\theta_k,\psi^{(i)}\right)\cdot p\left(N_{1:k-1}|\theta_k,\psi^{(i)}\right)\cdot p\left(\theta_k|\psi^{(i)}\right)}{p(N_k|N_{1:k-1})\cdot p(N_{1:k-1})} \overset{3}{=} \frac{p\left(N_k|\theta_k,\psi^{(i)}\right)\cdot p(N_{1:k-1})\cdot p\left(\theta_k|N_{1:k-1},\psi^{(i)}\right)}{p(N_k|N_{1:k-1})\cdot p(N_{1:k-1})} \overset{4}{=} \frac{p\left(N_k|\theta_k,\psi^{(i)}\right)\cdot p\left(\theta_k|N_{1:k-1},\psi^{(i)}\right)}{p\left(N_k|N_{1:k-l},\psi^{(i)}\right)} \sim N\left(\theta_{k|k},W_{k|k}\right)$. Where (1) and (3) follow Bayes' rule, (2) stems from N_k and $N_{1:k-1}$ being independent given θ_k and (4) is simple algebra.

We are going to take the log of both sides and differentiate with respect to θ_k so we can ignore the denominator.

We use $p\left(N_k|\theta_k,\psi^{(i)}\right) = \prod_l \exp\left[-\lambda_k\left(l\right)\Delta + n_k\left(l\right)\cdot\log\left(\lambda_k\left(l\right)\Delta\right)\right]$ and $p\left(\theta_k|N_{1:k-1},\psi^{(i)}\right) \sim N\left(\theta_{k|k-1},W_{k|k-1}\right)$ to get

$$-\frac{1}{2}\left(\theta_k-\theta_{k|k}\right)^T W_{k|k}^{-1}\left(\theta_k-\theta_{k|k}\right) = \sum_{l=1}^{T}\left[-\lambda_k\left(l\right)\Delta + n_k\left(l\right)\cdot\log\left(\lambda_k\left(l\right)\Delta\right)\right] - \frac{1}{2}\left(\theta_k-\theta_{k|k-1}\right)^T W_{k|k-1}^{-1}\left(\theta_k-\theta_{k|k-1}\right) + constants \tag{1.6}$$

We now derivate with respect to θ_k twice. First to get the linear term (the mean) and second to get the variance. (remember that W and W^{-1} are symmetric)

$$\begin{aligned}\frac{\partial}{\partial\theta_{k,r}} \to \left[W_{k|k}^{-1}\cdot\left(\theta_k-\theta_{k|k}\right)\right]_r &= \left[W_{k|k-1}^{-1}\cdot\left(\theta_k-\theta_{k|k-1}\right)\right]_r - \sum_{l=1}^{T}\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta_{k,r}}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right] \\ &= \left[W_{k|k-1}^{-1}\cdot\left(\theta_k-\theta_{k|k-1}\right)\right]_r - \sum_{l=(r-1)\frac{T}{R}}^{r\frac{T}{R}}\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right] \end{aligned} \tag{1.7}$$

This should hold for $\theta_k = \theta_{k|k-1}$ so we insert it and get

$$-W_{k|k}^{-1}\cdot\left(\theta_{k|k-1}-\theta_{k|k}\right) = \sum_{l=1}^{T}\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right]|_{\theta=\theta_{k|k-1}} \tag{1.8}$$

which we solve and get:

$$\theta_{k|k} = \theta_{k|k-1} + W_{k|k}\cdot\sum_{l=1}^{T}\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta_k}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right] = \theta_{k|k-1} + W_{k|k}\cdot\begin{pmatrix}\sum_{l=1}^{\frac{T}{R}}\left(n_k\left(l\right)-\lambda_k\left(l\right)\cdot\Delta\right)\\ \cdot\\ \cdot\\ \sum_{l=T\cdot\frac{R-1}{R}+1}^{T}\left(n_k\left(l\right)-\lambda_k\left(l\right)\cdot\Delta\right)\end{pmatrix}|_{\theta=\theta_{k|k-1}} \tag{1.9}$$

We now differentiate again:

$$\frac{\partial}{\partial\theta_k}\left\{W_{k|k}^{-1}\cdot\left(\theta_k-\theta_{k|k}\right) = W_{k|k-1}^{-1}\cdot\left(\theta_k-\theta_{k|k-1}\right) - \sum_{l=1}^{T}\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta_k}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right]\right\}$$

$$W_{k|k}^{-1} = W_{k|k-1}^{-1} - \sum_{l=1}^{T}\left\{\frac{\partial^2\log\left(\lambda_k\left(l\right)\right)}{\partial\theta^2}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right] - \Delta\lambda_k\left(l\right)\cdot\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta}\cdot\left(\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta}\right)^T\right\}|_{\theta=\theta_{k|k-1}}$$

$$\left[W_{k|k}^{-1}\right]_{ab} = \left[W_{k|k-1}^{-1}\right]_{ab} + \Delta\cdot\delta_{ab}\sum_{l=(a-1)\frac{T}{R}+1}^{a\frac{T}{R}}\lambda_k\left(l\right)\cdot|_{\theta=\theta_{k|k-1}} \tag{1.10}$$

This is the first order (in Δ) correction to the Kalman filter for non-Gaussian observation.

So, we starts sequentially (in 'k'), from equation # followed by equations #,# and calculate $\theta_{k|k}, W_{k|k}$.

Note: There is a point here that is not clear. In deriving the last equation (#) we insert $\theta = \theta_{k|k-1}$ but this is arbitrary. The correction to $W_{k|k}$ is θ dependent which seems wrong to me.

(It will also be the same result if we develop λ_k around $\theta_{k|k-1}$ ... so it's fine)

1.4.2 Smoothing algorithm (Kalman smoother)

This is a reverse sequence a.k.a. the Kalman smoother (Jazwinski, page 217, equation 7.86, see also in Shumway and Stoffer 1982)

This is developed here for $\theta_{k+1} - \theta_k \sim N(0, \Sigma)$ and assumes that we have $\theta_{k|k}, W_{k|k} \forall k$ (from previous stages). This algorithm will have to be revisited when introducing learning algorithms in section 7.

We start by assuming that $\theta_{k|l}, \theta_{k+1|l}$ $(l > k)$ both maximize the Gaussian joint distribution $p\left(\theta_k, \theta_{k+1}|N_{1:l}, \psi^{(i)}\right)$. Now,

$$p\left(\theta_k, \theta_{k+1}|N_{1:l}, \psi^{(i)}\right) = \frac{p\left(\theta_k, \theta_{k+1}, N_{1:l}|\psi^{(i)}\right)}{p\left(N_{1:l}\right)} = \frac{p\left(N_{1:k}\right)}{p\left(N_{1:l}\right)} \cdot p\left(\theta_k, \theta_{k+1}, N_{\mathbf{k+1}:l}|N_{1:k}, \psi^{(i)}\right)$$

$$= \frac{p\left(N_{1:k}\right)}{p\left(N_{1:l}\right)} \cdot p\left(N_{k+1:l}|\theta_k, \theta_{k+1}, N_{1:k}, \psi^{(i)}\right) \cdot p\left(\theta_k, \theta_{k+1}|N_{1:k}, \psi^{(i)}\right) \tag{1.11}$$

Since the process is assumed Markov (including the observations) we use: $p\left(N_{k+1:l}|\theta_k, \theta_{k+1}, N_{1:k}, \psi^{(i)}\right) = p\left(N_{k+1:l}|\theta_{k+1}, \psi^{(i)}\right)$ and $p\left(\theta_k, \theta_{k+1}|N_{1:k}, \psi^{(i)}\right) = p\left(\theta_{k+1}|\theta_k, N_{1:k}, \psi^{(i)}\right) \cdot p\left(\theta_k|N_{1:k}, \psi^{(i)}\right) = p\left(\theta_{k+1}|\theta_k, \psi^{(i)}\right) \cdot p\left(\theta_k|N_{1:k}, \psi^{(i)}\right)$.

Combining the terms above with equation # we get:

$$p\left(\theta_k, \theta_{k+1}|N_{1:l}, \psi^{(i)}\right) = c\left(\theta_{k+1}\right) \cdot p\left(\theta_{k+1}|\theta_k, \psi^{(i)}\right) \cdot p\left(\theta_k|N_{1:k}, \psi^{(i)}\right) \tag{1.12}$$

Where $c\left(\theta_{k+1}\right)$ is independent of θ_k. We now insert $p\left(\theta_{k+1}|\theta_k, \psi^{(i)}\right) \sim N\left(\theta_k, \Sigma\right)$ and $p\left(\theta_k|N_{1:k}, \psi^{(i)}\right) \sim N\left(\theta_{k|k}, W_{k|k}\right)$ and require that $\theta_{k|l}, \theta_{k+1|l}$ minimize:

$$d\left(\theta_{k+1}\right) + \left(\theta_{k+1} - \theta_k\right)^T \cdot \Sigma^{-1} \cdot \left(\theta_{k+1} - \theta_k\right) + \left(\theta_k - \theta_{k|k}\right)^T \cdot W_{k|k}^{-1} \cdot \left(\theta_k - \theta_{k|k}\right) \tag{1.13}$$

Where $d\left(\theta_{k+1}\right)$ is independent of θ_k. If we assume that $\theta_{k+1|l}$ is known we only need to minimize

$$\left(\theta_{k+1|l} - \theta_k\right)^T \cdot \Sigma^{-1} \cdot \left(\theta_{k+1|l} - \theta_k\right) + \left(\theta_k - \theta_{k|k}\right)^T \cdot W_{k|k}^{-1} \cdot \left(\theta_k - \theta_{k|k}\right) \tag{1.14}$$

With repect to θ_k. We take the derivative and set it to zero (remember that the matrices are symmetric):

$$0 = \left(\Sigma^{-1} + W_{k|k}^{-1}\right) \cdot \theta_{k|l} - \Sigma^{-1} \cdot \theta_{k+1|l} - W_{k|k}^{-1} \cdot \theta_{k|k}$$

which solves to

$$\theta_{k|l} = \left(\Sigma^{-1} + W_{k|k}^{-1}\right)^{-1} \cdot \left(\Sigma^{-1} \cdot \theta_{k+1|l} + W_{k|k}^{-1} \cdot \theta_{k|k}\right) \tag{1.15}$$

Using the identities in appendix 8.3 this resolves into:

$$\theta_{k|l} = \theta_{k|k} + A_k \cdot \left(\theta_{k+1|l} - \theta_{k+1|k}\right) \tag{1.16}$$

$$A_k = W_{k|k} W_{k+1|k}^{-1} \tag{1.17}$$

Proof: Set the matrices in appendix 8.3 to be $M = I, R = \Sigma, P = W_{k|k}$ and get

$$\theta_{k|l} = \left(\Sigma^{-1} + W_{k|k}^{-1}\right)^{-1} \Sigma^{-1} \cdot \theta_{k+1|l} + \left(\Sigma^{-1} + W_{k|k}^{-1}\right)^{-1} W_{k|k}^{-1} \cdot \theta_{k|k}$$

$$= W_{k|k} \cdot \left(W_{k|k} + \Sigma\right)^{-1} \theta_{k+1|l} + \left(I - W_{k|k} \cdot \left(W_{k|k} + \Sigma\right)^{-1}\right) \theta_{k|k}$$

and all that is left is to identify $\theta_{k|k} = \theta_{k+1|k}$ and $W_{k|k} + \Sigma = W_{k+1|k}$
Next we develop the smoothing of the covariance matrix. We define $\theta_{k|q}^{\sim} = \theta_k - \theta_{k|q}$ and subtract equation # from θ_k to get:

$$\theta_{k|l}^{\sim} + A_k \theta_{k+1|l} = \theta_{k|k}^{\sim} + A_k \theta_{k|k}$$

We square both sides (multiply by transpose) and compute the expextation to get

$$W_{k|l} + A_k \underset{=0}{\left\langle \theta_{k+1|l} \theta_{k|l}^{\sim T} \right\rangle} + \underset{=0}{\left\langle \theta_{k|l}^{\sim} \theta_{k+1|l}^{T} \right\rangle} A_k^T + A_k \left\langle \theta_{k+1|l} \theta_{k+1|l}^{T} \right\rangle A_k^T = W_{k|k} + A_k \underset{=0}{\left\langle \theta_{k|k} \theta_{k|k}^{\sim T} \right\rangle} + \underset{=0}{\left\langle \theta_{k|k}^{\sim} \theta_{k|k}^{T} \right\rangle} A_k^T + A_k \left\langle \theta_{k|k} \theta_{k|k}^{T} \right\rangle A_k^T$$

and so, by the identity $\left\langle \theta_k \theta_k^T \right\rangle = W_{k|q} + \left\langle \theta_{k|q} \theta_{k|q}^T \right\rangle$ we get

$$W_{k|l} = W_{k|k} + A_k \cdot \left(W_{k+1|l} - W_{k+1|k}\right) \cdot A_k^T \tag{1.18}$$

Backwards (reverse) calculation:
We start with $\theta_{K|K}$ and $W_{K|K}$ that resulted from the forward filter (section 4.1) and iterate equations #,# from K...1 to obtain $\theta_{k|K}$ and $W_{k|K}$ $\forall k$. From these quantities we easily get:

$$E\left(\theta_k^2 | N, \psi^{(i)}\right) = \int p\left(\theta | N, \psi^{(i)}\right) \cdot \theta_k^2 d^K \theta = W_{k|K} + \theta_{k|K}^T \theta_{k|K} \tag{1.19}$$

1.4.3 state space covariance algorithm (de Jong and Mackinnon 1988)

Here we compute $W_{k,u|K} \equiv \int p\left(\theta | N, \psi'\right) \cdot \left(\theta_k - \theta_{k|K}\right) \cdot \left(\theta_u - \theta_{u|K}\right) d^K \theta$. We assume that $1 \leq k \leq u \leq K$ and use $\theta_{k|l} = \theta_{k|k} + A_k \cdot \left(\theta_{k+1|l} - \theta_{k+1|k}\right)$ (equation # above).

Lemma: The orthogonal projection of θ_k on the subspace $N_1, ..., N_k, \theta_{k+1} - \theta_{k+1|k}, \epsilon_{s+1} ... \epsilon_K$ (with $\epsilon_{k+1} = \theta_{k+2} - \theta_{k+1}$ being the realization of the random additions in the brownian motion) is:

$$\hat{\theta_k} = \theta_{k|k} + A_k \cdot \left(\theta_{k+1} - \theta_{k+1|k}\right)$$

Proof: The inner product is $\langle X, Y \rangle = E\left(XY\right)$.
Both $\theta_{k+1} - \theta_{k+1|k}$ and $\epsilon_{s+1} ... \epsilon_K$ have mean zero and are uncorrelated to each other and to $N_1, ..., N_k$. So the projection breaks (see appendix) to

$$\hat{\theta_k} = \theta_{k|k} + Cov\left(\theta_k, \theta_{k+1} - \theta_{k+1|k}\right) \cdot W_{k+1|k}^{-1} \cdot \left(\theta_{k+1} - \theta_{k+1|k}\right) \tag{1.20}$$

and

$$Cov\left(\theta_k, \theta_{k+1} - \theta_{k+1|k}\right) \cdot W_{k+1|k}^{-1} = Cov\left(\theta_k, \theta_k\right) \cdot W_{k+1|k}^{-1} = A_k \tag{1.21}$$

Hence,

$$W_{k,u|K} \equiv Cov\left[\theta_k - \theta_{k|K}, \theta_u - \theta_{u|K}\right] = Cov\left[\theta_k - \theta_{k|K}, \theta_u\right] = Cov\left[\theta_k - \hat{\theta_k} + \hat{\theta_k} - \theta_{k|K}, \theta_u\right] \overset{1}{=} Cov\left[\hat{\theta_k} - \theta_{k|K}, \theta_u\right] \tag{1.22}$$

$$= Cov\left[A_k \cdot \left(\theta_{k+1} - \theta_{k+1|K}\right), \theta_u\right] = A_k W_{k+1,u|K} \tag{1.23}$$

Where the equality '1' stems from the orthogonality of the error in different steps. Namely, $\theta_k - \hat{\theta_k}$ is a projection of θ_k on a space that is uncorrelated to θ_u. (See appendix)

1.4.4 Calculating $\langle \exp(\theta) \rangle$

The last quantity we need to calculate in the E-step is $\int p(\theta|N, \psi') \cdot \exp(\theta_{k,r})\, d^K\theta$. Here we expand $\exp\theta_{k,r}$ around $\theta_{k|K}$ in a taylor series to the second order and get:

$$\exp(\theta_{k,r}) \approx \exp\left(\theta_{k|K,r}\right) + \left(\left(\theta_k - \theta_{k|K}\right) \cdot \nabla_\theta\right) \exp(\theta) \,|_{\theta=\left(0,0,\ldots,\theta_{k|K,r},0,..\right)} + \frac{1}{2}\left(\theta_k - \theta_{k|K}\right)\left[\left(\theta_k - \theta_{k|K}\right)\nabla_\theta\left(\nabla_\theta \exp\theta\right)\right] |_{\theta=\left(0,0,\ldots,\theta_{k|K,r},0,..\right)} \tag{1.24}$$

and than take the expectation according to $\theta_k \sim N\left(\theta_{k|K}, W_{k|K}\right)$ and get:

$$E\left[\exp\theta_{k,r}|N, \psi^{(i)}\right] \approx \exp\theta_{k|K,r} + \frac{1}{2}\left[W_{k|K}\right]_{rr}\exp\theta_{k|K,r} \tag{1.25}$$

1.5 M-Step - i'th iteration

The goal here is to find $\psi^{(i+1)} = \arg\max_\psi Q\left(\psi, \psi^{(i)}\right)$. Remember,

$$Q\left(\psi, \psi^{(i)}\right) = E_{p\left(\theta|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\sum_{l=1}^{T}\left(-\lambda_k(l)\Delta + n_k(l)\cdot\log\left(\lambda_k(l)\Delta\right)\right)\right]$$

$$-\frac{1}{2}KR\log 2\pi - \frac{1}{2}K\log|\Sigma| - \frac{1}{2}E_{p\left(\theta|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\left(\theta_k - \theta_{k-1}\right)^T\Sigma^{-1}\left(\theta_k - \theta_{k-1}\right)\right] \tag{1.26}$$

With (reminder)

$$\lambda_k\left(l|\theta_k, \gamma, H_k\right) = \exp\left\{\sum_{r=1}^{R}\theta_{k,r}g_{,r}(l)\right\}\exp\left\{\sum_{j=1}^{l}\gamma_j n_{k,l-j}\right\} \tag{1.27}$$

1.5.1 Updating γ (Newton-Raphson)

Newton raphson iteration for finding the extermum of $f(x)$ by $x_{n+1} = x_n - \frac{f'(x_n)}{f^{(2)}(x_n)}$.

$Q\left(\psi, \psi^{(i)}\right)$'s dependence on γ separates quite nicely.

$$\frac{\partial Q}{\partial \gamma_j} = \sum_{k,l}\left(-\lambda_k(l)\cdot\Delta\cdot n_k(l-j) + n_k(l)\cdot n_k(l-j)\right)$$

$$\frac{\partial^2 Q}{\partial\gamma_s\partial\gamma_j} = -\Delta\sum_{k,l}\lambda_k\left(l\right)\cdot n_k\left(l-j\right)\cdot n_k\left(l-s\right)$$

These derivatives are used to follow an iterative algorithm whos 'm' iteration goes:

$$\gamma_{m+1}^{(i)} = \gamma_m^{(i)} - \left[\nabla_\gamma^2 Q\right]^{-1}\cdot\nabla_\gamma Q|_{\gamma=\gamma_m^{(i)}} \tag{1.28}$$

and the stopping criterion is $\left|\gamma_{m+1}^{(i)} - \gamma_m^{(i)}\right| < 10^{-2}$

1.5.2 Updating Σ

$Q\left(\psi,\psi^{(i)}\right)$'s dependence on Σ separates completely. Furthermore, there is a closed form solution to:

$$\Sigma^{(i+1)} = \arg\max_{\Sigma} S\left(\Sigma\right) = \arg\max_{\Sigma}\left\{-\frac{1}{2}K\log\left|\Sigma\right| - \frac{1}{2}E_{p\left(\theta|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\left(\theta_k-\theta_{k-1}\right)^T\Sigma^{-1}\left(\theta_k-\theta_{k-1}\right)\right]\right\} \tag{1.29}$$

In the original setting Σ is a diagonal matrix. So, in this case we can assume $\Sigma = \begin{pmatrix} d_1 & & 0 \\ & \cdot & \\ & & \cdot \\ 0 & & d_R \end{pmatrix}$ and solve $\frac{\partial S}{\partial d_j} = 0$:

$$0 = \frac{K}{d_j} - E_{p\left(\theta|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\left(\theta_k-\theta_{k-1}\right)^T\begin{pmatrix} 0 & & 0 \\ & \cdot & \\ & & -\frac{1}{d_j^2} & \\ 0 & & 0 \end{pmatrix}\left(\theta_k-\theta_{k-1}\right)\right] = \frac{K}{d_j} - \frac{1}{d_j^2}E_{p\left(\theta|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\left(\theta_k-\theta_{k-1}\right)_j^T\left(\theta_k-\theta_{k-1}\right)_j\right]$$

which is solved to:

$$d_j = \frac{1}{K}E_{p\left(\theta|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\left(\theta_k-\theta_{k-1}\right)_j^T\left(\theta_k-\theta_{k-1}\right)_j\right] \tag{1.30}$$

In the general case we use the fact that Σ is symmetric and invertible so there exists a diagonal matrix $D = \begin{pmatrix} d_1 & & 0 \\ & \cdot & \\ & & \cdot \\ 0 & & d_R \end{pmatrix}$ and an orthogonal matrix O such that $\Sigma = ODO^{-1}$. Introducing this into equation # we get

$S\left(\Sigma\right) = -\frac{1}{2}K\log\left|\Sigma\right| - \frac{1}{2}E_{p\left(\theta|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\left(\theta_k-\theta_{k-1}\right)^T\Sigma^{-1}\left(\theta_k-\theta_{k-1}\right)\right]$

$= -\frac{1}{2}K\log\left|D\right| - \frac{1}{2}\int d^R\theta p\left(\theta|N,\psi^{(i)}\right)\left[\sum_{k=1}^{K}\left(\theta_k-\theta_{k-1}\right)^T O^T D^{-1} O\left(\theta_k-\theta_{k-1}\right)\right]$

We make the change $y_k = O\theta_k$ which has a Jacobian with determinant 1 and get

$$S\left(D\right) = -\frac{1}{2}K\log\left|D\right| - \frac{1}{2}E_{p\left(O^T y|N,\psi^{(i)}\right)}\left[\sum_{k=1}^{K}\left(y_k-y_{k-1}\right)^T D^{-1}\left(y_k-y_{k-1}\right)\right] \tag{1.31}$$

which solves as above. All we need to do now is transform back to Σ via:

$$\Sigma^{(i+1)} = \frac{1}{K} E_{p\left(\theta|N,\psi^{(i)}\right)} \left[\sum_{k=1}^{K} O \begin{pmatrix} \left(O^T\theta_k^T - O^T\theta_{k-1}^T\right)_1 \left(O\theta_k - O\theta_{k-1}\right)_1 & & 0 \\ & \cdot & \\ & & \cdot \\ 0 & & d_R \end{pmatrix} O^{-1} \right] \tag{1.32}$$

$$= \frac{1}{K} E_{p\left(\theta|N,\psi^{(i)}\right)} \sum_k \left(\theta_k - \theta_{k-1}\right)^T \left(\theta_k - \theta_{k-1}\right) \tag{1.33}$$

1.5.3 Updating θ_0

This too has a closed form solution. The part of $Q\left(\psi, \psi^{(i)}\right)$ pertaining θ_0 is $-\frac{1}{2} E_{p\left(\theta|N,\psi^{(i)}\right)} \left[\left(\theta_1 - \theta_0\right)^T \Sigma^{-1} \left(\theta_1 - \theta_0\right) \right] = -\sum_{ab} \frac{1}{2} E_{p\left(\theta|N,\psi^{(i)}\right)} \left[\left(\theta_1 - \theta_0\right)_a \Sigma_{ab}^{-1} \left(\theta_1 - \theta_0\right)_b \right]$. We derivate by θ_{0q} and equate to zero (taking into consideration that Σ^{-1} is symmetric:

$$-E_{p\left(\theta|N,\psi^{(i)}\right)} \left[\Sigma^{-1} \left(\theta_0 - \theta_1\right) \right]_q = 0 \tag{1.34}$$

But, since Σ is invertible we obtain:

$$\theta_0^{(i+1)} = E_{p\left(\theta|N,\psi^{(i)}\right)} \left[\theta_1\right] \tag{1.35}$$

This concludes the M-step (and the original SSGLM). Next we include the stimulus features.

1.6 Introducing stimulus features

Here we add another dimension to the CIF, the stimulus features. Thus, the change is:

$$\lambda_k \left(l|\theta_k, \gamma, H_k\right) = \exp\left\{ \sum_{r=1}^{R} \sum_{\alpha=0}^{F} \theta_{k,\alpha,r} g_{,r}\left(l\right) f\left(x_k\right)_\alpha \right\} \exp\left\{ \sum_{j=1}^{l} \gamma_j n_{k,l-j} \right\} \tag{1.36}$$

Where, x_k is the stimulus at thr k'th trial, $f\left(x_\alpha\right) \in \{-1, 1\}^{F+1}$ is it's features and f_0 is always '1' to account for the feature independent cases. This change is equivalent to a dynamic PSTH for each feature.

Next we will see the changes in the algorithms this change introduces.

1.6.1 log-likelihood function

Here there's no change. We continue to assume

$$p\left(\theta_{k+1}|\theta_k\right) \sim \mathbb{N}\left(0, \Sigma\right) \tag{1.37}$$

But now Σ is symmetric and block diagonal because we construct $\theta_k = \left(\theta_{k,r=1,\alpha=0}, \theta_{k,r=1,\alpha=1}, ..., \theta_{k,r=1,\alpha=F}, \theta_{k,r=2,\alpha=0}, ...\right)$. For ease of use we identify $\theta_{k,r,\alpha} = \theta_k\left(\left(r-1\right)F + \alpha + 1\right)$The log-likelihood of observed spikes and hidden process (L) doesn't change.

1.6.2 E-step

Forward filter algorithm (Kalman++ Eden et al 2004) The development of the forward filter starts the same. The differences start when taking derivatives of λ. Here we introduce the changes.

$$\frac{\partial}{\partial\theta_{k,r,\alpha}} \rightarrow \left[W_{k|k}^{-1}\cdot\left(\theta_k-\theta_{k|k}\right)\right]_{r,\alpha} = \left[W_{k|k-1}^{-1}\cdot\left(\theta_k-\theta_{k|k-1}\right)\right]_{r,\alpha} - \sum_{l=1}^{T}\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta_{k,r}}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right]$$

$$= \left[W_{k|k-1}^{-1}\cdot\left(\theta_k-\theta_{k|k-1}\right)\right]_{r,\alpha} - \sum_{l=(r-1)\frac{T}{R}}^{r\frac{T}{R}}\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right]\cdot f\left(x_k\right)_\alpha \tag{1.38}$$

This should hold for $\theta_k=\theta_{k|k-1}$ so we insert it and get

$$\theta_{k|k}=\theta_{k|k-1}+W_{k|k}\cdot\sum_{l=1}^{T}\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta_k}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right]=\theta_{k|k-1}+W_{k|k}\cdot\begin{pmatrix}\sum_{l=1}^{\frac{T}{R}}\left(n_k\left(l\right)-\lambda_k\left(l\right)\cdot\Delta\right)\cdot\overrightarrow{f}\left(x_k\right)\\ \cdot\\ \cdot\\ \sum_{l=T\cdot\frac{R-1}{R}+1}^{T}\left(n_k\left(l\right)-\lambda_k\left(l\right)\cdot\Delta\right)\cdot\overrightarrow{f}\left(x_k\right)\end{pmatrix}|_{\theta=\theta_{k|k-1}} \tag{1.39}$$

Another derivative yields:

$$W_{k|k}^{-1}=W_{k|k-1}^{-1}-\sum_{l=1}^{T}\left\{\frac{\partial^2\log\left(\lambda_k\left(l\right)\right)}{\partial\theta^2}\cdot\left[n_k\left(l\right)-\lambda_k\left(l\right)\Delta\right]-\Delta\lambda_k\left(l\right)\cdot\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta}\cdot\left(\frac{\partial\log\left(\lambda_k\left(l\right)\right)}{\partial\theta}\right)^T\right\}|_{\theta=\theta_{k|k-1}}$$

$$\left[W_{k|k}^{-1}\right]_{(a,\alpha),(b,\beta)}=\left[W_{k|k-1}^{-1}\right]_{(a,\alpha),(b,\beta)}+\Delta\cdot\delta_{ab}\sum_{l=(a-1)\frac{T}{R}+1}^{a\frac{T}{R}}\lambda_k\left(l\right)\cdot f_\alpha\left(x_k\right)\cdot f_\beta\left(x_k\right)|_{\theta=\theta_{k|k-1}} \tag{1.40}$$

Smoothing algorithm (Kalman smoother) Here there are no changes.

state space covariance algorithm (de Jong and Mackinnon 1988) Here there are no changes.

1.6.3 Calculating $\langle\exp\left(\theta\right)\rangle$

The last quantity we need to calculate in the E-step is $\int p\left(\theta|N,\psi'\right)\cdot\exp\left(\theta_{k,r}\right)d^K\theta$. Here we expand $\exp\theta_{k,r}$ around $\theta_{k|K}$ in a taylor series to the second order and get:

$$\exp\left(\sum_{r,\alpha}\theta_{k,r,\alpha}g_{,r}\left(l\right)f\left(x_k\right)_\alpha\right)\approx\exp\left(\sum_{r,\alpha}\theta_{k|K,r,\alpha}g_{,r}\left(l\right)f\left(x_k\right)_\alpha\right)$$

$$+\left(\left(\theta_k-\theta_{k|K}\right)\cdot\nabla_\theta\right)\exp\left(\sum\theta gf\right)|_{\theta=\theta_{k|K}}+\frac{1}{2}\left(\theta_k-\theta_{k|K}\right)\left[\left(\theta_k-\theta_{k|K}\right)\nabla_\theta\left(\nabla_\theta\exp\sum\theta gf\right)\right]|_{\theta=\theta_{k|K}} \tag{1.41}$$

and than take the expectation according to $\theta_k\sim N\left(\theta_{k|K},W_{k|K}\right)$ and get:

$$E\left[\exp\left(\sum_{r,\alpha}\theta_{k,r,\alpha}g_{,r}\left(l\right)f\left(x_k\right)_\alpha\right)|N,\psi^{(i)}\right]\approx\exp\left(\sum_{\alpha}\theta_{k|K,r,\alpha}f\left(x_k\right)_\alpha\right)\cdot\left[1+\frac{1}{2}\sum_{\alpha,\beta}\left[W_{k|K}\right]_{(r,\alpha),(r,\beta)}\cdot f\left(x_k\right)_\alpha\cdot f\left(x_k\right)_\beta\right] \quad (1.42)$$

where r is such that $g_r\left(l\right)=1$

1.6.4 M-step

Updating γ (Newton-Raphson) No change.

Updating Σ No change.

Updating θ_0 No change.

1.7 Introducing learning algorithms

Learning algorithms can be introduced via the drift term in the stochastic waights update. Our approach will be to iterate the optimization of the stochastic processes with the updating of deterministic algorithmic properties (e.g. learning rate) with the goal of minimizing the random components (e.g. Σs)

1.8 Appendices

1.8.1 Normal distribution

$$p\left(\mathbf{x}\right)=\left(2\pi\right)^{-\frac{d}{2}}\left|\mathbf{P}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\mu\right)^T\mathbf{P}^{-1}\left(\mathbf{x}-\mu\right)\right] \quad (1.43)$$

Where the mean is μ and the covariance matrix is $P_{ij}=cov\left(x_i,x_j\right)$

1.8.2 The convolution of gaussians

We want to compute $\int p\left(\theta_k|\theta_{k-1},\psi^{(i)}\right)p\left(\theta_{k-1}|N_{1:k-1},\psi^{(i)}\right)d\theta_{k-1}\sim\int\exp\left(-\frac{1}{2}\frac{\left(\theta_k-\theta_{k-1}\right)^2}{2\Sigma}\right)\otimes\exp\left(-\frac{1}{2}\frac{\left(\theta_{k-1}-\theta_{k-1|k-1}\right)^2}{2W_{k-1|k-1}}\right)d\theta_{k-1}=\int\exp\left(-\frac{1}{2}\frac{\left(u+\theta_{k-1|k-1}-\theta_k\right)^2}{2\Sigma}\right)\otimes\exp\left(-\frac{1}{2}\frac{u^2}{2W_{k-1|k-1}}\right)du$

A simple fourier analysis (namely, the fact that the charachteristic function of the normal distribution $N\left(\mu,\Sigma\right)$ is $\exp\left(i\mu^Tt-\frac{1}{2}t^T\Sigma t\right)$) shows that the result is also Gaussian with mean $\theta_{k-1|k-1}$ and variance $W_{k-1|k-1}+\Sigma$

1.8.3 Matrix equalities

Let P,R,M be square matrices (reversible)

Eq1: $\left(I+PM^TR^{-1}M\right)^{-1}=I-PM^T\left(MPM^T+R\right)^{-1}M$

We simply check: $\left(I+PM^TR^{-1}M\right)\cdot\left(I-PM^T\left(MPM^T+R\right)^{-1}M\right)=I+PM^TR^{-1}M-PM^T\left(MPM^T+R\right)^{-1}M-PM^TR^{-1}MPM^T\left(MPM^T+R\right)^{-1}M=I+PM^TR^{-1}\left(MPM^T+R-R-MPM^T\right)\left(MPM^T+R\right)^{-1}M=I$

Eq2: Multiply by P from the right and get:

$\left(I+PM^TR^{-1}M\right)^{-1}P=P-PM^T\left(MPM^T+R\right)^{-1}MP$

Eq3: Multiply by M^TR^{-1} from the right:

$\left(I+PM^TR^{-1}M\right)^{-1}PM^TR^{-1}=PM^TR^{-1}-PM^T\left(MPM^T+R\right)^{-1}MPM^TR^{-1}=PM^T\left(I-\left(MPM^T+R\right)^{-1}MPM^T\right)R^{-1}=PM^T\left(MPM^T+R\right)^{-1}\left(MPM^T+R-MPM^T\right)R^{-1}=PM^T\left(MPM^T+R\right)^{-1}$

Eq4: $\left(I+PM^TR^{-1}M\right)^{-1}P=\left(P^{-1}+M^TR^{-1}M\right)^{-1}$

Proof: Take $()^{-1}$ from the left side, $\left[\left(I+PM^TR^{-1}M\right)^{-1}P\right]^{-1}=P^{-1}\cdot\left(I+PM^TR^{-1}M\right)=P^{-1}+M^TR^{-1}M.$

Eq5: Since the left hand side of #4 and #2 is the same we get $\left(P^{-1}+M^TR^{-1}M\right)^{-1}=P-PM^T\left(MPM^T+R\right)^{-1}MP$

Eq6: Similarly, we insert Eq #4 into #3 and get

$\left(P^{-1}+M^TR^{-1}M\right)^{-1}M^TR^{-1}=\left(I+PM^TR^{-1}M\right)^{-1}PM^TR^{-1}=PM^T\left(MPM^T+R\right)^{-1}$

1.8.4 Orthogonal projection of random variables

We can treat functions of random variables as inner product vector spaces. More specifically we define the σ-algebra of a random variable X as:

$$\sigma\left(X\right)=span\left\{\delta_{x_1}\left(X\right),...,\delta_{x_n}\left(X\right)\right\} \tag{1.44}$$

with $A=\{x_1,...,x_n\}$ being the possible values of X and $\delta_{x_i}\left(X\right)=\begin{cases}1 & X=x_i\\ 0 & X\neq x_i\end{cases}$ is an indicator function.

Defining the inner product, $\langle X,Y\rangle=E\left(XY\right)$, we see that $\left\{\delta_{x_i}\left(X\right)\right\}_{i=1}^{n}$ is an orthogonal basis of $\sigma\left(X\right)$. Also, by definition $\left\langle\delta_{x_i}\left(X\right),\delta_{y_j}\left(Y\right)\right\rangle=P\left(X=x_i,Y=y_j\right)$.

The orthogonal projection of $\delta_{x_i}\left(X\right)$ on $\sigma\left(Y\right)$ is thus

$$\sum_{j=1}^{m}\frac{\left\langle\delta_{x_i}\left(X\right),\delta_{y_j}\left(Y\right)\right\rangle}{\left\langle\delta_{y_j}\left(Y\right),\delta_{y_j}\left(Y\right)\right\rangle}\delta_{y_j}\left(Y\right)=\sum_{j=1}^{m}\frac{\left\langle\delta_{x_i}\left(X\right),\delta_{y_j}\left(Y\right)\right\rangle}{\left\langle\delta_{y_j}\left(Y\right)\right\rangle}\delta_{y_j}\left(Y\right)=E\left(\delta_{x_i}\left(X\right)|Y\right) \tag{1.45}$$

and this holds true to any function of X.

Example 1 (conditional expectance): Let $\sigma_L\left(Y\right)=span\left\{Y-E\left(Y\right),1\right\}$ (an orthogonal basis). This means that projecting X on $\sigma_L\left(Y\right)$ gives $Cov\left(X,Y\right)\cdot Var\left(Y\right)^{-1}\cdot\left(Y-E\left(Y\right)\right)+E\left(X\right)$. If X,Y are jointly normal this results holds for $\sigma\left(Y\right)$ itself and we get:

$$E\left(X|Y\right)=E\left(X\right)+Cov\left(X,Y\right)\cdot Var\left(Y\right)^{-1}\cdot\left(Y-E\left(Y\right)\right) \tag{1.46}$$

Example 2 (orthogonality): As in any vector orthogonal projection we get that $X-Cov\left(X,Y\right)\cdot Var\left(Y\right)^{-1}\cdot\left(Y-E\left(Y\right)\right)$ is orthogonal to $Y-E\left(Y\right)$.

Example 3 (uncorrelated variables): Let Y,Z be uncorrelated ($Cov\left(Y,Z\right)=0$). The projection of X on $\left(Y,Z\right)-\left\langle \left(Y,Z\right)\right\rangle$ is $\left(Cov\left(X,Y\right),Cov\left(X,Z\right)\right)\cdot Var\left(Y,Z\right)^{-1}\cdot\left(\begin{array}{c} Y-E\left(Y\right)\\ Z-E\left(Z\right)\end{array}\right)=Cov\left(X,Y\right)\cdot Var\left(Y\right)^{-1}\cdot\left(Y-E\left(Y\right)\right)+Cov\left(X,Z\right)\cdot Var\left(Z\right)^{-1}\cdot\left(Z-E\left(Z\right)\right)$

Example 4 If $E\left(YZ\right)=\left\langle Y,Z\right\rangle =0$ then the projection of X on Y,Z is $\left(\left\langle X,Y\right\rangle ,\left\langle X,Z\right\rangle \right)\cdot\left(\begin{array}{cc} \left\langle Y,Y\right\rangle & \left\langle Z,Y\right\rangle \\ \left\langle Z,Y\right\rangle & \left\langle Z,Z\right\rangle \end{array}\right)^{-1}\cdot\left(\begin{array}{c} Y\\ Z\end{array}\right)=\left\langle X,Y\right\rangle \cdot\left\langle Y,Y\right\rangle ^{-1}\cdot Y+\left\langle X,Z\right\rangle \cdot\left\langle Z,Z\right\rangle ^{-1}\cdot Z$

1.8.5 Orthogonal projection and discrete optimal linear smoothing (Meditch 1967)

Let:

$$\begin{aligned} x_{k+1} &= \phi_{k+1,k}\cdot x_{k}+u_{k}\\ z_{k} &= H_{k}\cdot x_{k}+\nu_{k}\end{aligned}\tag{1.47}$$

define the dynamics and observation process. $x_{k}\in\mathbb{R}^{n},z_{k}\in\mathbb{R}^{m},\phi\in\mathbb{R}^{n\times n},H\in\mathbb{R}^{m\times m}$ and $\left\langle u_{j}u_{k}^{T}\right\rangle =Q_{k}\cdot\delta_{kj},\left\langle \nu_{j}\nu_{k}^{T}\right\rangle =R_{k}\cdot\delta_{kj}$

Also, $\left\langle x_{0}\right\rangle =0$ and $\left\langle x_{0}x_{0}^{T}\right\rangle =P_{0}$.

Estimation The estimation of x_{k} using data points 1...j is $\hat{x}_{k|j}$.

Error The estimation error is $\widetilde{x}_{k|j}=x_{k}-\hat{x}_{k|j}$. An optimal estimation minimizes the squared error $\left\langle \widetilde{x}_{k|j}^{2}\right\rangle$.

$\hat{x}_{k|j}$ is a member of the vector space $Y_{j}=\left\{ \sum_{i=1}^{j}B_{i}z_{i}|B_{i}\in\mathbb{R}^{n\times m}\right\}$

Orthogonal projections Define the orthogonal projection of x_{k} on Y_{j} as $\bar{x}_{k|j}$ and it follows:
- $x_{k}-\bar{x}_{k|j}\perp Y_{j}$
- if $x_{k}-\zeta\perp Y_{j}$ and $\zeta\in Y_{j}$ then $\zeta=\bar{x}_{k|j}$

Theorem 1: $\bar{x}_{k|j}=\hat{x}_{k|j}$

Proof 1: - $\hat{x}_{k|j}\in Y_{j}$ by definition
- Rewrite $\left\langle \left(x_{k}-\hat{x}_{k|j}\right)^{2}\right\rangle =\left\langle \left(x_{k}-\bar{x}_{k|j}\right)^{2}\right\rangle +2\cdot\left\langle \left(x_{k}-\bar{x}_{k|j}\right)\cdot\left(\bar{x}_{k|j}-\hat{x}_{k|j}\right)\right\rangle +\left(\bar{x}_{k|j}-\hat{x}_{k|j}\right)^{2}$.

The middle part is zero by the orthogonality of $\left(x_{k}-\bar{x}_{k|j}\right)$ to all members of the vector space Y_{j} (including $\left(\bar{x}_{k|j}-\hat{x}_{k|j}\right)$) and the last part is non-negative.

From the optimality of $\hat{x}_{k|j}$ we conclude that $\bar{x}_{k|j}=\hat{x}_{k|j}$.

Theorem 2 (Kalman): $\hat{x}_{k+1|k}=\phi_{k+1,k}\cdot\hat{x}_{k}$ and $M_{k+1}=\phi_{k,k+1}P_{k}\phi_{k,k+1}^{T}+Q_{k}$
with the covariance matrix $M_{k+1}=\left\langle \tilde{x}_{k+1|k}\cdot\tilde{x}_{k+1|k}^{T}\right\rangle$

Define $\hat{z}_{k|j}=H_{k}\cdot\hat{x}_{k|j}$ and $\tilde{z}_{k|j}=z_{k}-\hat{z}_{k|j}$ and the vector space $Z_{k+1}=\left\{ K_{k+1}\cdot\tilde{z}_{k+1|k}|K\in\mathbb{R}^{n\times m}\right\}$ and get

Lemma: $Z_{k+1}$is orthogonal to Y_{k}.

Proof: consider the basic vector $\tilde{z}_{k+1|k} = z_{k+1} - H_{k+1}\phi_{k+1,k}\hat{x}_k$(Kalman) which can be expanded to $= H_{k+1}x_{k+1} + \nu_{k+1} - H_{k+1}\phi_{k+1,k}\hat{x}_k = H_{k+1}\phi_{k+1k} \cdot (x_k - \hat{x}_k) + H_{k+1}u_k + \nu_{k+1}$. Now, u_k and ν_{k+1} are gaussian with mean zero and are trivially orthogonal to Y_k and $x_k - \hat{x}_k$ is orthogonal to Y_k from Theorem 1.

1.9 References

"Stochastic processes and filtering theory" - Andrew H.Jazwinski (AP 1970 Vol 64)