Modifications to the CPM for the paper *Difference Detection in LC-MS Data for Protein Biomarker Discovery***, by J. Listgarten, R. Neal, S. Roweis, P. Wong, and A. Emili**

1 Multi-Dimensional Time Series

In the original CPM paper (Listgarten *et al.*, 2005), the CPM was a generative model for a set of K scalar time series, $\vec{x}^k = (x_1^k, x_2^k, ..., x_{N^k}^k)$. We here extend the CPM to accomodate vector time series rather than just scalar time series, and we do so in the simplest way possible. Please refer to the original CPM paper for details upon which these extensions depend.

Let a multi-dimensional observed LC-MS run (*i.e.*, with more than one *m/z* bin – not just the TIC) be $\mathcal{X}^k = (\vec{x}_1^k, \vec{x}_2^k, ..., \vec{x}_N^k)$, with entries $\mathcal{X}_{id}^k$ where i runs from 1 to N (length of the time series), d runs from 1 to D, the dimensionality of the vector at each time point (in the Difference Detection paper, $D = 4$), and k runs from 1 to K, the number of experimental LC-MS samples. Let the latent trace be $\mathcal{Z} = (\vec{z}_1, \vec{z}_2, ..., \vec{z}_M)$, with entries, $\mathcal{Z}_{id}$, where i runs from 1 to M (the number of hidden time states), and d runs from 1 to D.

The emission probabilities of the original CPM become multi-dimensional Gaussians with (spherical) covariance, Σ

$$\begin{aligned} A_{\pi_i}(\vec{x}_i^k) &\equiv p(\vec{x}_i^k | \pi_i^k, \vec{z}_{\tau_i^k}) = \mathcal{N}(\vec{x}_i^k; \vec{z}_{\tau_i^k}\phi_i u_k, \Sigma) \\ &= (2\pi)^{-\frac{D}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\vec{x}_i^k - \vec{z}_{\tau_i^k}\phi_i u_k)^T \Sigma^{-1} (\vec{x}_i^k - \vec{z}_{\tau_i^k}\phi_i u_k)\right] \end{aligned} \tag{1}$$

and the log likelihood, $\mathcal{L}^{cp}$, stays the same as in the CPM, except for i) the emission probabilities just listed, and ii) the smooting prior/penalty on the latent trace, which now sums over all dimensions, D:

$$\mathcal{L}^{\text{class}} \equiv \sum_{k=1}^{K} \left(\log p(\pi_1) + \sum_{i=1}^{N} \log A_{\pi_i}(\vec{x}_i^k) + \sum_{i=2}^{N} \log T_{\pi_{i-1},\pi_i}^k \right)$$

$$\begin{aligned} \mathcal{L}^{cp} &\equiv \mathcal{L}^{\text{class}} + \mathcal{P} \\ &= \mathcal{L}^{\text{class}} + \sum_{k=1}^{K} \log \mathcal{D}(d_v^k | \{\eta_v^k\}) + \log \mathcal{D}(s_v | \{\eta_v'\}) - \lambda \sum_{d=1}^{D} \sum_{j=1}^{\tau-1} (\mathcal{Z}_{j+1,d} - \mathcal{Z}_{j,d})^2. \end{aligned}$$

The E-Step remains unchanged as do the updates for the scale and state transision probabilities (except that emisson probabilities are now multi-dimensional).

Letting the diagonal entries of the diagonal covariance matrix, Σ, be σ, then the updates for the spherical covariance matrices become:

$$\sigma^2 = \frac{\sum_{d=1}^{D} \sum_k \sum_{s=1}^{S} \sum_{i=1}^{N} \gamma_s^k(i) (\mathcal{X}_{i,d}^k - \mathcal{Z}_{\tau_s,d} u_k \phi_s)^2}{NSD}$$

where S denotes the number of hidden time states.

Lastly, note that the updates for the latent traces are independent for each of the D emission dimensions and so the derivates are essentially unchanged.

2 Replacing Hidden Scale States with A Set of Linear Spline Control Points

We eliminate the hidden scale states, and instead introduce scale flexibility by way of a set of parameters. So now the emission probabilities change again, slightly, to

$$\begin{aligned} A_{\pi_i}(\vec{x}_i^k) &\equiv p(\vec{x}_i^k|\pi_i^k, \vec{z}_{\tau_i^k}) = \mathcal{N}(\vec{x}_i^k; \vec{z}_{\tau_i^k}\phi_i u_{\tau_i}^k, \Sigma) \\ &= (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\vec{x}_i^k - \vec{z}_{\tau_i^k}\phi_i u_{\tau_i}^k)^T\Sigma^{-1}(\vec{x}_i^k - \vec{z}_{\tau_i^k}\phi_i u_{\tau_i}^k)\right] \end{aligned}$$

where all we have done is to change u_k to $u_{\tau_i}^k$ as compared to Equation 1 above, and these new $u_{\tau_i}^k$ will be defined by a linear spline, over latent time, as follows:

Each spline tells us by what amount to scale each observed point relative to the latent trace from which it is being generated. The more control points we use, the more local the scaling. The same scaling spline is used for all dimensions, d. For each spline we use G control points, evenly spaced in latent time. Note that the positions of the control points is fixed ahead of time and never changes during learning. The first control point is located at the first latent time point (*i.e.*, 1), and the last control point at the last latent time point (*i.e.*, M). For any latent time, τ which does not have a control point, let $c_{1\tau}$ be the latent time of the control point immediately to the left of τ and $c_{2\tau}$ be the latent time of the control point immediately to the right of hidden time state τ. Let μ_j^k be the value of the control point occuring at latent time j (only defined for those latent times which have a control point), then the scaling value defined by the spline at any latent time, τ, for observed time series k, represented by u_τ^k is given by

$$u_\tau^k = \frac{(\tau - c_{1\tau})\mu_{kc_{2\tau}} + (c_{2\tau} - \tau)\mu_{kc_{1\tau}}}{c_{2\tau} - c_{1\tau}}, \tag{2}$$

unless τ itself has a control point, in which case $u_\tau^k = \mu_\tau^k$.

During the M-step of training, we need to update the value at each control point. This is easily accomplished by taking the derivate of the log likelihood with respect to each control point, and feeding it, along with the log likelihood, to an optimization routine. For M-step updates of other parameters, we can simply replace u_k with $u_{\tau_i}^k$ which are now defined by the scaling spline. We make similar (trivial) substitutions for the E-step.

Bibliography

Jennifer Listgarten, Radford M. Neal, Sam T. Roweis, and Andrew Emili. Multiple alignment of continuous time series. In L. K. Saul et al, editor, Advances in Neural Information Processing Systems, volume 17. The MIT Press, 2005.