Model-based Clustering for Neural Populations

The goal for this research is to do model-based clustering for neural populations, by making use of features for each counting process observation.

1 Notations

Assume we can observe neural activities for N neurons, with counting observation up to T steps. Therefore, the observation is a *N-by-T* matrix, $\mathbf{Y} \in \mathbb{Z}_{\geq 0}^{N \times T}$,with each row represents the recording from single neuron. Denote the recording for neuron i as $\mathbf{y}_i = (y_{i1}, \ldots, y_{\mathrm{iT}})'$, $i = 1, \ldots N.$, with the cluster index for neuron i as $z_i \in \{1, \ldots\}$. The number of neurons in cluster j is $n_j = \sum_{i=1}^{N} I(z_i = j)$, and $\sum_{j=1,2,\ldots} n_j = N$. The proportion/ probability in cluster z_i is ρ_{z_i}.

2 Clustering Wrapper

The model-based clustering problem can be transformed into fitting the mixture model (MM). The likelihood for each cluster depends on how we model the counting observation, but fitting strategies for MM are the same for all models. Here, I choose to fit the MM by Gibbs sampler. Depending on whether the number of cluster is finite or not, there are two versions: finite mixture model (FMM) and Dirichlet process mixture model (DPMM).

2.1 Finite Mixture Model

Assume the number of cluster is J. The full likelihood for these N neurons is

$$L = \prod_{i=1}^{N} \rho_{z_i} f\left(\mathbf{y}_i | \mathbf{\Theta}_{z_i}\right) = \prod_{j=1}^{J} \rho_j^{n_j} \left[\prod_{i:z_i=j} f\left(\mathbf{y}_i | \mathbf{\Theta}_j\right) \right]$$

, where $\mathbf{\Theta}_j$ contains all parameters in cluster j defined by the specific model. Therefore, the parameters need to update are:

(1) Cluster indicator: $\{z_i\}_{i=1}^{N}$

(2) Cluster proportion: $\rho = (\rho_1, \ldots \rho_J)'$

(3) Model parameters: $\mathbf{\Theta}_j$

The (conditional) priors for clustering-related parameters:

(1) Cluster indicator $\{z_i\}_{i=1}^N$: $P(z_i = j) = \rho_j$

(2) Cluster proportion $\rho = (\rho_1, \dots \rho_J)'$:

$$\rho \sim Dir(\delta_1, \dots \delta_J)$$

, where $\delta_1 = \ldots = \delta_J = 1$

So, the MCMC(Gibbs sampler) iteration for FMM is:

(1) Update $\{z_i\}_{i=1}^N$:

$$P\left(z_i = j \middle| \mathbf{y}_i, \{\mathbf{\Theta}_j\}_{j=1}^J\right) \propto \rho_j f(\mathbf{y}_i|\mathbf{\Theta}_j)$$

(2) Update $\rho = (\rho_1, \dots \rho_J)^{'}$:

$$\rho| \ \{\mathbf{y}_i\}_{i=1}^N, \ \{z_i\}_{i=1}^N, \{\mathbf{\Theta}_j\}_{j=1}^J \sim Dir(\delta_1 + n_1, \dots \delta_J + n_J)$$

(3) Update $\mathbf{\Theta}_j$: this is defined by the specific model. When there's no $z_i = j$, just sample $\mathbf{\Theta}_j$ from priors or by other observation-independent ways.

2.2 Dirichlet Process Mixture Model

Since calculation of posterior predictive distribution can be hard or even impossible for complicated models, instead of using the popular CRP representation of DP (Neal, 2020), I choose to use the slice sampler (Walker, 2007).

Use the "stick-breaking" construction for cluster proportion, i.e.

$$\rho_1 = \eta_1$$

$$\rho_j = (1 - \eta_1) \cdot \ldots \cdot (1 - \eta_{j-1}) \eta_j$$

$$\eta_j \sim Beta(1, \alpha)$$

In the slice sampler for DPMM, the parameters need to update are:

(1) "stick-breaking" elements: η_j

(2) Augment latent variable: $\{u_i\}_{i=1}^N$

(3) Model parameters: $\mathbf{\Theta}_j$

(4) Cluster indicator: $\{z_i\}_{i=1}^N$

So, the MCMC(Gibbs sampler) iteration for DPMM is:

(1) update η_j, for $j = 1, \ldots, z^* = max\left\{z_i\right\}_{i=1}^N$ as

$$\eta_j | \left\{z_i\right\}_{i=1}^N, \ldots \sim Beta(n_j + 1,\ N - \sum_{l=1}^{j} n_l + \alpha)$$

(2) update $\left\{u_i\right\}_{i=1}^N$:

$$u_i | \rho, \ldots \sim U(0,\ \rho_{z_i})$$

(3) update η_j, for $j = z^* + 1, \ldots,\ s^*$. s^* is the smallest value, s.t. $\sum_{j=1}^{s^*} \rho_j > 1 - \min\left\{u_1, \ldots, u_N\right\}$

$$\eta_j \sim Beta(1, \alpha)$$

(4) Update $\mathbf{\Theta}_j$: this is defined by the specific model. When there's no $z_i = j$, just sample $\mathbf{\Theta}_j$ from priors or by other observation-independent ways.

(5) Update $\left\{z_i\right\}_{i=1}^N$

$$P\left(z_i = j \middle| \mathbf{y}_i, \left\{\mathbf{\Theta}_j\right\}, \rho, \left\{u_i\right\}_{i=1}^N\right) = \frac{f\left(\mathbf{y}_i | \mathbf{\Theta}_j\right)}{\sum_{j:\rho_j > u_i} f\left(\mathbf{y}_i | \mathbf{\Theta}_j\right)}$$

3 linear Dynamical System Model

Here, I model the observations by a linear dynamical system (LDS) model.

LDS models the multi-dimensional time series using a lower dimensional latent representation of the system, which evolves over time according to linear dynamics. By specifying the linear dynamics and process noise covariance, we can also handle the interactions between different neural populations (clusters).

3.1 Model Details

Denote the latent vector in cluster j as $\mathbf{x}_t^{(j)} \in \mathbb{R}^{p_j}$. For simplicity, assume all $p_j = p$. Each observation follows a Poisson distribution:

$$\log \lambda_{\text{it}} = d_i + \mathbf{c'}_i \mathbf{x}_t^{(z_i)}$$

$$y_{\text{it}} \sim Poisson(\lambda_{\text{it}})$$

, where $\mathbf{c}_i \in \mathbb{R}^p$ and $\mathbf{x}_t^{(z_i)} \in \mathbb{R}^p$.

Although the loading (d_i and $\mathbf{c}_i$) is determined by neuron index i, the distribution is also cluster-dependent. That is,

$$(d_i, \mathbf{c}_i')' \sim N(\boldsymbol{\mu}_{\text{dc}}^{(z_i)}, \mathbf{\Sigma}_{\text{dc}}^{(z_i)})$$

By doing this, the loading within each cluster is also correlated.

Denote all latent states as $\mathbf{x}_t = \left(\mathbf{x}_t'^{(1)}, \mathbf{x}_t'^{(2)}, \dots\right)'$ and they evolve linearly with a Gaussian noise:

$$\mathbf{x}_1 \sim N(\mathbf{x}_0, \mathbf{Q}_0)$$

$$\mathbf{x}_{t+1}|\mathbf{x}_t \sim N(\mathbf{A}\mathbf{x}_t + \mathbf{b}, \mathbf{Q})$$

For simplicity, assume $\mathbf{Q}_0$ is known (e.g. $\mathbf{Q}_0 = \mathbf{I} \times 10^{-2}$).

If we assume process noise covariance is block diagonal (Joshua et al., 2020), we can write things as:

$$\mathbf{x}_{t+1}^{(j)}|\mathbf{x}_t^{(1)}, \mathbf{x}_t^{(2)}, \dots \sim N(\sum_{l=1,\dots} \mathbf{A}_{j\leftarrow l}\mathbf{x}_t^{(l)} + \mathbf{b}_j, \mathbf{Q}^{(j)})$$

Notice $\{\mathbf{A}_{j\leftarrow l}\}$ forms the full transition matrix as:

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{1\leftarrow 1} & \mathbf{A}_{1\leftarrow 2} & \dots \\ \mathbf{A}_{2\leftarrow 1} & \mathbf{A}_{2\leftarrow 2} & \dots \\ \dots & \dots & \dots \end{pmatrix}$$

Denote the j^{th} row block of $\mathbf{A}$ as $\mathbf{A}_j = \begin{pmatrix} \mathbf{A}_{j\leftarrow 1} & \mathbf{A}_{j\leftarrow 2} & \dots \end{pmatrix}$. Then, $\sum_{l=1,\dots} \mathbf{A}_{j\leftarrow l}\mathbf{x}_t^{(l)} + \mathbf{b}_j = \mathbf{A}_j\mathbf{x}_t + \mathbf{b}_j$.

If we further let $\mathbf{Q}$ be diagonal, with the k^{th} row of $\mathbf{x}_t$, $\mathbf{A}$, $\mathbf{b}$ denoted as x_{kt}, $\mathbf{a}_k$, b_k. The corresponding process noise variance is q_k. Then:

$$x_{k,t+1}|x_{\text{kt}} \sim N\left(\mathbf{a}_k'\mathbf{x}_t + b_k, q_k\right)$$

For completeness, Gibbs samplers for all three versions (no constraint, block diagonal and diagonal $\mathbf{Q}$) are shown below.

Since the progress noise is independent at each step, $f\left(\mathbf{y}_i|\mathbf{\Theta}_j\right) = \prod_{t=1}^{T} P(y_{\text{it}}|\mathbf{\Theta}_j)$, where $P(\cdot)$ is the Poisson density and $\mathbf{\Theta}_j$ contains all parameters in cluster j.

3.2 Conditional Priors for Parameters

The parameters need to estimate:

(1) Latent vectors: $\{\mathbf{x}_t\}_{t=1}^{T}$

(2) Initials: $\mathbf{x}_0$

(3) Linear mapping (loading) for latent vectors: $\{d_i\}_{i=1}^{N}$ and $\{\mathbf{c}_i\}_{i=1}^{N}$

(4) Mean and covariance for loading in each cluster: $\left\{\boldsymbol{\mu}_{\text{dc}}^{(j)}\right\}_j$ and $\left\{\mathbf{\Sigma}_{\text{dc}}^{(j)}\right\}_j$

(5) Linear dynamics for latent vectors: $\mathbf{A}$ and $\mathbf{b}$

(6) Process noise: $\mathbf{Q}$

The conditional priors for these parameters:

(1) Latent vectors $\{\mathbf{x}_t\}_{t=1}^T$: the conditional prior is defined by

$$\mathbf{x}_1 \sim N(\mathbf{x}_0, \mathbf{Q}_0)$$

$$\mathbf{x}_{t+1}|\mathbf{x}_t \sim N(\mathbf{A}\mathbf{x}_t + \mathbf{b}, \mathbf{Q})$$

(2) Initials $\mathbf{x}_0$: assume there are J clusters,

$$\mathbf{x}_0 \sim N(\boldsymbol{\mu}_{\mathbf{x}_{00}},\ \boldsymbol{\Sigma}_{\mathbf{x}_{00}})$$

, where $\boldsymbol{\mu}_{\mathbf{x}_{00}} = \mathbf{0}_{\mathrm{Jp}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}_{00}} = \mathbf{I}_{\mathrm{Jp}}$

(3) Linear mapping (loading) for latent vectors $\{d_i\}_{i=1}^N$ and $\{\mathbf{c}_i\}_{i=1}^N$:

$$(d_i, \mathbf{c}_i')' \sim N(\boldsymbol{\mu}_{\mathrm{dc}}^{(z_i)}, \boldsymbol{\Sigma}_{\mathrm{dc}}^{(z_i)})$$

(4) Mean and covariance for loading in each cluster $\left\{\boldsymbol{\mu}_{\mathrm{dc}}^{(j)}\right\}_j$ and $\left\{\boldsymbol{\Sigma}_{\mathrm{dc}}^{(j)}\right\}_j$:

$$\boldsymbol{\mu}_{\mathrm{dc}}^{(j)} \sim N(\delta_{dc0}, \mathbf{T}_{dc0})$$

, where $\boldsymbol{\delta}_{dc0} = \mathbf{0}_{p+1}$ and $\mathbf{T}_{dc0} = \mathbf{I}_{p+1}$

$$\boldsymbol{\Sigma}_{\mathrm{dc}}^{(j)} \sim W^{-1}(\Psi_{dc0}, \nu_{dc0})$$

, where $\nu_{dc0} = p + 1 + 2$ and $\Psi_{dc0} = \mathbf{I}_{p+1} \times 10^{-4}$

(5) Linear dynamics for latent vectors $\mathbf{A}$ and $\mathbf{b}$: if the number of cluster is J

(a) No constraints on $\mathbf{Q}$: denote $\widetilde{\mathbf{A}} = vec(\mathbf{A})$

$$\left(\mathbf{b}', \widetilde{\mathbf{A}}\right)' \sim N(\boldsymbol{\mu}_{BA_0}, \boldsymbol{\Sigma}_{BA_0})$$

, where $\boldsymbol{\mu}_{BA_0} = \left(\mathbf{0}_{\mathrm{Jp}}', vec(\mathbf{I}_{\mathrm{Jp}})'\right)'$ and $\boldsymbol{\Sigma}_{BA_0} = 0.25\mathbf{I}_{Jp(1+Jp)}$

(b) Assume $\mathbf{Q}$ is block-diagonal: denote $\widetilde{\mathbf{A}}_j = vec(\mathbf{A}_j)$

$$\left(\mathbf{b}_j', \widetilde{\mathbf{A}}_j'\right)' \sim N(\boldsymbol{\mu}_{bA_0}, \boldsymbol{\Sigma}_{bA_0})$$

, where $\boldsymbol{\mu}_{bA_0} = \left(\mathbf{0}_{p+(j-1)p^2}', vec(\mathbf{I}_p)', \mathbf{0}_{(J-j)p^2}'\right)'$ and $\boldsymbol{\Sigma}_{bA_0} = 0.25\mathbf{I}_{p(1+pJ)}$

(c) Assume $\mathbf{Q}$ is diagonal:

$$(b_k, \mathbf{a}_k')' \sim N(\mu_{ba_k0}, \boldsymbol{\Sigma}_{ba0})$$

, where $\mu_{ba_k0} = (0, \mathbf{e}_k')'$ and $\boldsymbol{\Sigma}_{ba0} = 0.25\mathbf{I}_{Jp+1}$

(6) Process noise $\mathbf{Q}$: if the number of cluster is J

(a) No constraints on $\mathbf{Q}$:

$$\mathbf{Q} \sim W^{-1}(\Psi_{\mathbf{Q}_0}, \nu_{\mathbf{Q}_0})$$

, where $\nu_{\mathbf{Q}_0} = Jp + 2$ and $\Psi_{\mathbf{Q}_0} = \mathbf{I}_{\mathrm{Jp}} \times 10^{-4}$. (To make the mean of $\mathbf{Q}$ loosely centered around $\mathbf{I}_{Jp} \times 10^{-4}$)

(b) Assume $\mathbf{Q}$ is block-diagonal:

$$\mathbf{Q}^{(j)} \sim W^{-1}(\Psi_0, \nu_0)$$

, where $\nu_0 = p + 2$ and $\Psi_0 = \mathbf{I}_p \times 10^{-4}$. (To make the mean of $\mathbf{Q}^{(j)}$ loosely centered around $\mathbf{I}_p \times 10^{-4}$)

(c) Assume $\mathbf{Q}$ is diagonal:

$$q_k \sim IG(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2})$$

, where $\nu_0 = 4$ and $\sigma_0^2 = 10^{-4}$.

3.3 MCMC (Gibbs Sampler)

3.3.1 Update $\{\mathbf{x}_t\}_{t=1}^T$

Use Laplace approximation and make use of the block tri-diagonal Hessian.

Denote t^{th} column of mean firing rate and observation as $\widetilde{\boldsymbol{\lambda}}_t = (\lambda_{1t}, \ldots, \lambda_{\mathrm{Nt}})'$ and $\widetilde{\mathbf{y}}_t = (y_{1t}, \ldots, y_{\mathrm{Nt}})'$. The linear mapping matrix for all observations is $\mathbf{C}$, such that $\log \widetilde{\boldsymbol{\lambda}}_t = \mathbf{d} + \mathbf{C}\mathbf{x}_t$. Let $\mathbf{x} = (\mathbf{x}'_1, \ldots, \mathbf{x}'_T)'$ and $f(\mathbf{x}) = \log P(\mathbf{x} | \{\mathbf{y}_i\}_{i=1}^N, \mathbf{C}, \mathbf{Q}_0, \mathbf{A}, \mathbf{b}, \mathbf{Q}, \ldots)$ The first and second derivative with respect to $\mathbf{x}$, for $t = 2, \ldots, T-1$:

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{x}_1} &= \mathbf{C}'\left(\widetilde{\mathbf{y}}_1 - \widetilde{\boldsymbol{\lambda}}_1\right) - \mathbf{Q}_0^{-1}(\mathbf{x}_1 - \mathbf{x}_0) + \mathbf{A}'\mathbf{Q}^{-1}(\mathbf{x}_2 - \mathbf{A}\mathbf{x}_1 - \mathbf{b}) \\
\frac{\partial f}{\partial \mathbf{x}_t} &= \mathbf{C}'\left(\widetilde{\mathbf{y}}_t - \widetilde{\boldsymbol{\lambda}}_t\right) - \mathbf{Q}^{-1}(\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1} - \mathbf{b}) + \mathbf{A}'\mathbf{Q}^{-1}(\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t - \mathbf{b}) \\
\frac{\partial f}{\partial \mathbf{x}_T} &= \mathbf{C}'\left(\widetilde{\mathbf{y}}_T - \widetilde{\boldsymbol{\lambda}}_T\right) - \mathbf{Q}^{-1}(\mathbf{x}_T - \mathbf{A}\mathbf{x}_{T-1} - \mathbf{b}) \\
\frac{\partial^2 f}{\partial \mathbf{x}_1 \partial \mathbf{x}_1'} &= -\mathbf{C}'\mathrm{Diag}\left(\widetilde{\boldsymbol{\lambda}}_1\right)\mathbf{C} - \mathbf{Q}_0^{-1} - \mathbf{A}'\mathbf{Q}^{-1}\mathbf{A} \\
\frac{\partial^2 f}{\partial \mathbf{x}_t \partial \mathbf{x}_t'} &= -\mathbf{C}'\mathrm{Diag}\left(\widetilde{\boldsymbol{\lambda}}_t\right)\mathbf{C} - \mathbf{Q}^{-1} - \mathbf{A}'\mathbf{Q}^{-1}\mathbf{A} \\
\frac{\partial^2 f}{\partial \mathbf{x}_T \partial \mathbf{x}_T'} &= -\mathbf{C}'\mathrm{Diag}\left(\widetilde{\boldsymbol{\lambda}}_T\right)\mathbf{C} - \mathbf{Q}^{-1} \\
\frac{\partial^2 f}{\partial \mathbf{x}_1 \partial \mathbf{x}_2'} &= \frac{\partial^2 f}{\partial \mathbf{x}_t \partial \mathbf{x}_{t+1}'} = \mathbf{A}'\mathbf{Q}^{-1} \qquad\qquad \frac{\partial^2 f}{\partial \mathbf{x}_t \partial \mathbf{x}_{t-1}'} = \mathbf{Q}^{-1}\mathbf{A}
\end{aligned}$$

So, the gradient is:

$$\nabla = \frac{\partial f}{\partial \mathbf{x}} = \left(\left(\frac{\partial f}{\partial \mathbf{x}_1} \right)', \ \ldots, \left(\frac{\partial f}{\partial \mathbf{x}_T} \right)' \right)'$$

And the block tri-diagonal Hessian:

$$H = \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}'} = \begin{pmatrix} \frac{\partial^2 f}{\partial \mathbf{x}_1 \partial \mathbf{x}_1'} & \mathbf{A}'\mathbf{Q}^{-1} & 0 & \cdots & 0 \\ \mathbf{Q}^{-1}\mathbf{A} & \frac{\partial^2 f}{\partial \mathbf{x}_2 \partial \mathbf{x}_2'} & \mathbf{A}'\mathbf{Q}^{-1} & \cdots & \vdots \\ 0 & \mathbf{Q}^{-1}\mathbf{A} & \frac{\partial^2 f}{\partial \mathbf{x}_3 \partial \mathbf{x}_3'} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \frac{\partial^2 f}{\partial \mathbf{x}_T \partial \mathbf{x}_T'} \end{pmatrix}$$

Use Newton-Raphson to find $\boldsymbol{\mu}_{\mathbf{x}} = \text{argmax}_{\mathbf{x}} \left(f \left(\mathbf{x} \right) \right)$ and $\boldsymbol{\Sigma}_{\mathbf{x}} = - \left[\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}'} \ |_{\mathbf{X}=\boldsymbol{\mu}_{\mathbf{X}}} \right]^{-\mathbf{1}}$, such that $(P(\mathbf{x}| \left\{ \mathbf{y}_i \right\}_{i=1}^{N}, \ldots) \ \approx \ N \left(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}} \right)$. When using Newton-Raphson (NR), $H \backslash \nabla$ in MATLAB will make use of block tri-diagonal structure automatically.

However, NR is not robust to bad initials. At the first few iterations, simply using fitting from previous step may lead to infinite Hessian. When the initial from previous step fails, use the approximation at recursive priors, , i.e. the adaptive smoother estimates, as the initial. The adaptive smoother estimates are from backward RTS smoother from adaptive filter, and the details about Poisson adaptive filter can be found in Eden et al., 2004.

To sample efficiently and make best use of sparse covariance, use Cholesky decomposition of $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} = \mathbf{R}\mathbf{R}'$: sample $\mathbf{Z} \sim N(\mathbf{R}'\mu_{\mathbf{x}}, \mathbf{I})$, then $\mathbf{x} = \left(\mathbf{R}' \right)^{-1} \mathbf{Z} \sim N(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}})$.

3.3.2 Update $\mathbf{x}_0$

$$P \left(\mathbf{x}_0 | \mathbf{x}_1, \ \mathbf{Q}_0 \ldots \right) \propto N(\mathbf{x}_1 | \mathbf{x}_0, \ \mathbf{Q}_0) N(\mathbf{x}_0 | \boldsymbol{\mu}_{\mathbf{x}_{00}}, \ \boldsymbol{\Sigma}_{\mathbf{x}_{00}})$$

By conjugacy, $\mathbf{x}_0 | \mathbf{x}_1, \ \mathbf{Q}_0 \ldots \sim N(\boldsymbol{\mu}_{\mathbf{x}_0}, \ \boldsymbol{\Sigma}_{\mathbf{x}_0})$

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{x}_0} &= \left[\boldsymbol{\Sigma}_{\mathbf{x}_{00}}^{-1} + \mathbf{Q}_0^{-1} \right]^{-1} \\ \boldsymbol{\mu}_{\mathbf{x}_0} &= \boldsymbol{\Sigma}_{\mathbf{x}_0} \left(\boldsymbol{\Sigma}_{\mathbf{x}_{00}}^{-1} \boldsymbol{\mu}_{\mathbf{x}_{00}} + \mathbf{Q}_0^{-1} \mathbf{x}_1 \right) \end{aligned}$$

3.3.3 Update $\{d_i\}_{i=1}^N$ and $\{\mathbf{c}_i\}_{i=1}^N$

To update efficiently, use Laplace approximation again. Denote $(d_i, \mathbf{c}_i')' = \boldsymbol{\zeta}_i \in \mathbb{R}^{p+1}$ and $\left(1, \mathbf{x}_t^{\prime(z_i)}\right) = \widetilde{\mathbf{x}}_t^{\prime(z_i)}$.

$$P\left(\boldsymbol{\zeta}_i \middle| \mathbf{y}_i,\ \left\{\mathbf{x}_t^{(z_i)}\right\}_{t=1}^T, \dots\right) = \exp f\left(\boldsymbol{\zeta}_i\right) \approx N\left(\boldsymbol{\zeta}_i | \boldsymbol{\mu}_{\boldsymbol{\zeta}_i}, \boldsymbol{\Sigma}_{\boldsymbol{\zeta}_i}\right)$$

$$\frac{\partial f}{\partial \boldsymbol{\zeta}_i} = \frac{\partial l}{\partial \boldsymbol{\zeta}_i} - {\boldsymbol{\Sigma}_{\mathrm{dc}}^{(z_i)}}^{-1}\left(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_{\mathrm{dc}}^{(z_i)}\right) = \left[\sum_{t=1}^T \widetilde{\mathbf{x}}_t^{(z_i)}\left(y_{\mathrm{it}} - \lambda_{\mathrm{it}}\right)\right] - {\boldsymbol{\Sigma}_{\mathrm{dc}}^{(z_i)}}^{-1}\left(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_{\mathrm{dc}}^{(z_i)}\right)$$

$$\frac{\partial^2 f}{\partial \boldsymbol{\zeta}_i \partial \boldsymbol{\zeta}_i'} = \frac{\partial^2 l}{\partial \boldsymbol{\zeta}_i \partial \boldsymbol{\zeta}_i'} - {\boldsymbol{\Sigma}_{\mathrm{dc}}^{(z_i)}}^{-1} = -\left[\sum_{t=1}^T \lambda_{\mathrm{it}} \widetilde{\mathbf{x}}_t^{(z_i)} \widetilde{\mathbf{x}}_t^{\prime(z_i)}\right] - {\boldsymbol{\Sigma}_{\mathrm{dc}}^{(z_i)}}^{-1}$$

, where l is Poisson log-likelihood. Then use Newton-Raphson to find $\boldsymbol{\mu}_{\boldsymbol{\zeta}_i} = \mathrm{argmax}_{\boldsymbol{\zeta}_i}\left(f\left(\boldsymbol{\zeta}_i\right)\right)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\zeta}_i} = -\left[\frac{\partial^2 f}{\partial \boldsymbol{\zeta}_i \partial \boldsymbol{\zeta}_i'} |_{\boldsymbol{\zeta}_i = \boldsymbol{\mu}_{\boldsymbol{\zeta}_i}}\right]^{-1}$. If initial as the previous step fits fails, simply use prior mean of $\boldsymbol{\mu}_{\boldsymbol{\zeta}_i}$, i.e. $\boldsymbol{\delta}_{dc0}$.

3.3.4 Update $\left\{\mu_{\mathbf{dc}}^{(j)}\right\}_j$ and $\left\{\boldsymbol{\Sigma}_{\mathbf{dc}}^{(j)}\right\}_j$

As above, denote $(d_i, \mathbf{c}_i')' = \boldsymbol{\zeta}_i \in \mathbb{R}^{p+1}$.

(1) Mean $\left\{\boldsymbol{\mu}_{\mathrm{dc}}^{(j)}\right\}_j$: by conjugacy, $\boldsymbol{\mu}_{\mathrm{dc}}^{(j)} \sim N\left(\boldsymbol{\delta}_{\mathrm{dc}}, \mathbf{T}_{\mathrm{dc}}\right)$

$$\mathbf{T}_{\mathrm{dc}}^{-1} = \left(\mathbf{T}_{dc0}^{-1} + n_j {\boldsymbol{\Sigma}_{\mathrm{dc}}^{(j)}}^{-1}\right)^{-1}$$

$$\boldsymbol{\delta}_{\mathrm{dc}} = \mathbf{T}_{\mathrm{dc}}\left(\mathbf{T}_{dc0}^{-1}\boldsymbol{\delta}_{dc0} + {\boldsymbol{\Sigma}_{\mathrm{dc}}^{(j)}}^{-1} \sum_{i:z_i=j} \boldsymbol{\zeta}_i\right)$$

(2) Covariance $\left\{\boldsymbol{\Sigma}_{\mathrm{dc}}^{(j)}\right\}_j$: by conjugacy, $\boldsymbol{\Sigma}_{\mathrm{dc}}^{(j)} \sim W^{-1}\left(\Psi_{\mathrm{dc}}, \nu_{\mathrm{dc}}\right)$

$$\nu_{\mathrm{dc}} = n_j + \nu_{dc0}$$

$$\Psi_{\mathrm{dc}} = \Psi_{dc0} + \sum_{i:z_i=j}\left(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_{\mathrm{dc}}^{(j)}\right)\left(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_{\mathrm{dc}}^{(j)}\right)'$$

3.3.5 Update A and b

Assume there are J clusters.

(1) No constraints on $\mathbf{Q}$:

$$P\left(\left(\mathbf{b}', \widetilde{\mathbf{A}}'\right)' \middle| \{\mathbf{x}_t\}_{t=1}^T, \mathbf{Q}, \dots\right) \propto \left[\prod_{t=2}^T N\left(\mathbf{x}_t | \mathbf{A}\mathbf{x}_{t-1} + \mathbf{b}, \mathbf{Q}\right)\right] P\left(\left(\mathbf{b}', \widetilde{\mathbf{A}}'\right)'\right)$$

Let $\left(\mathbf{b}', \widetilde{\mathbf{A}}'\right)' = \boldsymbol{\gamma} \in \mathbb{R}^{Jp(1+Jp)}$, $\left(1,\ \mathbf{x}'_{t-1}\right) \bigotimes \mathbf{I}_{\mathrm{Jp}} = \widetilde{\mathbf{G}}_{t-1} \in \mathbb{R}^{Jp \times Jp(1+pJ)}$ and $\widetilde{\mathbf{G}} = \left(\widetilde{\mathbf{G}}'_1, \ldots, \widetilde{\mathbf{G}}'_{T-1}\right)'$. By conjugacy, $\boldsymbol{\gamma} | \left\{\mathbf{x}_t\right\}_{t=1}^T, \mathbf{Q}, \ldots \sim N(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}})$

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \left[\boldsymbol{\Sigma}_{BA_0}^{-1} + \widetilde{\mathbf{G}}' \mathbf{I}_{T-1} \bigotimes (\mathbf{Q})^{-1} \widetilde{\mathbf{G}}\right]^{-1}$$
$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \left(\boldsymbol{\Sigma}_{BA_0}^{-1} \boldsymbol{\mu}_{BA_0} + \widetilde{\mathbf{G}}' \mathbf{I}_{T-1} \bigotimes (\mathbf{Q})^{-1} \left(\mathbf{x}_2', \ldots, \mathbf{x}_T'\right)'\right)$$

(2) Assume $\mathbf{Q}$ is block-diagonal:

$$P\left(\left(\mathbf{b}'_j, \widetilde{\mathbf{A}}'_j\right)' \middle| \left\{\mathbf{x}_t^{(j)}\right\}_{t=1}^T, \mathbf{Q}^{(j)}, \ldots\right) \propto \left[\prod_{t=2}^T N\left(\mathbf{x}_t^{(j)} | \mathbf{A}_j \mathbf{x}_{t-1} + \mathbf{b}_j, \mathbf{Q}^{(j)}\right)\right] P\left(\left(\mathbf{b}'_j, \widetilde{\mathbf{A}}'_j\right)'\right)$$

Notice that $\mathbf{A}_j \mathbf{x}_{t-1} + \mathbf{b}_j = \left(\left(1,\ \mathbf{x}'_{t-1}\right) \bigotimes \mathbf{I}_p\right) \left(\mathbf{b}'_j, \widetilde{\mathbf{A}}'_j\right)'$, then the problem is reduced to Bayesian linear regression. Denote $\left(1,\ \mathbf{x}'_{t-1}\right) \bigotimes \mathbf{I}_p = \widetilde{\mathbf{X}}_{t-1} \in \mathbb{R}^{p \times p(1+pJ)}$, $\widetilde{\mathbf{X}} = \left(\widetilde{\mathbf{X}}'_1, \ldots, \widetilde{\mathbf{X}}'_{T-1}\right)'$ and $\left(\mathbf{b}'_j, \widetilde{\mathbf{A}}'_j\right)' = \gamma_j \in \mathbb{R}^{p(1+pJ)}$. Then

$$\begin{aligned}\prod_{t=2}^T N\left(\mathbf{x}_t^{(j)} | \mathbf{A}_j \mathbf{x}_{t-1} + \mathbf{b}_j, \mathbf{Q}^{(j)}\right) &= \prod_{t=2}^T N\left(\mathbf{x}_t^{(j)} | \widetilde{\mathbf{X}}_{t-1} \gamma_j, \mathbf{Q}^{(j)}\right) \\ &= N\left(\left(\mathbf{x}_2'^{(j)}, \ldots, \mathbf{x}_T'^{(j)}\right)' | \widetilde{\mathbf{X}} \boldsymbol{\gamma}_j,\ \mathbf{I}_{T-1} \bigotimes \mathbf{Q}^{(j)}\right)\end{aligned}$$

By conjugacy, $P\left(\boldsymbol{\gamma}_j \middle| \left\{\mathbf{x}_t^{(j)}\right\}_{t=1}^T, \mathbf{Q}^{(j)}, \ldots\right) = N\left(\boldsymbol{\gamma}_j | \boldsymbol{\mu}_{\boldsymbol{\gamma}_j}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_j}\right)$

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_j} = \left[\boldsymbol{\Sigma}_{bA_0}^{-1} + \widetilde{\mathbf{X}}' \mathbf{I}_{T-1} \bigotimes \left(\mathbf{Q}^{(j)}\right)^{-1} \widetilde{\mathbf{X}}\right]^{-1}$$
$$\boldsymbol{\mu}_{\boldsymbol{\gamma}_j} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_j} \left(\boldsymbol{\Sigma}_{bA_0}^{-1} \boldsymbol{\mu}_{bA_0} + \widetilde{\mathbf{X}}' \mathbf{I}_{T-1} \bigotimes \left(\mathbf{Q}^{(j)}\right)^{-1} \left(\mathbf{x}_2'^{(j)}, \ldots, \mathbf{x}_T'^{(j)}\right)'\right)$$

(3) Assume $\mathbf{Q}$ is diagonal:

$$P\left(\left(b_k, \mathbf{a}'_k\right)' | \left\{x_{kt}\right\}_{t=1}^T, q_k,\ \ldots\right) \propto \left[\prod_{t=2}^T N\left(x_{\mathrm{kt}} | \mathbf{a}'_k \mathbf{x}_{t-1} + b_k, q_k\right)\right] P\left(\left(b_k, \mathbf{a}'_k\right)'\right)$$

Rewrite $\mathbf{a}'_k \mathbf{x}_{t-1} + b_k = \left(1, \mathbf{x}'_{t-1}\right)\left(b_k, \mathbf{a}'_k\right)'$. Then the problem is reduced to Bayesian linear regression. Denote $\left(b_k, \mathbf{a}'_k\right)' = \boldsymbol{\gamma}_k \in \mathbb{R}^{Jp+1}$ and $\widetilde{\mathbf{X}} = \left(\left(1, \mathbf{x}'_1\right)', \ldots, \left(1, \mathbf{x}'_{T-1}\right)'\right)'$. Then by conjugacy, $P\left(\boldsymbol{\gamma}_k | \left\{x_{\mathrm{kt}}\right\}_{t=1}^T, q_k,\ \ldots\right) =$

$N\left(\boldsymbol{\gamma}_k|\boldsymbol{\mu}_{\boldsymbol{\gamma}_k},\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_k}\right)$

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_k}=\left[\boldsymbol{\Sigma}_{ba0}^{-1}+\frac{\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}}{q_k}\right]^{-1}$$
$$\boldsymbol{\mu}_{\boldsymbol{\gamma}_k}=\boldsymbol{\Sigma}_{\gamma_k}\left(\boldsymbol{\Sigma}_{ba0}^{-1}\mu_{ba0}+\widetilde{\mathbf{X}}'\left(x_{k2},\ldots,x_{\mathrm{kT}}\right)'/q_k\right)$$

3.3.6 Update Q

(1) No constraints on $\mathbf{Q}$:
Let $\mathbf{A}\mathbf{x}_{t-1}+\mathbf{b}=\boldsymbol{\mu}_{x_t}$

$$P\left(\mathbf{Q}\middle|\left\{\mathbf{x}_t\right\}_{t=1}^{T},\mathbf{A},\mathbf{b},\ldots\right)\propto\left[\prod_{t=2}^{T}N\left(\mathbf{x}_t|\boldsymbol{\mu}_{\mathbf{x}_t},\ \mathbf{Q}\right)\right]W^{-1}(\mathbf{Q}|\Psi_{\mathbf{Q}_0},\nu_{\mathbf{Q}_0})$$

By conjugacy,

$$P\left(\mathbf{Q}\middle|\left\{\mathbf{x}_t^{(j)}\right\}_{t=1}^{T},\ \mathbf{A},\mathbf{b},\ldots\right)=W^{-1}\left(\mathbf{Q}|\Psi_0+\sum_{t=2}^{T}\left(\mathbf{x}_t-\boldsymbol{\mu}_{\mathbf{x}_t}\right)\left(\mathbf{x}_t-\boldsymbol{\mu}_{\mathbf{x}_t}\right)',\ T-1+\nu_0\right)$$

(2) Assume $\mathbf{Q}$ is block-diagonal:
Let $\mathbf{A}_j\mathbf{x}_{t-1}+\mathbf{b}_j{=}\boldsymbol{\mu}_{\mathbf{x}_t^{(j)}}$,

$$P\left(\mathbf{Q}^{(j)}\middle|\left\{\mathbf{x}_t^{(j)}\right\}_{t=1}^{T},\ \mathbf{A}_j,\mathbf{b}_j,\ldots\right)\propto\left[\prod_{t=2}^{T}N\left(\mathbf{x}_t^{(j)}|\boldsymbol{\mu}_{\mathbf{x}_t^{(j)}},\mathbf{Q}^{(j)}\right)\right]W^{-1}\left(\mathbf{Q}^{(j)}|\Psi_0,\nu_0\right)$$

By conjugacy,

$$P\left(\mathbf{Q}^{(j)}\middle|\left\{\mathbf{x}_t^{(j)}\right\}_{t=1}^{T},\ \mathbf{A}_j,\mathbf{b}_j,\ldots\right)=W^{-1}\left(\mathbf{Q}^{(j)}|\Psi_0+\sum_{t=2}^{T}\left(\mathbf{x}_t^{(j)}-\boldsymbol{\mu}_{\mathbf{x}_t^{(j)}}\right)\left(\mathbf{x}_t^{(j)}-\boldsymbol{\mu}_{\mathbf{x}_t^{(j)}}\right)',\ T-1+\nu_0\right)$$

(3) Assume $\mathbf{Q}$ is diagonal:
Let $\mathbf{a}_k^{'}\mathbf{x}_{t-1}+b_k=\mu_{x_{\mathrm{kt}}}$,

$$P\left(q_k|\left\{x_{\mathrm{kt}}\right\}_{t=1}^{T},b_k,\mathbf{a}_k^{'},\ \ldots\right)\propto\left[\prod_{t=2}^{T}N\left(x_{\mathrm{kt}}|\mu_{x_{\mathrm{kt}}},q_k\right)\right]\mathrm{IG}\left(q_k\middle|\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}\right)$$

By conjugacy,

$$P\left(q_k|\left\{x_{\mathrm{kt}}\right\}_{t=1}^{T},b_k,\mathbf{a}_k^{'},\ \ldots\right)=IG\left(q_k|\frac{\nu_0+T-1}{2},\frac{\nu_0\sigma_0^2+\sum_{t=2}^{T}\left(x_{\mathrm{kt}}-\mu_{x_{\mathrm{kt}}}\right)^2}{2}\right)$$

4 Simulations

There are two set of simulation examples. The first example is generated from LDS model directly, while in the second example the latents are generated directly without explicit specifying linear dynamics. In all of the following results, the covariance of process noise $\mathbf{Q}$ is assumed to be block diagonal. The code for non-constrained and diagonal $\mathbf{Q}$ can be found in XXX and XXX. The fitting results for all these three are similar, because the underlying true $\mathbf{Q}$ is diagonal.

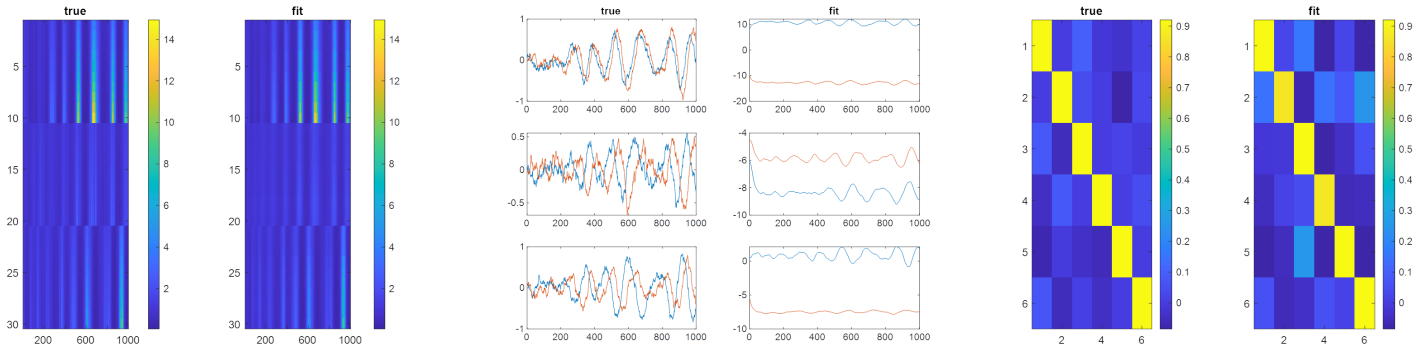
Before doing the clustering, I first assume the cluster labels are known to see the fitting performance. All the fitting results with labels are shown in means from iteration 50 to 100.

4.1 Simulation 1: Generate from LDS Directly

In this simulation, there are 3 clusters with 10 neurons in each cluster. The dimension of latents in each cluster is 2. In the linear dynamics, the bias term $\mathbf{b}$ is zero. There's no within-population interaction but has some weak between-population interactions. In other words, the linear dynamics matrix $\mathbf{A}$ is roughly diagonal. The details of simulation can be found in the first section of XXX.

4.1.1 Labeled Data: No Clustering

The code for fitting is XXX, and some results are in Figure 1.



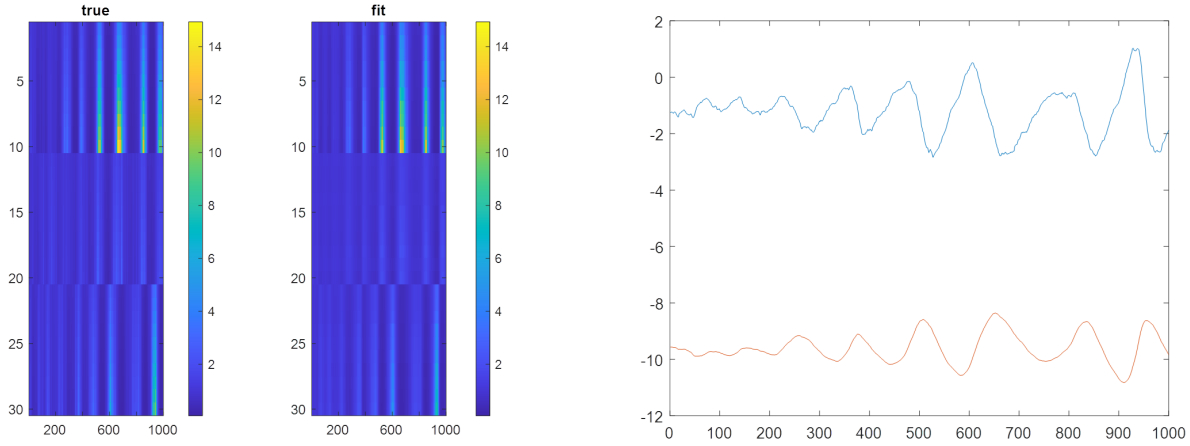
(a) overall mean firing rate (b) latents (c) dynamics $\mathbf{A}$

Figure 1: LDS sample with labels

The fitting for overall mean firing rate is perfect and the dynamics pattern is mostly recovered. The fitting of latents is OK up to scale. In other words, the model captures the oscillating pattern of latents.

But the latent patterns of these three clusters seems quite similar. That means the differences in the mean firing rates may be purely explained by the loading,

i.e. $\mathbf{d}$ and $\mathbf{C}$. To see that, I set the number of cluster to be 1 and the results are shown in Figure 2.



(a) overall mean firing rate

(b) latents

Figure 2: LDS sample with labels, one cluster

Again, the fitting for overall mean firing rate is perfect. This is why it's important to make the loading, $\mathbf{d}$ and $\mathbf{C}$, also be cluster-dependent (loading within clusters are correlated). If the loading only depends on neuron index and will not change for different clustering assignments, it's impossible to do clustering (at least in this case), since the loading is enough to capture all the patterns.

4.1.2 Unlabeled Data: Clustering

To give the full path of clustering, I show results in GIFs. The code for FMM and DPMM can be found in XXX and XXX.

There are three results:

(1) Fit by FMM with $\mathbf{J} = 3$, starting from random cluster assignments:

(2) Fit by FMM with $\mathbf{J} = 3$, starting from all neurons in single cluster:

(3) Fit by DPMM ($\alpha = X$), starting from each neuron forms its own cluster:

I didn't show the DPMM starting from single cluster. Since in my current implementation, the cluster generation is not efficient. In other words, the newly generated cluster will usually not be sampled and the algorithm usually gets stuck in 1 or 2 clusters.

This is a big problem... That means if we occasionally combine two clusters into one, we may never correct the mistake. Further, when the data grows, the

number of cluster will be hard to grow. Fix that later.

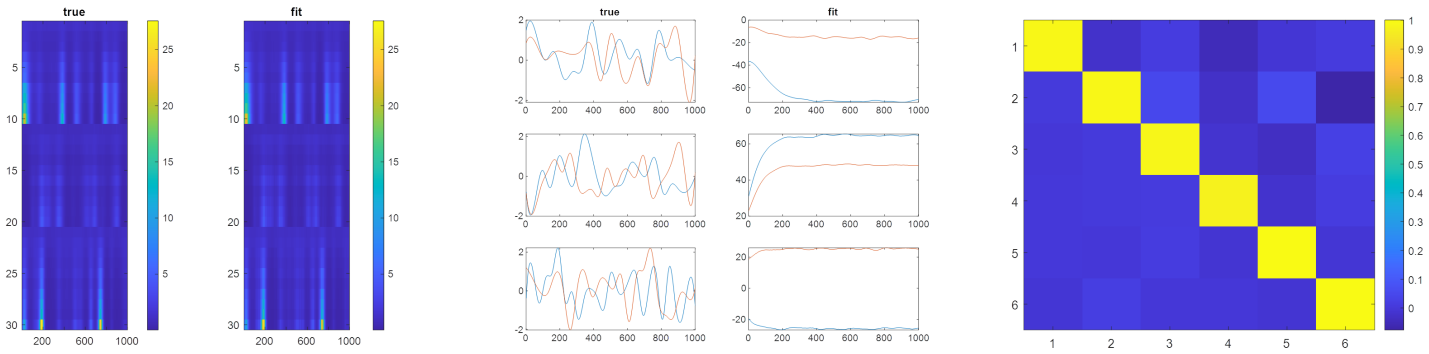
4.2 Simulation 2: Generate Latents, without Specifying Linear Dynamics

In this simulation, there are again 3 clusters with 2 latents in each. Besides set 10 neurons in each cluster, I further simulate 50 neurons in each cluster to see performance in a larger scale. Now, the latents are generated directly without specifying the underlying linear dynamics of latents. However, each latent is generated independently, so the linear dynamics matrix **A** should be roughly diagonal. The details of simulation can be found in the first section of XXX.

4.2.1 Labeled Data: No Clustering

As in simulation 1, the data is fitted by the true cluster assignment (3 clusters) and forcing all neurons belong to single cluster. The code can be found in XXX.

(1) Smaller dataset (10 neurons each):
The results with true cluster assignment is shown in Figure 3 And results



(a) overall mean firing rate (b) latents (c) dynamics **A**

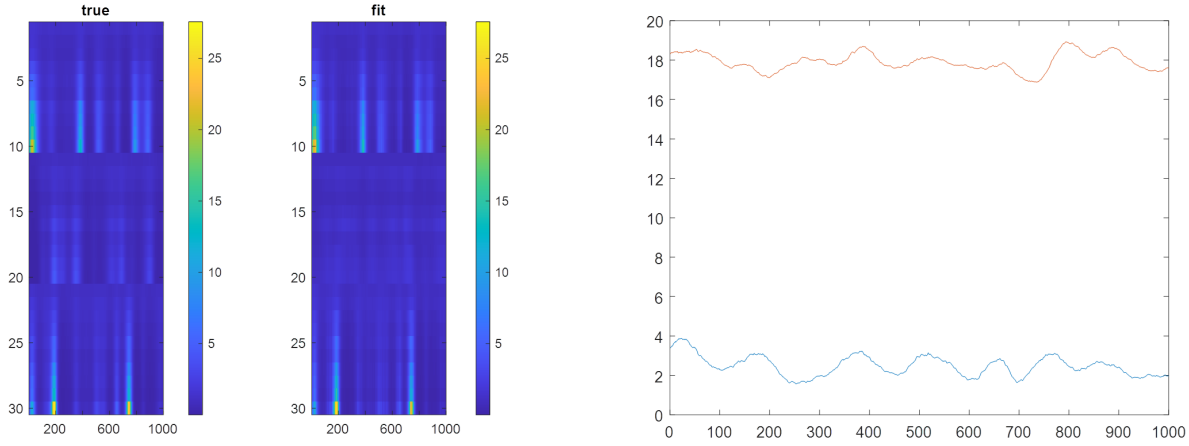
Figure 3: unspecified dynamics with labels, 10 neurons each

by forcing all neurons to be in single cluster (Figure 4)

Again, the fittings for overall mean firing rate are perfect in both cases and the dynamics is roughly diagonal. The oscillating pattern of fitted latents with true cluster index is not significant, because of plot scaling. The pattern is easier to see in the second fitting.

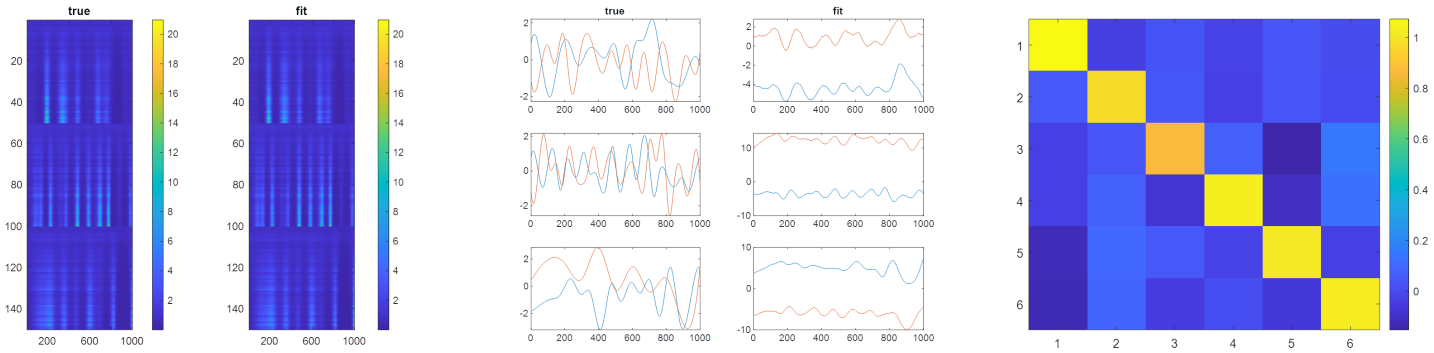
(2) Larger dataset (50 neurons each):
The results with true cluster assignment is shown in Figure 5 And results by forcing all neurons to be in single cluster (Figure 6)

Now we can see the pattern more clearly.



(a) overall mean firing rate (b) latents

Figure 4: unspecified dynamics with labels, one cluster, 10 neurons each



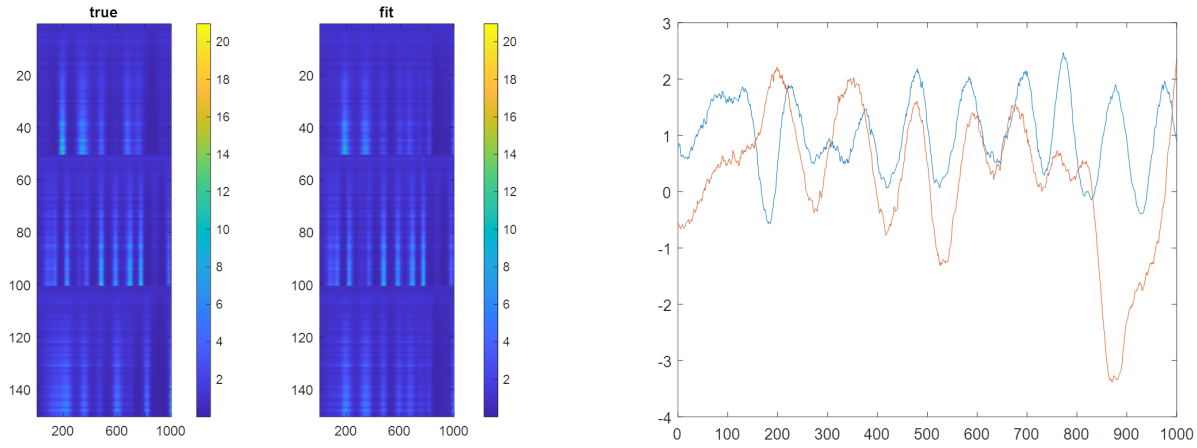
(a) overall mean firing rate (b) latents (c) dynamics **A**

Figure 5: unspecified dynamics with labels, 50 neurons each

4.2.2 Unlabeled Data: Clustering

As previous, in each example, I show results from (1) FMM starting from random cluster assignment ($J = 3$), (2) FMM starting from single cluster and (3) DPMM starting from multiple clusters. The DPMM starting from single cluster need to be fixed later.

(1) Smaller dataset (10 neurons each):



(a) overall mean firing rate

(b) latents

Figure 6: unspecified dynamics with labels, one cluster, 50 neurons each

(a) Fit by FMM with $\mathbf{J} = 3$, start from random cluster assignments:

(b) Fit by FMM with $\mathbf{J} = 3$, start from single cluster:

(c) Fit by DP ($\alpha = X$), start from assuming each neuron forms its own cluster:

(2) Larger dataset (50 neurons each):

(a) Fit by FMM with $\mathbf{J} = 3$, start from random cluster assignments:

(b) Fit by FMM with $\mathbf{J} = 3$, start from single cluster:

(c) Fit by DP ($\alpha = X$), start from random assignment for 10 clusters:

The performance is not bad. The neuron at top of each cluster cannot be clustered correctly by its nature: the signals are too weak to be clearly clustered.

5 TODO

(1) Find a more efficient way to generate new cluster parameters, otherwise the newly generated ones will always be rejected.

(2) improve DPMM