Factor Analysis with Poisson Output

Gopal Santhanam[1]
Byron Yu[1]
Krishna V. Shenoy[1,2]
[1]Department of Electrical Engineering,[2]Neurosciences Program
Stanford University
Stanford, CA 94305, USA
{gopals,byronyu,shenoy}@stanford.edu

Maneesh Sahani
Gatsby Computational Neuroscience Unit
University College London
17 Queen Square, London WC1N 3AR, UK
maneesh@gatsby.ucl.ac.uk

Technical Report NPSL-TR-06-1

March 19, 2006

Abstract

We derive a modified version of factor analysis for data that is poisson (rather than gaussian) distributed. This modified approach may better fit certain classes of data, including neuronal spiking data commonly collected in electrophysiology experiments.

1 Introduction

Factor analysis and other similar dimensionality reduction approaches (e.g., PCA or SPCA) are derived using a state-space model. The latent state is modeled as a gaussian distribution. The observed output is modeled as a linear function of the latent state with additive gaussian noise. This approach can provide the benefit of reducing the dimensionality of the observed, but noisy, data to a small number of underlying factors. These factors may then be used to provide meaningful predictions on new data.

For count, or point process, data, the gaussian output noise model used in factor analysis may not provide a good description of the data. Instead, we modify the output noise model to be poisson. Additionally, we extend the state-space model to incorporate a mixture of gaussians rather than a single gaussian distribution. This extension can serve to better model the latent state, especially when there is an *a priori* expectation that data is clustered. Once trained, the model can be used to make predictions of the latent (or unobserved) states for new observed data.

We dub our new approach as "Factor Analysis with Poisson Output", or FAPO for short.

2 Generative Model

The generative model for FAPO is given below.

$$\mathbf{x} \mid s \sim \mathcal{N}\left(\boldsymbol{\mu}_s, \Sigma_s\right) \tag{1}$$

$$y^i \mid \mathbf{x} \sim \text{Poisson}(h(\mathbf{c}^i \cdot \mathbf{x} + d^i)\Delta) \quad \text{for} \quad i \in 1, \ldots, q \tag{2}$$

The random variable s is the mixture component indicator and has a discrete probability distribution over $\{1,\ldots,M\}$ (i.e., $P(s) = \pi_s$). Given s, the latent state vector, $\mathbf{x} \in \mathbb{R}^{p\times 1}$, is gaussian distributed with mean $\boldsymbol{\mu}_s$ and covariance Σ_s. The outputs, $y^i \in \mathbb{N}_0$, are generated from a poisson distribution where h is a link function mapping $\mathbb{R} \to \mathbb{R}_+$, $\mathbf{c}^i \in \mathbb{R}^{p\times 1}$ and $d^i \in \mathbb{R}$ are constants, and $\Delta \in \mathbb{R}$ is the time bin width. We collect the counts from all q simultaneously observed variables into a vector $\mathbf{y} \in \mathbb{N}_0^{q\times 1}$, whose ith element is y^i. The choice of the link function h is discussed in the following section.

In this work, we assume that all of the parameters of the model, namely π_s, $\boldsymbol{\mu}_s$, Σ_s, $\mathbf{c}^i$, and d^i for $s \in \{1,\ldots,M\}$ and $i \in \{1,\ldots,q\}$, are unknown. The goal is to learn the parameters so that the model can be used to make predictions of $\mathbf{x}$ and s for a new $\mathbf{y}$.

3 System Identification

The procedure of system identification, or "model training," requires learning the parameters from the observed data. The observed data includes N observations of $\mathbf{y}$, an i.i.d. sequence $(\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N)$ denoted by $\{\mathbf{y}\}$, and N observations of the mixture component indicators, s, an i.i.d. sequence $(s_1,s_2,\ldots,s_N)$ denoted by $\{s\}$. The latent state vectors are hidden and not observed.

This situation is an *unsupervised* problem, although not completely unsupervised; the system identification can be more challenging if s is also unknown. This latter scenario is beyond the scope of this article. Once the model is trained, however, we estimate the most likely $\mathbf{x}$ *and* s with new observed data y, as described in the following section.

The standard approach to system identification in the presence of unobserved latent variables is the **Expectation-Maximization** (or **EM**) algorithm. The algorithm maximizes the likelihood of the model parameters (i.e., $\theta = \{\pi_s, \boldsymbol{\mu}_{1,\ldots,M}, \Sigma_{1,\ldots,M}, \mathbf{c}^{1,\ldots,q}, d^{1,\ldots,q}\}$) over the observed data. The algorithm is iterative and each iteration is performed in two parts, the expectation (E) step and the maximization (M) step. Iterations are performed until the likelihood converges.

3.1 E-step

The E-step of EM requires computing the expected log joint likelihood, $E\left[\log P\left(\{\mathbf{x}\},\{\mathbf{y}\},\{s\} \mid \theta\right)\right]$, over the posterior distribution of the hidden state vector, $P\left(\{\mathbf{x}\} \mid \{\mathbf{y}\}, s, \theta^k\right)$, where θ^k are the parameter estimates at the kth EM iteration. Since the observations are i.i.d. we can equivalently maximize the sum of the individual expected log joint likelihoods, $E\left[\log P\left(\mathbf{x},\mathbf{y},s \mid \theta\right)\right]$.

The posterior distribution can be expressed as follows:

$$P\left(\mathbf{x} \mid \mathbf{y}, s, \theta^k\right) \propto P\left(\mathbf{y} \mid \mathbf{x}, \theta^k\right) P\left(\mathbf{x} \mid s, \theta^k\right). \tag{3}$$

Because $P(\mathbf{y} \mid \mathbf{x})$ is a product of poissons rather than a gaussian, the state posterior $P(\mathbf{x} \mid \mathbf{y})$ will not be of a form that allows for easy computation of the log joint likelihood. Instead, as in [1], we approximated this posterior with a gaussian centered at the mode of $\log P(\mathbf{x} \mid \mathbf{y})$ and whose covariance is given by the negative inverse hessian of the log posterior at that mode. Certain choices of h, including $h_1(z) = e^z$ and $h_2(z) = \log(1+e^z)$, lead to a log posterior that is strictly concave in $\mathbf{x}$. In these cases, the unique mode can easily be found by Newton's method.

$$\begin{aligned}
\log P(\mathbf{x} \mid \mathbf{y}, s, \theta) &= \log P(\mathbf{y} \mid \mathbf{x}, \theta) + \log P(\mathbf{x} \mid s, \theta) + C_1 \\
&= \left(\sum_{i=1}^{q} \log P\left(y^i \mid \mathbf{x}\right)\right) + \log \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_s, \Sigma_s\right) + C_2 \\
&= \left(\sum_{i=1}^{q} -h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right)\Delta + y^i \log\left(h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right)\Delta\right) - \log\left(y^i!\right)\right) + \log \mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}_s, \Sigma_s\right) + C_3 \\
&= \left(\sum_{i=1}^{q} -h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right)\Delta + y^i \log h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right)\right) - \frac{1}{2}\mathbf{x}'\Sigma_s^{-1}\mathbf{x} + \boldsymbol{\mu}_s'\Sigma_s^{-1}\mathbf{x} + C_4
\end{aligned} \tag{4}$$

Taking the gradient and hessian of (4) with respect to $\mathbf{x}$, results in the following expressions.

$$\nabla_{\mathbf{x}} \log P(\mathbf{x} \mid \mathbf{y}, s, \theta) = \left(\sum_{i=1}^{q} -\nabla_{\mathbf{x}} h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right) \Delta + y^i \nabla_{\mathbf{x}} \log h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right) \right) - \Sigma_s^{-1} \mathbf{x} + \Sigma_s^{-1} \boldsymbol{\mu}_s$$

$$\nabla^2_{\mathbf{x}} \log P(\mathbf{x} \mid \mathbf{y}, s, \theta) = \left(\sum_{i=1}^{q} -\nabla^2_{\mathbf{x}} h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right) \Delta + y^i \nabla^2_{\mathbf{x}} \log h\left(\mathbf{c}^i \cdot \mathbf{x} + d^i\right) \right) - \Sigma_s^{-1}$$

Letting, $\zeta^i = \mathbf{c}^i \cdot \mathbf{x} + d^i$. For the aforementioned versions of h_1 and h_2, the gradient and hessians are

$$\begin{aligned} \nabla_{\mathbf{x}} \log P(\mathbf{x} \mid \mathbf{y}, s, \theta) &= \left(\sum_{i=1}^{q} -e^{\zeta^i} \mathbf{c}^i \Delta + y^i \mathbf{c}^i \right) - \Sigma_s^{-1} \mathbf{x} + \Sigma_s^{-1} \boldsymbol{\mu}_s && (5) \\ &= \left(\sum_{i=1}^{q} \left(-e^{\zeta^i} \Delta + y^i \right) \mathbf{c}^i \right) - \Sigma_s^{-1} \mathbf{x} + \Sigma_s^{-1} \boldsymbol{\mu}_s && (6) \end{aligned}$$

$$\nabla^2_{\mathbf{x}} \log P(\mathbf{x} \mid \mathbf{y}, s, \theta) = \left(\sum_{i=1}^{q} -e^{\zeta^i} \mathbf{c}^i (\mathbf{c}^i)' \Delta \right) - \Sigma_s^{-1} \qquad (7)$$

and

$$\begin{aligned} \nabla_{\mathbf{x}} \log P(\mathbf{x} \mid \mathbf{y}, s, \theta) &= \left(\sum_{i=1}^{q} -\frac{e^{\zeta^i}}{1 + e^{\zeta^i}} \mathbf{c}^i \Delta + y^i \frac{e^{\zeta^i}}{\left(1 + e^{\zeta^i}\right) \log\left(1 + e^{\zeta^i}\right)} \mathbf{c}^i \right) - \Sigma_s^{-1} \mathbf{x} + \Sigma_s^{-1} \boldsymbol{\mu}_s \\ &= \left(\sum_{i=1}^{q} \left(-\Delta + y^i \frac{1}{\log\left(1 + e^{\zeta^i}\right)} \right) \frac{e^{\zeta^i}}{1 + e^{\zeta^i}} \mathbf{c}^i \right) - \Sigma_s^{-1} \mathbf{x} + \Sigma_s^{-1} \boldsymbol{\mu}_s && (8) \end{aligned}$$

$$\begin{aligned} \nabla^2_{\mathbf{x}} \log P(\mathbf{x} \mid \mathbf{y}, s, \theta) = \sum_{i=1}^{q} \Bigg[& -y^i \frac{\frac{e^{\zeta^i}}{1+e^{\zeta^i}} \mathbf{c}^i}{\left[\log\left(1 + e^{\zeta^i}\right)\right]^2} \frac{e^{\zeta^i}}{1 + e^{\zeta^i}} (\mathbf{c}^i)' x \\ & + \left(-\Delta + y^i \frac{1}{\log\left(1 + e^{\zeta^i}\right)} \right) \frac{e^{\zeta^i} \left(1 + e^{\zeta^i}\right) \mathbf{c}^i - e^{2\zeta^i} \mathbf{c}^i}{\left(1 + e^{\zeta^i}\right)^2} (\mathbf{c}^i)' \Bigg] - \Sigma_s^{-1} \\ = \sum_{i=1}^{q} \Bigg[& -y^i \frac{e^{2\zeta^i}}{\left[\log\left(1 + e^{\zeta^i}\right)\right]^2 \left(1 + e^{\zeta^i}\right)^2} \mathbf{c}^i (\mathbf{c}^i)' \\ & + \left(-\Delta + y^i \frac{1}{\log\left(1 + e^{\zeta^i}\right)} \right) \frac{e^{\zeta^i}}{\left(1 + e^{\zeta^i}\right)^2} \mathbf{c}^i (\mathbf{c}^i)' \Bigg] - \Sigma_s^{-1} \\ = \sum_{i=1}^{q} & \left[\left(-y^i \frac{e^{\zeta^i}}{\left[\log\left(1 + e^{\zeta^i}\right)\right]^2} - \Delta + y^i \frac{1}{\log\left(1 + e^{\zeta^i}\right)} \right) \frac{e^{\zeta^i}}{\left(1 + e^{\zeta^i}\right)^2} \mathbf{c}^i (\mathbf{c}^i)' \right] - \Sigma_s^{-1} \\ = \sum_{i=1}^{q} & \left[\left(-\Delta + y^i \frac{1}{\log\left(1 + e^{\zeta^i}\right)} \left(1 - \frac{e^{\zeta^i}}{\log\left(1 + e^{\zeta^i}\right)} \right) \right) \frac{e^{\zeta^i}}{\left(1 + e^{\zeta^i}\right)^2} \mathbf{c}^i (\mathbf{c}^i)' \right] - \Sigma_s^{-1}, \qquad (9) \end{aligned}$$

respectively.

For observation n, let Q_n be a gaussian distribution in $\mathbb{R}^p$ that approximates $P\left(\mathbf{x}_n \mid \mathbf{y}_n, s_n, \theta^k\right)$ and has mean $\boldsymbol{\xi}_n$ and covariance Ψ_n. The expectation of the log joint likehood for a given observation can be expressed as follows:

$$\mathcal{E}_n = E_{Q_n}\left[\log P\left(\mathbf{x}_n, \mathbf{y}_n, s_n \mid \theta\right)\right] \tag{10}$$

$$= E_{Q_n}\left[\left(\sum_{i=1}^{q} \log P\left(y_n^i \mid \mathbf{x}_n\right)\right) + \log P\left(\mathbf{x}_n \mid s_n\right) + \log P(s_n)\right] \tag{11}$$

$$\begin{aligned} = E_{Q_n}\Bigg[& \left(\sum_{i=1}^{q} -h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta + y_n^i \log\left(h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta\right) - \log\left(y_n^i!\right)\right) \\ & - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log\left(\left|\Sigma_{s_n}\right|\right) - \frac{1}{2}\mathbf{x}_n' \Sigma_{s_n}^{-1}\mathbf{x}_n + \boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1}\mathbf{x}_n - \frac{1}{2}\boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1}\boldsymbol{\mu}_{s_n} \\ & + \log P(s_n)\Bigg]. \end{aligned} \tag{12}$$

The terms that do not depend on $\mathbf{x}_n$ or any component of θ can be grouped as a constant, C, outside the expectation. Doing so, and also moving terms that do not depend on $\mathbf{x}_n$ outside the expectation, we have

$$\begin{aligned} \mathcal{E}_n = E_{Q_n} & \left[\left(\sum_{i=1}^{q} -h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta + y_n^i \log\left(h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta\right)\right) - \frac{1}{2}\mathbf{x}_n' \Sigma_{s_n}^{-1}\mathbf{x}_n + \boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1}\mathbf{x}_n\right] \\ & - \frac{1}{2}\boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1}\boldsymbol{\mu}_{s_n} - \frac{1}{2}\log\left(\left|\Sigma_{s_n}\right|\right) + C \end{aligned}$$

$$\begin{aligned} = E_{Q_n} & \left[\sum_{i=1}^{q} -h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta + y_n^i \log\left(h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta\right)\right] \\ & - \frac{1}{2}E_{Q_n}\left[\mathbf{x}_n' \Sigma_{s_n}^{-1}\mathbf{x}_n\right] + \boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1} E_{Q_n}\left[\mathbf{x}_n\right] \\ & - \frac{1}{2}\boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1}\boldsymbol{\mu}_{s_n} - \frac{1}{2}\log\left(\left|\Sigma_{s_n}\right|\right) + C \end{aligned} \tag{13}$$

$$\begin{aligned} = E_{Q_n} & \left[\sum_{i=1}^{q} -h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta + y_n^i \log\left(h\left(\mathbf{c}^i \cdot \mathbf{x}_n + d^i\right)\Delta\right)\right] \\ & - \frac{1}{2}\mathbf{Tr}\left(\Sigma_{s_n}^{-1}\left(\Psi_n + \boldsymbol{\xi}_n \boldsymbol{\xi}_n'\right)\right) + \boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1}\boldsymbol{\xi}_n \\ & - \frac{1}{2}\boldsymbol{\mu}_{s_n}' \Sigma_{s_n}^{-1}\boldsymbol{\mu}_{s_n} - \frac{1}{2}\log\left(\left|\Sigma_{s_n}\right|\right) + C, \end{aligned} \tag{14}$$

where (13) is simplified to (14) by using the following relationship:

$$\begin{aligned} E_{Q_n}\left[\mathbf{x}_n' \Sigma_{s_n}^{-1}\mathbf{x}_n\right] &= E_{Q_n}\left[\mathbf{Tr}\left(\mathbf{x}_n' \Sigma_{s_n}^{-1}\mathbf{x}_n\right)\right] \\ &= E_{Q_n}\left[\mathbf{Tr}\left(\Sigma_{s_n}^{-1}\mathbf{x}_n \mathbf{x}_n'\right)\right] \\ &= \mathbf{Tr}\left(\Sigma_{s_n}^{-1}\ E_{Q_n}[\mathbf{x}_n \mathbf{x}_n']\right) \\ &= \mathbf{Tr}\left(\Sigma_{s_n}^{-1}\left(\Psi_n + \boldsymbol{\xi}_n \boldsymbol{\xi}_n'\right)\right). \end{aligned}$$

Because the posterior state distributions are approximated as gaussians in the E-step, the expectation in (14) is a gaussian integral that involves non-linear functions g and h and cannot be computed analytically in general. Fortunately, this high-dimensional integral can be reduced to a one-dimensional gaussian integrals (with mean $\mathbf{c}^i \cdot \boldsymbol{\xi}_n$ and variance $(\mathbf{c}^i)' \Psi_n (\mathbf{c}^i)$).

The expectation of the log joint likelihood over all of the N observations is simply the sum of the individual

$\mathcal{E}_n$ terms:

$$
\begin{aligned}
\mathcal{E} &= E_Q\left[\log P(\{\mathbf{x}\},\{\mathbf{y}\},\{s\} \mid \theta)\right] \\
&= \sum_{n=1}^{N} \mathcal{E}_n .
\end{aligned}
$$

3.2 M-step

The M-step requires finding (learning) the $\hat{\theta}^{k+1}$ that satisfies:

$$\hat{\theta}^{k+1} = \arg\max_{\theta} E_Q\left[\log P(\{\mathbf{x}\},\{\mathbf{y}\},\{s\} \mid \theta)\right]. \tag{15}$$

This can achieved by differentiating $\mathcal{E}$ with respect to the parameters, θ, as shown below. The indicator function, $I(s_n = s)$ will prove useful. Also, let $N_s = \sum_{n=1}^{N} I(s_n = s)$.

- Prior probability of mixture component identification s:

$$\pi_s = \frac{1}{N}\sum_{n=1}^{N} I(s_n = s) \tag{16}$$

- State vector mean, for mixture component identification s:

$$\frac{\partial \mathcal{E}}{\partial \boldsymbol{\mu}_s} = \sum_{n=1}^{N} I(s_n = s)\left(\Sigma_s^{-1}\boldsymbol{\xi}_n - \Sigma_s^{-1}\boldsymbol{\mu}_s\right) = 0$$

$$\boldsymbol{\mu}_s^{k+1} = \frac{1}{N_s}\sum_{n=1}^{N} I(s_n = s)\boldsymbol{\xi}_n \tag{17}$$

- State vector covariance, for mixture component identification s:

$$
\begin{aligned}
\frac{\partial \mathcal{E}}{\partial \Sigma_s} &= \sum_{n=1}^{N} I(s_n = s) \\
&\qquad \frac{\partial}{\partial \Sigma_s}\left(-\frac{1}{2}\mathbf{Tr}\left(\Sigma_s^{-1}\left(\Psi_n + \boldsymbol{\xi}_n\boldsymbol{\xi}_n'\right)\right) + \boldsymbol{\mu}_s'\Sigma_s^{-1}\boldsymbol{\xi}_n - \frac{1}{2}\boldsymbol{\mu}_s'\Sigma_s^{-1}\boldsymbol{\mu}_s - \frac{1}{2}\log(|\Sigma_s|)\right) \\
&= \sum_{n=1}^{N} I(s_n = s)\left(\Sigma_s^{-1}\left(\frac{1}{2}\left(\Psi_n + \boldsymbol{\xi}_n\boldsymbol{\xi}_n'\right)' - \boldsymbol{\mu}_s\boldsymbol{\xi}_n' + \frac{1}{2}\boldsymbol{\mu}_s\boldsymbol{\mu}_s'\right)\Sigma_s^{-1} - \frac{1}{2}\Sigma_s^{-1}\right) = 0
\end{aligned}
$$

$$\frac{N_s}{2}\Sigma_s^{-1} = \sum_{n=1}^{N} I(s_n = s)\Sigma_s^{-1}\left(\frac{1}{2}\left(\Psi_n + \boldsymbol{\xi}_n\boldsymbol{\xi}_n'\right) - \boldsymbol{\mu}_s\boldsymbol{\xi}_n' + \frac{1}{2}\boldsymbol{\mu}_s\boldsymbol{\mu}_s'\right)\Sigma_s^{-1}$$

$$
\begin{aligned}
\Sigma_s^{k+1} &= \frac{1}{N_s}\sum_{n=1}^{N} I(s_n = s)\left(\Psi_n + \boldsymbol{\xi}_n\boldsymbol{\xi}_n'\right) - \frac{2}{N_s}\boldsymbol{\mu}_s^{k+1}\sum_{n=1}^{N} I(s_n = s)\boldsymbol{\xi}_n' + \frac{1}{N_s}\boldsymbol{\mu}_s^{k+1}(\boldsymbol{\mu}_s^{k+1})'\sum_{n=1}^{N} I(s_n = s) \\
&= \frac{1}{N_s}\sum_{n=1}^{N} I(s_n = s)\left(\Psi_n + \boldsymbol{\xi}_n\boldsymbol{\xi}_n'\right) - \boldsymbol{\mu}_s^{k+1}(\boldsymbol{\mu}_s^{k+1})' \qquad (18)
\end{aligned}
$$

- Observation mapping constants:

We want to maximize the following objective function, with respect to $\mathbf{c}^i$ and d^i:

$$\begin{aligned}\tilde{\mathcal{E}} &= \sum_{n=1}^{N} E_{Q_n}\left[\sum_{i=1}^{q} -h\left(\mathbf{c}^i\cdot\mathbf{x}_n + d^i\right)\Delta + y_n^i \log\left(h\left(\mathbf{c}^i\cdot\mathbf{x}_n + d^i\right)\Delta\right)\right]\\ &= \sum_{n=1}^{N}\sum_{i=1}^{q} E_{Q_n}\left[-h\left(\mathbf{c}^i\cdot\mathbf{x}_n + d^i\right)\Delta + y_n^i \log\left(h\left(\mathbf{c}^i\cdot\mathbf{x}_n + d^i\right)\Delta\right)\right]\\ &= \sum_{i=1}^{q}\sum_{n=1}^{N} E_{Q_n}\left[-\Delta\cdot h\left(\mathbf{c}^i\cdot\mathbf{x}_n + d^i\right) + y_n^i \log h\left(\mathbf{c}^i\cdot\mathbf{x}_n + d^i\right)\right] + C.\end{aligned} \tag{19}$$

First, let us instead examine the following more general problem: maximize the objective function $E_{\mathbf{x}}[g(\mathbf{c}\cdot\mathbf{x}+d)]$ with respect to $\mathbf{c}$ and d, where g is concave and $\mathbf{x}$ is gaussian distributed with mean $\boldsymbol{\xi}$ and covariance Ψ. Defining the new variables $\tilde{\mathbf{c}} = [\mathbf{c}'\ d]'$ and $\tilde{\mathbf{x}} = [\mathbf{x}'\ 1]'$, the objective function can be equivalently expressed as

$$\begin{aligned}O &= E_{\mathbf{x}}[g\left(\mathbf{c}'\mathbf{x}+d\right)]\\ &= \int_{\mathbf{x}} g\left(\tilde{\mathbf{c}}'\tilde{\mathbf{x}}\right)\mathcal{N}_{\mathbf{x}}(\tilde{\boldsymbol{\xi}},\tilde{\Psi})\,d\mathbf{x}\\ &= \int_{z} g(z)\,\mathcal{N}_z(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}},\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}})\,dz,\end{aligned} \tag{20}$$

where $\tilde{\boldsymbol{\xi}} = [\boldsymbol{\xi}'\ 1]'$ and $\tilde{\Psi}$ is a matrix in $\mathbb{R}^{(p+1)\times(p+1)}$ with the upper-left $p \times p$ sub-matrix equal to Ψ and the rest of the elements set to zero. This objective function can be maximized using Newton's method since it is concave in $\tilde{\mathbf{c}}$. However, to perform this optimization method, we require the gradient and the hessian of O. The gradient can be obtained as shown below.

$$\begin{aligned}\frac{\partial O}{\partial\tilde{\mathbf{c}}} &= \frac{\partial}{\partial\tilde{\mathbf{c}}}\int_z g(z)\,\mathcal{N}_z(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}},\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}})\,dz\\ &= \int_z g(z)\,\frac{\partial}{\partial\tilde{\mathbf{c}}}\mathcal{N}_z(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}},\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}})\,dz\\ &= \int_z g(z)\,\frac{\partial}{\partial\tilde{\mathbf{c}}}\left[\frac{1}{\sqrt{2\pi\cdot\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}\exp\left(-\frac{\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{2\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right)\right]dz\end{aligned} \tag{21}$$

Taking the partial derivative in (21) requires the following quantities.

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\frac{1}{\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}\right) = \frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^{-\frac{1}{2}} = -\frac{1}{2}\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^{-\frac{3}{2}}2\tilde{\Psi}\tilde{\mathbf{c}} = -\frac{\tilde{\Psi}\tilde{\mathbf{c}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\frac{1}{\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}$$

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\exp\left(-\frac{\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{2\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right) = \exp\left(-\frac{\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{2\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right)\frac{\partial}{\partial\tilde{\mathbf{c}}}\left(-\frac{\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{2\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right)$$

$$\begin{aligned}\frac{\partial}{\partial\tilde{\mathbf{c}}}\left(-\frac{\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{2\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right) &= -\frac{2\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)\left(-\tilde{\boldsymbol{\xi}}\right)2\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}} - 4\tilde{\Psi}\tilde{\mathbf{c}}\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{4\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^2}\\ &= -\frac{-\tilde{\boldsymbol{\xi}}(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}) - \tilde{\Psi}\tilde{\mathbf{c}}\frac{\left(z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}} = \frac{z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\left(\tilde{\boldsymbol{\xi}} + \frac{z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}}\right)\end{aligned}$$

Using the equations above, we can reduce (21) to

$$\frac{\partial O}{\partial\tilde{\mathbf{c}}} = \int_z g(z)\left[-\frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}} + \frac{z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\left(\tilde{\boldsymbol{\xi}} + \frac{z-\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}}\right)\right]\mathcal{N}_z(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}},\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}})\,dz. \tag{22}$$

While there does not exist a convenient analytic solution to the above integral, it can be accurately and reasonably efficiently approximated using gaussian quadrature [2], [3]. Specifically, gaussian quadrature rules state that

$$\int_z f(z)\mathcal{N}_z(\mu,\sigma^2)\,dz \approx \sum_{j=1}^{J} w_j f(Z_j) \quad \text{for } Z_j = \mu + \gamma_j \sigma, \tag{23}$$

for any function f, where w_j are the quadrature weights and γ_j are the normalized quadrature points.

Identifying the function f in (22) and substituting $z = Z_j = \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}} + \gamma_j\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}$, the quadrature function for the gradient is

$$\begin{aligned} f_1(\gamma_j) &= g(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}} + \gamma_j\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}})\left[-\frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}} + \frac{\gamma_j}{\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}\tilde{\boldsymbol{\xi}} + \frac{\gamma_j^2}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}}\right] \\ &= g(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}} + \gamma_j\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}})\frac{1}{\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}\left[\gamma_j\tilde{\boldsymbol{\xi}} + \frac{1}{\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}\left(\gamma_j^2 - 1\right)\tilde{\Psi}\tilde{\mathbf{c}}\right]. \end{aligned} \tag{24}$$

Likewise, the same procedure must be performed to find the hessian of the objective function:

$$\frac{\partial^2 O}{\partial\tilde{\mathbf{c}}^2} = \int_z g(z)\frac{\partial}{\partial\tilde{\mathbf{c}}}\left[\mathcal{N}_z(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}, \tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}})\left[-\frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}} + \frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\left(\tilde{\boldsymbol{\xi}} + \frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}}\right)\right]'\right] dz. \tag{25}$$

The following quantities will be useful.

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}\tilde{\mathbf{c}}\right)' = \frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^{-1}\tilde{\mathbf{c}}'\tilde{\Psi}\right) = -\frac{2}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^2}\tilde{\Psi}\tilde{\mathbf{c}}\tilde{\mathbf{c}}'\tilde{\Psi} + \frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi}$$

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\boldsymbol{\xi}'\right) = \frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^{-1}\left(z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)\boldsymbol{\xi}'\right) = -\frac{2}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^2}\tilde{\Psi}\tilde{\mathbf{c}}\left(z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)\boldsymbol{\xi}' - \frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}'$$

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\left(\frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right)^2\tilde{\mathbf{c}}'\tilde{\Psi}\right) = \left(\frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right)^2\tilde{\Psi} + \frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\left(\frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right)^2\right)\tilde{\mathbf{c}}'\tilde{\Psi}$$

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\left(\left(\frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\right)^2\right) = 2\frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\left(-\frac{2\left(z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^2}\tilde{\Psi}\tilde{\mathbf{c}} - \frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\boldsymbol{\xi}}\right) = -4\frac{\left(z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}\right)^2}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^3}\tilde{\Psi}\tilde{\mathbf{c}} - 2\frac{z - \tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^2}\tilde{\boldsymbol{\xi}}$$

Next define a function $\mathbf{a}(\cdot)$ as

$$\mathbf{a}(\gamma_j) = \frac{1}{\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}\left[\gamma_j\tilde{\boldsymbol{\xi}} + \frac{1}{\sqrt{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}}\left(\gamma_j^2 - 1\right)\tilde{\Psi}\tilde{\mathbf{c}}\right]. \tag{26}$$

Substituting these expressions into (25), the quadrature function for the hessian is:

$$\begin{aligned} f_2(\gamma_j) = g(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}} + \gamma_j\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}})\Bigg[&\mathbf{a}(\gamma_j)\mathbf{a}(\gamma_j)' + \frac{2}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^2}\tilde{\Psi}\tilde{\mathbf{c}}\tilde{\mathbf{c}}'\tilde{\Psi} - \frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi} - \frac{2\gamma_j}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^{\frac{3}{2}}}\tilde{\Psi}\tilde{\mathbf{c}}\boldsymbol{\xi}' \\ &- \frac{1}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\boldsymbol{\xi}}\tilde{\boldsymbol{\xi}}' + \frac{\gamma_j^2}{\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}}\tilde{\Psi} - \frac{4\gamma_j^2}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^2}\tilde{\Psi}\tilde{\mathbf{c}}\tilde{\mathbf{c}}'\tilde{\Psi} - \frac{2\gamma_j}{\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)^{\frac{3}{2}}}\boldsymbol{\xi}\tilde{\mathbf{c}}'\tilde{\Psi}\Bigg]. \end{aligned} \tag{27}$$

For certain choices of h it is possible to compute the gradient and hessian analytically. To illustrate, we start with the following form

$$\int_z g(z)\mathcal{N}_z(\mu,\sigma^2)\,dz = \frac{1}{\sqrt{2\pi\sigma^2}}\int_z g(z)\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)dz \tag{28}$$

for $g(z) = -\exp(z)$. The classic method to solve this integral is to "complete the squares" in the exponent.

$$\begin{aligned}
&\int_z -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\left(\frac{(z-\mu)^2}{2\sigma^2}-z\right)\right)dz \\
&= \int_z -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{z^2-2\mu z+\mu^2-2\sigma^2 z}{2\sigma^2}\right)dz \\
&= \int_z -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{z^2-\left(2\mu+2\sigma^2\right)z+\mu^2}{2\sigma^2}\right)dz \\
&= \int_z -\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{z^2-2\left(\mu+\sigma^2\right)z+\mu^2+2\sigma^2\mu+\sigma^4-\mu^2-2\sigma^2\mu-\sigma^4+\mu^2}{2\sigma^2}\right)dz \\
&= -\exp\left(\mu+\frac{1}{2}\sigma^2\right)\int_z \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(z-(\mu+\sigma)\right)^2}{2\sigma^2}\right)dz \\
&= -\exp\left(\mu+\frac{1}{2}\sigma^2\right)\int_z \mathcal{N}\left(\mu+\sigma,\sigma^2\right)dz \\
&= -\exp\left(\mu+\frac{1}{2}\sigma^2\right) \qquad (29)
\end{aligned}$$

Relating this form back to (20), the gradient with respect to $\tilde{\mathbf{c}}$ is

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\left[-\exp\left(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}+\frac{1}{2}\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)\right)\right] = -\exp\left(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}+\frac{1}{2}\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)\right)\left(\tilde{\boldsymbol{\xi}}+\tilde{\Psi}\tilde{\mathbf{c}}\right). \tag{30}$$

Likewise, the hessian can be computed as follows.

$$\frac{\partial}{\partial\tilde{\mathbf{c}}}\left[-\exp\left(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}+\frac{1}{2}\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)\right)\left(\tilde{\boldsymbol{\xi}}+\tilde{\Psi}\tilde{\mathbf{c}}\right)'\right] = -\exp\left(\tilde{\mathbf{c}}'\tilde{\boldsymbol{\xi}}+\frac{1}{2}\left(\tilde{\mathbf{c}}'\tilde{\Psi}\tilde{\mathbf{c}}\right)\right)\left(\left(\tilde{\boldsymbol{\xi}}+\tilde{\Psi}\tilde{\mathbf{c}}\right)\left(\tilde{\boldsymbol{\xi}}+\tilde{\Psi}\tilde{\mathbf{c}}\right)'+\tilde{\Psi}\right) \tag{31}$$

For another choice of $g(x)$, $\tilde{\mathbf{c}}'\boldsymbol{\xi}$, the gradient is trivially $\boldsymbol{\xi}$ and the hessian is the zero matrix.

4 Inference

Once the model parameters have been chosen, the generative model can be used to make inferences on the training data or new observations. For the training data, the hidden state vector $\mathbf{x}$ is the only variable that must be inferred. The posterior distribution of $\mathbf{x}$ can be approximated by a gaussian, exactly as described previously. This results in a distribution Q with mean $\boldsymbol{\xi}$ and covariance Ψ. Therefore, the maximum *a posteriori* estimate estimate of $\mathbf{x}$ is simply $\boldsymbol{\xi}$.

When performing inference for a new observation, the mixture component identification, s, is assumed to be unknown. The posterior distributions of both s and $\mathbf{x}$, given the data, $\mathbf{y}$, are potentially of interest. The first of these distributions can be expressed as follows:

$$\begin{aligned}
P\left(s\mid\mathbf{y},\hat{\theta}\right) &\propto P\left(\mathbf{y}\mid s,\hat{\theta}\right)P\left(s\mid\hat{\theta}\right) \\
&= \pi_s\left(\int_{\mathbf{x}} P\left(\mathbf{y},\mathbf{x}\mid s,\hat{\theta}\right)d\mathbf{x}\right) \\
&= \pi_s\left(\int_{\mathbf{x}} P\left(\mathbf{y}\mid\mathbf{x},\hat{\theta}\right)P\left(\mathbf{x}\mid s,\hat{\theta}\right)d\mathbf{x}\right) \\
&= \pi_s E_{\mathbf{x}\mid s}\left[P\left(\mathbf{y}\mid\mathbf{x},\hat{\theta}\right)\right]. \qquad (32)
\end{aligned}$$

where the expectation in (32) is of a product of poissons with respect to a gaussian distribution that has mean $\hat{\boldsymbol{\mu}}_s$ and covariance $\hat{\Sigma}_s$. This expectation can be computed using sampling techniques or Laplace's method.

To infer $\mathbf{x}$ given the data, the following derivation applies:

$$\begin{aligned} P\left(\mathbf{x} \mid \mathbf{y}, \hat{\theta}\right) &= \sum_{s=1}^{M} P\left(\mathbf{x} \mid \mathbf{y}, s, \hat{\theta}\right) P\left(s \mid \mathbf{y}, \hat{\theta}\right) \\ &= \sum_{s=1}^{M} P\left(\mathbf{x} \mid \mathbf{y}, s, \hat{\theta}\right) \pi_s E_{\mathbf{x}|s}\left[P\left(\mathbf{y} \mid \mathbf{x}, \hat{\theta}\right)\right]. \end{aligned} \tag{33}$$

References

[1] A.C. Smith and E.N. Brown. Estimating a state-space model from point process observations. *Neural Comput*, 15(5):965–991, 2003.

[2] S.J. Julier and J.K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Proc. AeroSense: 11th Int. Symp. Aerospace/Defense Sensing, Simulation and Controls*, pages 182–193, 1997.

[3] U.N. Lerner. *Hybrid Bayesian networks for reasoning about complex systems*. PhD thesis, Stanford University, Stanford, CA, 2002.