ized by $\rho_i$. Our objective is to recover the set of transformation parameters $\{\rho_i\}_{i=1}^N$, such that the aligned data set $\{y_i = \tau(x_i, \rho_i^{-1})\}_{i=1}^N$ is more coherent. In the process, we also learn a clustering assignment $\{z_i\}_{i=1}^N$ of the data points. Here $\rho_i^{-1}$ is defined as the parameter vector generating the inverse of the transformation that would be generated by the parameter vector $\rho$ (i.e. $x_i = \tau(\tau(x_i, \rho_i^{-1}), \rho_i)$).

2 Bayesian Joint Alignment

The Bayesian alignment (BA) model assumes a unimodal data set (in $\S$ 3 this assumption is relaxed). Consequently there is a single set of parameters ($\theta$ and $\rho$) that generate the entire data set (see Figure 3). Under this model, every observed data item, $x_i$, is generated by transforming a canonical data item, $y_i$, with transformation, $\rho_i$. More formally, $x_i = \tau(y_i, \rho_i)$, where $y_i \sim F_D(\theta)$ and $\rho_i \sim F_T(\varphi)$. The auxiliary variable $y_i$ is not shown in the graphical model for simplicity. Given the Bayesian setting, the parameters $\theta$ and $\varphi$ are random variables, with their respective prior distributions, $H_D(\lambda)$ and $H_T(\alpha)$.\footnote{Here we assume that the hyperparameters $\alpha$ and $\lambda$ are fixed, but they can be learned or sampled if necessary.}

The model does not assume that there exists a single perfect canonical example that explains all the data, but uses a parametric distribution $F_D(\theta)$ to generate a slightly different canonical example, $y_i$, for each data item, $x_i$. This enables it to explain variability in the data set that may not be captured with the transformation function alone. The model treats the transformation function as a black-box operation, making it applicable to a wide range of data types (e.g. curves, images, and 3D MRI scans), as long as an appropriate transformation function is specified.

For both this model and the full joint alignment and clustering model introduced in the next section we use exponential family distributions for $F_D(\theta)$ and $F_T(\varphi)$ and their respective conjugate priors for $H_T(\alpha)$ and $H_D(\lambda)$. This allows us to use Rao-Blackwellized sampling schemes [4] by analytically integrating out the model parameters and caching sufficient statistics for efficient likelihood computations. Furthermore, the hyperparameters now play intuitive roles where they act as a pseudo data set and are easier to set or learn from data.

2.1 Learning

Given a data set $\{x_i\}_{i=1}^N$ we wish to learn the parameters of this model $\{\rho_i\}_{i=1}^N, \theta, \varphi$. We use a Rao-Blackwellized Gibbs sampler that integrates out the model parameters, $\theta$ and $\varphi$, and only samples the hidden variables, $\{\rho_i\}_{i=1}^N$. Such samplers typically speed-up convergence. The intuition is that the model parameters are implicitly updated with the sampling of every transformation parameter instead of once per Gibbs iteration. The resulting Gibbs sampler only iterates over the transformation parameters:

\[
\forall i = 1: N; \quad \rho_i^{(t)} \sim p(\rho_i | x_i, \rho_i^{(t-1)}, \alpha, \lambda) \\
\propto p(x_i | \rho_i, x_{-i}, \rho_i^{(t-1)}, \alpha, \lambda) \\
= p(x_i | \rho_i, x_{-i}, \rho_i^{(t)}, \alpha, \lambda) p(\rho_i | \rho_i^{(t)}, \alpha) \\
= p(y_i | y_i^{(t)}, \lambda) p(\rho_i | \rho_i^{(t)}, \alpha),
\]

where $y_i = \tau(x_i, \rho_i^{-1})$. The $t$ superscript in the above equations refers to the Gibbs iteration number.

Sampling $\rho_i$ is complicated by the fact that $p(y_i | y_i^{(t)}, \lambda)$ depends on the transformation function. Previous alignment research [21, 24], has shown that the gradient of an alignment objective function with respect to the transformations provides a strong indicator for how alignment should proceed. One option would be Hamiltonian Monte Carlo sampling [26] which uses the gradient as a drift factor to influence sampling. However, instead of relying on direct sampling techniques, we use approximations based on the posterior mode [13]. Such an approach is more direct since it is expected that the distribution will be tightly concentrated around the mode. Thus, at each iteration the transformation parameter is updated as follows:

\[
\rho_i = \arg \max_{\rho_i} p(y_i | y_i^{(t)}, \lambda) p(\rho_i | \rho_i^{(t)}, \alpha).
\]

Interestingly, the same learning scheme can be derived using the incremental variant [27] of hard-EM.

2.2 Model Characteristics

The objective function optimized in our model contains two key terms, a data term, $p(x_i | \rho, \theta)$, and a transformation term, $p(\rho | \varphi)$. The latter acts as a regularizer to penalize large transformations and prevent the data from being annihilated. One advantage of our model is that large transformations are penalized in a principled fashion. More specifically, the cost of a transformation, $\rho_i$ is based on the

![Figure 3: Graphical representation for our proposed Bayesian alignment model ($\S$ 2).](image-url)