A Unified Theory of Exact Inference and Learning in Exponential Family Latent Variable Models

Sacha Sokoloski sacha.sokoloski@uni-tuebingen.de
Hertie Institute for AI in Brain Health
University of Tübingen

Editor:

Abstract

Bayes' rule describes how to infer posterior beliefs about latent variables given observations, and inference is a critical step in learning algorithms for latent variable models (LVMs). Although there are exact algorithms for inference and learning for certain LVMs such as linear Gaussian models and mixture models, researchers must typically develop approximate inference and learning algorithms when applying novel LVMs. In this paper we study the line that separates LVMs that rely on approximation schemes from those that do not, and develop a general theory of exponential family, latent variable models for which inference and learning may be implemented exactly. Firstly, under mild assumptions about the exponential family form of a given LVM, we derive necessary and sufficient conditions under which the LVM prior is in the same exponential family as its posterior, such that the prior is conjugate to the posterior. We show that all models that satisfy these conditions are constrained forms of a particular class of exponential family graphical model. We then derive general inference and learning algorithms, and demonstrate them on a variety of example models. Finally, we show how to compose our models into graphical models that retain tractable inference and learning. In addition to our theoretical work, we have implemented our algorithms in a collection of libraries with which we provide numerous demonstrations of our theory, and with which researchers may apply our theory in novel statistical settings.

Keywords: Bayesian Inference, Bayesian Smoothing, Expectation-Maximization, Exponential Families

1 Introduction

Latent variable models (LVMs) describe data in terms of observable random variables that we measure directly, and latent random variables that we do not measure, but that help explain our observations. The probabilistic formulation of LVMs affords general strategies for two fundamental operations when applying them: (i) Bayes' rule describes how to infer posterior beliefs about latent variables given prior beliefs and observations; and (ii) the principle of maximum likelihood — or equivalently minimal cross-entropy — reduces learning to minimizing the cross-entropy of the LVM parameters given the data (Roweis and Ghahramani, 1999). Because LVMs can often be scaled to achieve arbitrary model complexity, the central limitation of LVMs is rarely whether they are sufficiently powerful to model a given dataset, but rather whether inference and learning for a given LVM are computationally tractable.

There are numerous techniques for approximate inference and learning with LVMs. Variational methods are arguably the most general framework (Neal and Hinton, 1998; Wainwright and Jordan, 2008; Kingma and Welling, 2013), and approaches outside of the classic variational framework include contrastive divergence (Hinton, 2002) and generative adversarial networks (Goodfellow et al., 2014; Radford et al., 2015). These techniques have been used to develop powerful models of hierarchical (Hinton et al., 2006; Salakhutdinov and Hinton, 2012; Vértes and Sahani, 2018) and dynamic latent structure (Taylor et al., 2011; Boulanger-Lewandowski et al., 2012; Durstewitz, 2017), which have in turn been applied to modelling biological neural circuits (Beck et al., 2012), cognitive modelling (Salakhutdinov et al., 2013; Lake et al., 2015), and image synthesis (Ho et al., 2020).

Nevertheless, there are well-known LVMs for which inference and learning can be implemented exactly — by exact inference and learning we mean, more or less, that there is a closed-form expression for the posterior over the latent variables, and that we may minimize the cross-entropy of the model parameters directly, rather than via a lower-bound that can introduce approximation errors (Shekhovtsov et al., 2021). For example, there are closed-form expressions for the posteriors of mixture models and linear Gaussian models (which includes principle component analysis and factor analysis as special cases), and exact implementations of expectation-maximization (EM) for training them (see Bishop, 2006).

In this paper we study the boundary that separates models that require approximation techniques for inference and learning from those that do not, and derive a general theory of exact learning and inference for a broad class of exponential family LVMs. We begin our approach by analyzing the conditions under which priors and posteriors over latent variables share the same exponential family form. In general, priors are known as conjugate priors when they have the same parametric form as the posterior (Diaconis and Ylvisaker, 1979; Arnold et al., 1993), and conjugate priors are widely-applied to inferring posteriors over the parameters of exponential family models. In this study we generalize this notion of conjugacy to LVMs with priors and posteriors over arbitrary latent variables.

We show that if the likelihood of the observations given the latent variables is exponential family distributed, then the latent variable prior and posterior are conjugate if and only if (i) the LVM has a particular exponential-family form, and (ii) the LVM parameters satisfy a constraint. The exponential-family form in question has a long history (Besag, 1974; Arnold and Press, 1989; Arnold et al., 2001; Yang et al., 2015; Tansey et al., 2015), and models with this form have been variously referred to as conditionally specified distributions (Arnold and Press, 1989), vector space Markov random fields (Tansey et al., 2015), and exponential family harmoniums (Smolensky, 1986; Welling et al., 2005) — in this paper we refer to them simply as harmoniums. We then show that for harmoniums, the prior is conjugate to the posterior if and only if the likelihood satisfies an equation on its exponential family parameters. We refer to harmoniums that satisfy this equation as conjugated harmoniums.

We show that both mixture models and linear Gaussian models are forms of conjugated harmonium. Outside of these well-known cases, we also study harmoniums defined variously in terms of von Mises distributions, products of Poisson distributions, and Boltzmann machines, and show that they are conjugated under certain conditions. Finally, we show how to use our theory to construct graphical models out of harmonium components that retain tractable inference and learning.

Inference and learning algorithms for LVMs are often developed on a case-by-case basis, yet we show that many algorithms are special cases of a set of general algorithms for inference and learning with conjugated harmoniums. This not only facilitates theoretical unification, but also simplifies the implementation of learning and inference algorithms, as it allows programmers to implement one set of algorithms for a wide variety of cases. Indeed, we have developed a collection of Haskell libraries for numerical optimization based on our theory, and all the models that we demonstrate in this paper were implemented with these libraries (`https://github.com/alex404/goal`).

2 Background

In this section we present the necessary background for our theory of exponential family, latent variable models. We begin with a brief introduction to exponential families, followed by a review of the theory of exponential family harmoniums, and conclude with how to extend harmoniums with graphical model structure. We rely on several notational conventions to help convey our theory, and although we introduce this notation over the course of the text, we summarize the key features here.

In general, we use Latin letters (e.g. x or z) for observations and latent states, and Greek letters (e.g. θ or η) for model parameters. We use lowercase letters (e.g. x or θ) to denote scalars, bold, lowercase letters (e.g. $\mathbf{x}$ or $\boldsymbol{\theta}$) to denote vectors, and bold, capital letters (e.g. $\mathbf{X}$ or $\boldsymbol{\Theta}$) to denote matrices and multilinear transformations — when the form of the variable (e.g. scalar or vector) is not specified, we default to non-bold, lowercase letters. We use capital, italic letters (e.g. X and Z) to denote random variables, regardless if they are scalars, vectors, or matrices. Finally, we use calligraphic letters in the latter part of the Latin alphabet (e.g. $\mathcal{X}$ or $\mathcal{Z}$) to denote the sample space of a random variable, and capital Greek letters (e.g. Θ or H) to denote parameter spaces. Some exceptions to these patterns include, that we use P and p to denote probability distributions and densities of random variables, respectively, Q and q to denote model distributions and densities, respectively, $\mathcal{M}$ to denote statistical models. We may denote functions by either Latin or Greek letters depending on their role.

To indicate that a mathematical object is related to a particular random variable or sample space, we subscript it with capital, italic letters. For example, P_X and p_X denote the probability distribution and density of X, respectively. Similarly, $\mathcal{M}_X$ is a model on the sample space $\mathcal{X}$ of some random variable X of interest, parameterized e.g. by the space Θ_X, with contains elements $\boldsymbol{\theta}_X \in \Theta_X$. We denote the ith element of a vector $\boldsymbol{\theta}$ by θ_i, or e.g. of the vector $\boldsymbol{\theta}_X$ by $\theta_{X,i}$. We denote the ith row or jth column of $\boldsymbol{\Theta}$ by $\boldsymbol{\theta}_i$ or $\boldsymbol{\theta}_j$, respectively, and always state whether we are considering a row or column of the given matrix. When referring to the jth element of a vector $\boldsymbol{\theta}_i$ indexed by i, we write θ_{ij}. Finally, when indexing data points from a sample/dataset, or parameters that are tied to particular observations, we use parenthesized, superscript letters, e.g. $x^{(i)}$, or $\boldsymbol{\theta}_X^{(i)}$.

2.1 Exponential Families

For a thorough development of exponential family theory see Amari and Nagaoka (2007), and Wainwright and Jordan (2008).

Consider a random variable X on the sample space $\mathcal{X}$ with unknown distribution P_X, and suppose $X^{(1)}, \ldots, X^{(n)}$ is an independent and identically distributed sample from P_X. One strategy for modelling P_X based on the sample is to first define a "sufficient" statistic $\mathbf{s}_X \colon \mathcal{X} \to \mathrm{H}_X$ that captures features of interest about X. We then look for a probability distribution Q_X whose expectation of the sufficient statistic $\mathbb{E}_Q[\mathbf{s}_X(X)]$ matches the average of the sufficient statistic over the data $\frac{1}{n}\sum_{i=1}^n \mathbf{s}_X(X^{(i)})$, where the expected value of $f(X)$ under Q_X is defined by $\mathbb{E}_Q[f(X)] = \int_{\mathcal{X}} f dQ_X$. To further constrain the space of possible distributions we also assume that Q_X maximizes the entropy $E_Q[-\log q_X]$, where $q_X = dQ_X/d\mu_X$ is the density function (Radon-Nikodym derivative) of Q_X with respect to some base measure μ_X. Based on these assumptions, it can be shown that q_X must have the exponential family form

$$\log q_X(x) = \mathbf{s}_X(x) \cdot \boldsymbol{\theta}_X - \psi_X(\boldsymbol{\theta}_X), \tag{1}$$

where $\boldsymbol{\theta}_X$ are the natural parameters, and $\psi_X(\boldsymbol{\theta}_X) = \log \int_{\mathcal{X}} e^{\mathbf{s}_X(x)\cdot\boldsymbol{\theta}_X} d\mu_X(x)$ is the log-partition function. We also refer to $\boldsymbol{\eta}_X = \mathbb{E}_Q[\mathbf{s}_X(X)]$ as the mean parameters of q_X.

In general, $\partial_{\boldsymbol{\theta}_X}\psi_X(\boldsymbol{\theta}_X) = \mathbb{E}_Q[\mathbf{s}_X(X)] = \boldsymbol{\eta}_X$, so that we may easily identify the mean parameters $\boldsymbol{\eta}_X$ of q_X given its natural parameters $\boldsymbol{\theta}_X$. However, the natural parameters $\boldsymbol{\theta}_X$ for a given density q_X might not be unique, in the sense that different natural parameters might yield the same density function as defined by Equation 1. To address this we may further assume that the sufficient statistic $\mathbf{s}_X$ is minimal, in the sense that the component functions $\{s_{X,i}\}_{i=1}^{d_X}$ are non-constant and linearly independent, where d_X is the dimension of H_X. If the sufficient statistic $\mathbf{s}_X$ is minimal, then each $\boldsymbol{\theta}_X$ determines a unique density q_X, and $\partial_{\boldsymbol{\theta}_X}\psi_X$ is invertible.

A d_X-dimensional exponential family $\mathcal{M}_X$ is thus a manifold of probability densities defined by a sufficient statistic $\mathbf{s}_X$ and a base measure μ_X. The densities in $\mathcal{M}_X$ can be identified by their mean parameters $\boldsymbol{\eta}_X$ or natural parameters $\boldsymbol{\theta}_X$, and we denote the space of all mean and natural parameters by H_X and Θ_X respectively. A minimal exponential family $\mathcal{M}_X$ is an exponential family with a minimal sufficient statistic $\mathbf{s}_X$, and the parameter spaces Θ_X and H_X of a minimal exponential family are isomorphic. In this case, the transition functions between the two parameter spaces, known as the forward mapping $\boldsymbol{\tau}_X \colon \Theta_X \to \mathrm{H}_X$ and the backward mapping $\boldsymbol{\tau}_X^{-1} \colon \mathrm{H}_X \to \Theta_X$, are given by the gradient $\boldsymbol{\tau}_X(\boldsymbol{\theta}_X) = \partial_{\boldsymbol{\theta}_X}\psi_X(\boldsymbol{\theta}_X)$ and its inverse, respectively.

2.2 Exponential Family Harmoniums

An exponential family harmonium is a kind of product exponential family which includes various LVMs as special cases (Smolensky, 1986; Welling et al., 2005). Given two exponential families $\mathcal{M}_X$ and $\mathcal{M}_Z$ that model the distributions of X and Z, respectively, a harmonium $\mathcal{H}_{XZ}$ is an exponential family that models the joint distribution of X and Z. Where $\mathcal{M}_X$ and $\mathcal{M}_Z$ have base measures μ_X and μ_Z, and sufficient statistics $\mathbf{s}_X$ and $\mathbf{s}_Z$, respectively, the base measure of $\mathcal{H}_{XZ}$ is $\mu_{XZ} = \mu_X \cdot \mu_Z$, and its sufficient statistic is given by $\mathbf{s}_{XZ}(x, z) = (\mathbf{s}_X(x), \mathbf{s}_Z(z), \mathbf{s}_X(x) \otimes \mathbf{s}_Z(z))$, where $\otimes$ is the outer product operator. In words, $\mathbf{s}_{XZ}$ is the concatenation of all the component functions of $\mathbf{s}_X$, $\mathbf{s}_Z$, and $\mathbf{s}_X \otimes \mathbf{s}_Z$. We also note that $\mathcal{H}_{XZ}$ is minimal if both $\mathcal{M}_X$ and $\mathcal{M}_Z$ are minimal. More intuitively, $\mathcal{H}_{XZ}$ is the exponential

family that comprises all densities q_{XZ} with the form

$$\log q_{XZ}(x,z) = \mathbf{s}_X(x) \cdot \boldsymbol{\theta}_X + \mathbf{s}_Z(z) \cdot \boldsymbol{\theta}_Z + \mathbf{s}_X(x) \cdot \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z) - \psi_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ}), \quad (2)$$

where $\boldsymbol{\theta}_{XZ} = (\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})$ are the natural parameters of q_{XZ}, and ψ_{XZ} is the log-partition function of $\mathcal{H}_{XZ}$. We refer to the parameters $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_Z$ as biases, and $\boldsymbol{\Theta}_{XZ}$ as interactions.

Although harmoniums can certainly model the joint distribution of two observable variables, we focus on the setting where one of the variables is latent, and so it will prove helpful to begin developing the language of LVMs. From here on out we will assume that the variables X and Z denote observable and latent random variables, respectively, and we refer to e.g. $\mathbf{s}_X$ and $\mathbf{s}_Z$ as the observable and latent sufficient statistics, and $\boldsymbol{\theta}_X$ and $\boldsymbol{\theta}_Z$ as the observable and latent biases, respectively.

Harmonium densities have a simple log-linear structure, but their marginal densities do not. In particular, for $q_{XZ} \in \mathcal{H}_{XZ}$ with natural parameters $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_Z$, and $\boldsymbol{\Theta}_{XZ}$, it is easy to show that the observable density is given by

$$\log q_X(x) = \mathbf{s}_X(x) \cdot \boldsymbol{\theta}_X + \psi_Z(\boldsymbol{\theta}_Z + \mathbf{s}_X(x) \cdot \boldsymbol{\Theta}_{XZ}) - \psi_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ}), \quad (3)$$

and similarly the prior is given by

$$\log q_Z(z) = \mathbf{s}_Z(z) \cdot \boldsymbol{\theta}_Z + \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)) - \psi_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ}). \quad (4)$$

In contrast, the conditional densities of a harmonium do inherit a linear structure, and have the form of generalized linear models (Bishop, 2006; Yang et al., 2012). In particular, where the likelihood and posterior are defined by $q_{X|Z} = \frac{q_{XZ}}{q_Z}$ and $q_{Z|X} = \frac{q_{XZ}}{q_X}$, respectively, we may combine Equations 2, 3, and 4, to conclude that

$$\log q_{X|Z}(x \mid z) = \mathbf{s}_X(x) \cdot (\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)) - \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)), \quad (5)$$

and similarly that

$$\log q_{Z|X}(z \mid x) = \mathbf{s}_Z(z) \cdot (\boldsymbol{\theta}_Z + \mathbf{s}_X(x) \cdot \boldsymbol{\Theta}_{XZ}) - \psi_Z(\boldsymbol{\theta}_Z + \mathbf{s}_X(x) \cdot \boldsymbol{\Theta}_{XZ}). \quad (6)$$

When discussing conditional densities we denote e.g. the likelihood $q_{X|Z}$ at z by $q_{X|Z=z}$, so that $q_{X|Z=z} \in \mathcal{M}_X$ is the exponential family density with natural parameters $\boldsymbol{\theta}_{X|Z}(z) = \boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)$. We also write e.g. $q_{X|Z} \in \mathcal{M}_X$ as short hand for $q_{X|Z=z} \in \mathcal{M}_X, \forall z \in \mathcal{Z}$, to express that $q_{X|Z}$ is always a member of $\mathcal{M}_X$ for any z.

We defined harmoniums constructively as a product of component exponential families, but we may also define them intrinsically, as the most general families of densities with likelihoods and posteriors in pre-specified exponential families.

Theorem 1 *Suppose that $\mathcal{H}_{XZ}$ is the harmonium defined by the minimal exponential families $\mathcal{M}_X$ and $\mathcal{M}_Z$, and that q_{XZ} is an arbitrary joint density over the sample space $\mathcal{X} \times \mathcal{Z}$ with respect to the product measure $\mu_{XZ} = \mu_X \cdot \mu_Z$. Then $q_{X|Z} \in \mathcal{M}_X$ and $q_{Z|X} \in \mathcal{M}_Z$ if and only if $q_{XZ} \in \mathcal{H}_{XZ}$.*

Proof *See Arnold et al. (2001), Theorem 3.*

■

The leftward implication $\Longleftarrow$ of this theorem is a trivial consequence of Equations 5 and 6, but the rightward implication $\Longrightarrow$ is rather profound. We effectively suppose that q_{XZ} has a likelihood and a posterior with exponential family forms $q_{X|Z}(x \mid z) \propto e^{\mathbf{s}_X(x)\cdot\boldsymbol{\theta}_{X|Z}(z)}$ and $q_{Z|X}(z \mid x) \propto e^{\mathbf{s}_Z(z)\cdot\boldsymbol{\theta}_{Z|X}(x)}$ for arbitrary functions $\boldsymbol{\theta}_{X|Z}\colon \mathcal{Z} \to \Theta_X$ and $\boldsymbol{\theta}_{Z|X}\colon \mathcal{X} \to \Theta_Z$, respectively. Therefore, according to Theorem 1, the mere assumption that q_{XZ} exists is enough to ensure that $\boldsymbol{\theta}_{X|Z}$ and $\boldsymbol{\theta}_{Z|X}$ must have the linear expressions $\boldsymbol{\theta}_{X|Z}(z) = \boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)$ and $\boldsymbol{\theta}_{Z|X}(x) = \boldsymbol{\theta}_Z + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_X(x)$, respectively.

When developing an LVM for a particular statistical problem, there is often an intuitive choice of exponential family structure for the posterior and likelihood. For example, if we wish to cluster count data, it is natural to assume that the likelihood should be Poisson distributed, and the posterior should be a categorically distributed. Based on Theorem 1, this simple set of assumptions is sufficient to ensure that the proposed LVM is a harmonium.

2.3 Harmonium Graphical Models

More recent work has shown how to generalize the intrinsic characterization of harmoniums to Markov Random Fields (MRFs), which model multiple random variables with dependencies represented by a graph (Yang et al., 2015; Tansey et al., 2015). For now we forego the distinction between observable and latent variables, and consider the collection of random variables $X = X_1, \ldots, X_n$ with sample spaces $\mathcal{X}_1, \ldots, \mathcal{X}_n$, respectively. An MRF represents the dependencies between $X_1, \ldots, X_n$ with a graph $G = (V, E)$, where the vertices $V = \{1, \ldots, n\}$ index the random variables, and the edges E represent local dependencies between variables. A clique is a set of vertices in which every pair of vertices is connected by an edge, and the joint density q_X of an MRF can be factorized over its cliques so that

$$q_X(\mathbf{x}) \propto \prod_{C \in \mathcal{C}(G)} f_C(\mathbf{x}_C), \tag{7}$$

where $\mathbf{x}_C = \{x_i \colon i \in C\}$ is the set of variables indexed by the clique C, f_C are the clique potentials, and $\mathcal{C}(G)$ is the set of all cliques on the graph G.[1]

For each clique C, let $c_i \in C$ denote the ith index in C in ascending order. Given a set of exponential families $\mathcal{M}_{X_1}, \ldots, \mathcal{M}_{X_n}$ on the sample spaces $\mathcal{X}_1, \ldots, \mathcal{X}_n$, we define a harmonium graphical model (HGM) over $X_1, \ldots, X_n$ as the exponential family $\mathcal{H}_X^G$ with base measure $\mu_X = \mu_{X_i} \cdots \mu_{X_n}$ and sufficient statistic $\mathbf{s}_X^G = (\bigotimes_{i=1}^{|C|} \mathbf{s}_{X_{c_i}})_{C \in \mathcal{C}(G)}$ — in words, the flattened outer product between every random variable indexed by the given clique, concatenated over all cliques. More intuitively, $\mathcal{H}_X^G$ contains all densities of the form

$$\log q_X(\mathbf{x}) \propto \prod_{C \in \mathcal{C}(G)} e^{\boldsymbol{\Theta}_X^C\left(\left(\mathbf{s}_{X_{c_i}}(x_{c_i})\right)_{i=1}^{|C|}\right)}, \tag{8}$$

where each $\boldsymbol{\Theta}_X^C\colon \mathrm{H}_{X_{c_1}} \times \cdots \times \mathrm{H}_{X_{c_{|C|}}} \to \mathbb{R}$ is a multilinear function of the sufficient statistics $\mathbf{s}_{X_{c_i}}$ for every $c_i \in C$, that represents the interactions between the random variables $X_C = \{X_i \colon i \in C\}$ indexed by C. We refer to $\boldsymbol{\Theta}_X^C$ as the clique-wise interaction map, and note that it is proportional to the logarithm of the clique potential f_C.

1. Note that we specifically assume that $\mathcal{C}(G)$ includes all cliques, not only the maximal ones.

To express the conditional densities of an HGM $\mathcal{H}^G_X$, we first note that the conditional density of any X_i depends only on its neighbours, namely the random variables indexed by $N(i) = \{j \in V \colon \{i,j\} \in E\}$. The local structure of these dependencies is determined by the cliques that contain i, which we denote by $\mathcal{C}_i(G) = \{C \in \mathcal{C}(G) \colon i \in C\}$. Given these definitions, we may express the conditional density of each X_i as

$$q_{X_i|X_{N(i)}}(x_i \mid \mathbf{x}_{N(i)}) \propto \prod_{C \in \mathcal{C}_i(G)} e^{\mathbf{\Theta}^C_X \left(\left(\mathbf{s}_{X_{c_j}}(x_{c_j}) \right)^{|C|}_{j=1, c_j \neq i} \right)}. \tag{9}$$

In this equation observe that we apply $\mathbf{\Theta}^C_X$ to the sufficient statistics of every x_{c_j} indexed by C except for x_i. The omission of $\mathbf{s}_{X_i}(x_i)$ from the arguments of $\mathbf{\Theta}^C_X$ ensures that the return value of $\mathbf{\Theta}^C_X \left(\left(\mathbf{s}_{X_{c_j}}(x_{c_j}) \right)^{|C|}_{j=1, c_j \neq i} \right)$ is a vector in the parameter space Θ_{X_i}.

With the definitions of the joint and conditional densities of an HGM in place, we may now express its intrinsic definition.

Theorem 2 *Suppose $\mathcal{H}^G_X$ is the HGM defined by the graph G and the minimal exponential families $\mathcal{M}_{X_1}, \ldots, \mathcal{M}_{X_n}$, and that q_X is an arbitrary joint density over the sample space $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ with respect to the product measure $\mu_X = \mu_{X_1} \cdot \cdots \cdot \mu_{X_n}$. Then the conditional densities $q_{X_i|X_{N(i)}} \in \mathcal{M}_{X_i}$ for all i if and only if $q_X \in \mathcal{H}^G_X$.*

Proof *See Tansey et al. (2015), Theorem 1.* ■

To help us disambiguate HGMs from the "standard" harmonium we developed in Section 2.2, we refer to families of densities of the form Equation 2 as bivariate harmoniums when necessary, and note that the observable and latent variables may nevertheless be vector-valued.

Bivariate harmoniums can be formulated as special cases of HGM. In particular, we may derive Equation 2 from Equation 8 by representing a bivariate harmonium $\mathcal{H}^G_{XZ}$ with the graph $G = (\{1,2\}, \{\{1,2\}\})$ that contains two vertices and a single edge between them (Fig. 1**a**). The clique set of G is then $\mathcal{C}(G) = \{\{1\}, \{2\}, \{1,2\}\}$, so that the parameters of a given density $q_{XZ} \in \mathcal{H}^G_{XZ}$ are $\mathbf{\Theta}^{\{1\}}_X$, $\mathbf{\Theta}^{\{2\}}_Z$, and $\mathbf{\Theta}^{\{1,2\}}_{XZ}$, which correspond to the observable biases, the latent biases, and the interaction matrix, respectively. In Figure 1**b** we depict an example HGM with 3rd-order interactions. In Section 3.4.2 we will explore, conversely, how to interpret HGMs as forms of bivariate harmonium over observable and latent variables.

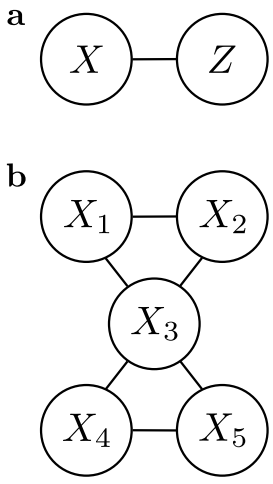

Figure 1: *Graphical models*

3 Results

Having reviewed the mathematical foundations, we now present our unified theory of exact inference and learning in exponential family, latent variable models. We first develop the

fundamental theorems of conjugated harmoniums. We then review several established models that can be formulated as conjugated harmoniums, and continue by showing how to fit conjugated harmoniums to data using of variety of algorithms. Finally, we extend our theory to models with graphical model structure, and show how to compose conjugated harmoniums into graphical models that retain tractable inference and learning.

3.1 Conjugated Harmoniums and Conjugating Likelihoods

The likelihood $q_{X|Z}$ and posterior $q_{Z|X}$ of a harmonium density q_{XZ} have simple linear structures, and are always in the exponential families $\mathcal{M}_X$ and $\mathcal{M}_Z$, respectively. In general, however, the observable density q_X and prior q_Z are not members of $\mathcal{M}_X$ and $\mathcal{M}_Z$, respectively. This is both a blessing and a curse, as on one hand, this allows the set of all observable densities to represent more complex distributions than the simpler set $\mathcal{M}_X$. On the other hand, because the prior may not be computationally tractable, various computations with harmoniums, such as sampling, learning, and inference, may also prove intractable.

3.1.1 Conjugated Harmoniums

Ideally, the set of observable densities would be more complex than $\mathcal{M}_X$ to ensure maximum representational power, while the priors would remain in $\mathcal{M}_Z$ to facilitate tractability. Perhaps surprisingly, some classes of harmoniums do indeed have this structure. In general, a prior and posterior are said to be conjugate if they have the same form. In the context of a harmonium density q_{XZ}, since the harmonium posterior $q_{Z|X} \in \mathcal{M}_Z$ by construction, the harmonium prior and posterior are conjugate if $q_Z \in \mathcal{M}_Z$.

Definition 3 (Conjugated Harmonium) *Where $\mathcal{H}_{XZ}$ is a harmonium defined by $\mathcal{M}_X$ and $\mathcal{M}_Z$, a harmonium density $q_{XZ} \in \mathcal{H}_{XZ}$ is conjugated if $q_Z \in \mathcal{M}_Z$, and the harmonium $\mathcal{H}_{XZ}$ is conjugated if every $q_{XZ} \in \mathcal{H}_{XZ}$ is conjugated.*

Trivially, any $q_{XZ} \in \mathcal{H}_{XZ}$ with parameters $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_Z$, and $\boldsymbol{\Theta}_{XZ}$ is conjugated if the interactions $\boldsymbol{\Theta}_{XZ} = \mathbf{0}$. This follows directly from Equation 4, and corresponds to the case where q_X and q_Z are independent — unfortunately, by the same logic, $q_X \in \mathcal{M}_X$ and so such a model offers no additional representational power. With the following lemma — the Conjugation Lemma — we present a necessary and sufficient condition on the parameters of q_{XZ} that ensures conjugation, yet which allows for q_X not to be in $\mathcal{M}_X$.

Lemma 4 (The Conjugation Lemma) *Suppose that $\mathcal{H}_{XZ}$ is a harmonium defined by the exponential families $\mathcal{M}_X$ and $\mathcal{M}_Z$, and that $q_{XZ} \in \mathcal{H}_{XZ}$ has parameters $(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})$. Then $q_{XZ} \in \mathcal{H}_{XZ}$ is conjugated if and only if there exists a vector $\boldsymbol{\rho}$ and a scalar χ such that*

$$\psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)) = \mathbf{s}_Z(z) \cdot \boldsymbol{\rho} + \chi, \tag{10}$$

for any $z \in \mathcal{Z}$.

Proof *On one hand, if we assume that $q_Z \in \mathcal{M}_Z$ with parameters $\boldsymbol{\theta}_Z^*$, then*

$$\begin{aligned} & q_Z(z) \propto e^{\boldsymbol{\theta}_Z \cdot \mathbf{s}_Z(x) + \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z))} \propto e^{\boldsymbol{\theta}_Z^* \cdot \mathbf{s}_Z(z)} \\ \implies & \quad \boldsymbol{\theta}_Z \cdot \mathbf{s}_Z(z) + \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)) = \boldsymbol{\theta}_Z^* \cdot \mathbf{s}_Z(z) + \chi \\ \implies & \quad \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)) = \mathbf{s}_Z(z) \cdot \boldsymbol{\rho} + \chi. \end{aligned}$$

for some χ*, and* $\boldsymbol{\rho} = \boldsymbol{\theta}^*_Z - \boldsymbol{\theta}_Z$.

On the other hand, if we first assume that Eq. 10 holds, then q_Z *is given by*

$$q_Z(z) \propto e^{\boldsymbol{\theta}_Z \cdot \mathbf{s}_Z(z) + \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z))} \propto e^{(\boldsymbol{\theta}_Z + \boldsymbol{\rho}) \cdot \mathbf{s}_Z(z)}, \tag{11}$$

which implies that $q_Z \in \mathcal{M}_Z$ *with parameters* $\boldsymbol{\theta}_Z + \boldsymbol{\rho}$. ■

Most of the computations and algorithm in our theory of conjugation involve operations on the parameters $\boldsymbol{\rho}$ and χ, and we refer to them as the conjugation parameters. We introduce two of these simpler computations with the following corollaries.

Corollary 5 *Suppose that* $\mathcal{H}_{XZ}$ *is a harmonium defined by the exponential families* $\mathcal{M}_X$ *and* $\mathcal{M}_Z$*, and that* $q_{XZ} \in \mathcal{H}_{XZ}$ *with parameters* $(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})$ *is conjugated with conjugation parameters* $\boldsymbol{\rho}$ *and* χ*. Then the parameters* $\boldsymbol{\theta}^*_Z$ *of* $q_Z \in \mathcal{M}_Z$ *are given by*

$$\boldsymbol{\theta}^*_Z = \boldsymbol{\theta}_Z + \boldsymbol{\rho}. \tag{12}$$

Proof *This follows from the second part of the proof of Lemma 4.* ■

This corollary shows that given the conjugation parameters, marginalizing the observable variables out of a conjugated harmonium density q_{XZ} reduces to vector addition. Assuming we can sample from densities in $\mathcal{M}_X$ and $\mathcal{M}_Z$, Corollary 5 also shows us that we may sample any conjugated $q_{XZ} \in \mathcal{H}_{XZ}$ by first sampling from $q_Z \in \mathcal{M}_Z$, and then $q_{X|Z} \in \mathcal{M}_X$.

Corollary 6 *Suppose that* $\mathcal{H}_{XZ}$ *is a harmonium defined by the exponential families* $\mathcal{M}_X$ *and* $\mathcal{M}_Z$*, that* $q_{XZ} \in \mathcal{H}_{XZ}$ *with parameters* $(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})$ *is conjugated with conjugation parameters* $\boldsymbol{\rho}$ *and* χ*, and that* $q_Z \in \mathcal{M}_Z$ *has parameters* $\boldsymbol{\theta}^*_Z$*. Then the log-partition function* ψ_{XZ} *satisfies*

$$\psi_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ}) = \psi_Z(\boldsymbol{\theta}^*_Z) + \chi, \tag{13}$$

where ψ_Z *is the log-partition function of* $\mathcal{M}_Z$.

Proof *Suppose* $q_{XZ} \in \mathcal{H}_{XZ}$ *is conjugated with conjugation parameters* $\boldsymbol{\rho}$ *and* χ*. Then by substituting Equation 10 into Equation 4,* $q_Z(z) = e^{\boldsymbol{\theta}_Z \cdot \mathbf{s}_Z(z) + \boldsymbol{\rho} \cdot \mathbf{s}_Z(z) + \chi - \psi_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})}$*, and by Corollary 5 we know that* $q_Z(z) = e^{\boldsymbol{\theta}^*_Z \cdot \mathbf{s}_Z(z) - \psi_Z(\boldsymbol{\theta}^*_Z)}$*, where* $\boldsymbol{\theta}^*_Z = \boldsymbol{\theta}_Z + \boldsymbol{\rho}$*. Therefore*

$$\begin{aligned} q_Z(z) &= e^{\boldsymbol{\theta}^*_Z \cdot \mathbf{s}_Z(z) - \psi_Z(\boldsymbol{\theta}^*_Z)} \\ &= e^{\boldsymbol{\theta}_Z \cdot \mathbf{s}_Z(z) + \boldsymbol{\rho} \cdot \mathbf{s}_Z(z) + \chi - \psi_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})} \\ \iff \qquad \psi_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ}) &= (\boldsymbol{\theta}_Z + \boldsymbol{\rho} - \boldsymbol{\theta}^*_Z) \cdot \mathbf{s}_Z(z) + \psi_Z(\boldsymbol{\theta}^*_Z) + \chi \\ &= \psi_Z(\boldsymbol{\theta}^*_Z) + \chi. \end{aligned}$$

■

The log-partition function ψ_Z of $\mathcal{M}_Z$ will often be tractable, which means that various computations that rely on ψ_{XZ}, such as evaluating the observable density q_X (Eq. 3), are also tractable. In particular, given the conjugated harmonium density $q_{XZ} \in \mathcal{H}_{XZ}$ with natural parameters $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_Z$, and $\boldsymbol{\Theta}_{XZ}$, and conjugation parameters $\boldsymbol{\rho}$ and χ, we may combine Equations 3 and 13 to conclude that

$$\log q_X(x) = \mathbf{s}_X(x) \cdot \boldsymbol{\theta}_X + \psi_Z(\boldsymbol{\theta}_Z + \mathbf{s}_X(x) \cdot \boldsymbol{\Theta}_{XZ}) - \psi_Z(\boldsymbol{\theta}_Z + \boldsymbol{\rho}) - \chi. \tag{14}$$

We may use Equation 13 to finesse the problem we laid out at the beginning of the section: on one hand, the observable density q_X of a conjugated harmonium density (Eq. 14) is not generally in $\mathcal{M}_X$, and yet on the other hand it is computable up to the computability of ψ_Z.

3.1.2 Conjugating Likelihoods

The left hand side of Equation 10 is simply the log-partition function of the likelihood $q_{X|Z}$ (Eq. 5) of a harmonium density q_{XZ}, and it can be more natural to interpret conjugation as a property of the likelihood $q_{X|Z}$ rather than the complete joint density q_{XZ}. Indeed, we may generalize the Conjugation Lemma by first considering an exponential family likelihood function $f_{X|Z} \colon \mathcal{Z} \to \mathcal{M}_X$ that does not presuppose a probabilistic structure over $\mathcal{Z}$. We then choose a family of priors $\mathcal{M}_Z$, so that given a particular prior $q_Z \in \mathcal{M}_Z$, we define the posterior $q_{Z|X}$ using Bayes' rule as

$$q_{Z|X}(z \mid x) \propto f_{X|Z}(x \mid z) q_Z(z). \tag{15}$$

From this perspective we may consider families of conjugate priors for a fixed likelihood.

Definition 7 (Conjugate Prior Family) *The exponential family $\mathcal{M}_Z$ is a conjugate prior family for the likelihood $f_{X|Z} \colon \mathcal{Z} \to \mathcal{M}_X$ if the posterior $q_{Z|X} \in \mathcal{M}_Z$ for any prior $q_Z \in \mathcal{M}_Z$, where the posterior is defined by Bayes' rule (Eq. 15).*

We may then generalize the conditions of the Conjugation Lemma to arrive at what we refer to as the Conjugation Theorem.

Theorem 8 (The Conjugation Theorem) *Suppose that $\mathcal{M}_X$ and $\mathcal{M}_Z$ are exponential families with minimal sufficient statistics $\mathbf{s}_X$ and $\mathbf{s}_Z$, respectively, and that $f_{X|Z} \colon \mathcal{Z} \to \mathcal{M}_X$, where $\mathcal{Z}$ is the sample space of $\mathcal{M}_Z$. Then $\mathcal{M}_Z$ is a conjugate prior family for the likelihood $f_{X|Z}$ if and only if $f_{X|Z}$ has the form*

$$f_{X|Z}(x \mid z) = e^{\mathbf{s}_X(x) \cdot (\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)) - \mathbf{s}_Z(z) \cdot \boldsymbol{\rho} - \chi}, \tag{16}$$

for some conjugation parameters χ and $\boldsymbol{\rho}$.

Proof *For $\implies$, suppose $\mathcal{M}_Z$ is a conjugate prior family for $f_{X|Z}$. Where $q_{XZ} = f_{X|Z} \cdot q_Z$ for some $q_Z \in \mathcal{M}_Z$, the likelihood of q_{XZ} is $f_{X|Z}$, and its posterior $q_{Z|X}$ is given by Bayes' rule (Eq. 15). Since $f_{X|Z} \in \mathcal{M}_X$ and $q_{Z|X} \in \mathcal{M}_Z$ by assumption, Theorem 1 implies that q_{XZ} is in the harmonium $\mathcal{H}_{XZ}$ defined by $\mathcal{M}_X$ and $\mathcal{M}_Z$, which implies that $f_{X|Z}$ has the form of Equation 5. Finally, since $q_Z \in \mathcal{M}_Z$, Equation 10 holds according to the Conjugation Lemma (Lm. 4), and Equation 16 follows directly by combining Equations 5 and 10.*

For $\Longleftarrow$, if Equation 16 holds, then for any $q_Z \in \mathcal{M}_Z$ with parameters $\boldsymbol{\theta}^*_Z$,

$$\begin{aligned} q_{Z|X}(z \mid x) &\propto f_{X|Z}(x \mid z) q_Z(z) \\ &= e^{\mathbf{s}_X(x)\cdot\boldsymbol{\theta}_X + \mathbf{s}_X(x)\cdot\boldsymbol{\Theta}_{XZ}\cdot\mathbf{s}_Z(z) - \mathbf{s}_Z(z)\cdot\boldsymbol{\rho} - \chi} e^{\mathbf{s}_Z(z)\cdot\boldsymbol{\theta}^*_Z} \\ &\propto e^{\mathbf{s}_Z(z)\cdot(\boldsymbol{\theta}^*_Z - \boldsymbol{\rho} + \mathbf{s}_X(x)\cdot\boldsymbol{\Theta}_{XZ})}, \end{aligned}$$

which implies $q_{Z|X} \in \mathcal{M}_Z$. ■

For the Conjugation Lemma we assume that the joint density q_{XZ} in question has the exponential family structure of a harmonium density. For the Conjugation Theorem, on the other hand, we merely assume that the likelihood in question $f_{X|Z}$ is exponential family distributed, and from that it follows that $f_{X|Z}$ must have the form of Equation 16 to support conjugate priors. Likelihoods with this form are important enough to earn a name.

Definition 9 (Conjugating Likelihood) *Where $\mathcal{M}_X$ is an exponential family and $\mathcal{Z}$ is a sample space, $f_{X|Z} \colon \mathcal{Z} \to \mathcal{M}_X$ is a conjugating likelihood if its has the form of Equation 16.*

We may intuitively interpret Bayes' rule as a function defined by the likelihood, where the prior is the input and the posterior is the output. If the prior is conjugate to the posterior, then Bayes' rule can be reapplied using the posterior as a new input, and this pattern will allow us to develop recursive algorithms for Bayesian inference (Sec. 3.4.1). Given a conjugating likelihood and a density from a conjugate prior family, the posterior has a simple expression.

Corollary 10 *Suppose that $f_{X|Z}$ is a conjugating likelihood with natural parameters $\boldsymbol{\theta}_X$ and $\boldsymbol{\Theta}_{XZ}$ and conjugation parameters $\boldsymbol{\rho}$ and χ, that $\mathcal{M}_Z$ is a conjugate prior family for $f_{X|Z}$, and that the prior $q_Z \in \mathcal{M}_Z$ has parameters $\boldsymbol{\theta}^*_Z$. Then the posterior $q_{Z|X=x} \in \mathcal{M}_Z$ has parameters*

$$\boldsymbol{\theta}_{Z|X}(x) = \boldsymbol{\theta}^*_Z + \mathbf{s}_X(x) \cdot \boldsymbol{\Theta}_{XZ} - \boldsymbol{\rho}. \tag{17}$$

Proof *See the proof of Theorem 8.* ■

Our last corollary for this section details exactly how the parameters of a conjugated harmonium relate to those of a conjugating likelihood.

Corollary 11 *Let $\mathcal{H}_{XZ}$ be the harmonium defined by $\mathcal{M}_X$ and $\mathcal{M}_Z$. Then $q_{XZ} \in \mathcal{H}_{XZ}$ is a conjugated harmonium density with natural parameters $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_Z$, and $\boldsymbol{\Theta}_{XZ}$ and conjugation parameters $\boldsymbol{\rho}$ and χ, if and only if $q_{X|Z}$ is a conjugating likelihood with natural parameters $\boldsymbol{\theta}_X$ and $\boldsymbol{\Theta}_{XZ}$ and conjugation parameters $\boldsymbol{\rho}$ and χ, and the prior $q_Z \in \mathcal{M}_Z$ with parameters $\boldsymbol{\theta}^*_Z = \boldsymbol{\theta}_Z + \boldsymbol{\rho}$.*

Proof *For $\Longrightarrow$ we apply Equation 10 to the harmonium likelihood (Eq. 5).*

For $\Longleftarrow$ we multiply the definitions of the conjugating likelihood and prior and see that

$$q_{X|Z}(x \mid z) \cdot p(z) \propto e^{\mathbf{s}_X(x)\cdot\boldsymbol{\theta}_X + \mathbf{s}_X(x)\cdot\boldsymbol{\Theta}_{XZ}\cdot\mathbf{s}_Z(z) - \mathbf{s}_Z(z)\cdot\boldsymbol{\rho} + \mathbf{s}_Z(z)\cdot(\boldsymbol{\theta}_Z + \boldsymbol{\rho})} \propto q_{XZ}(x, z).$$

■

3.2 A Collection of Conjugated Harmoniums

In this section we review a variety of well-known models, show how they can be expressed as harmoniums, and derive the conditions under which they are conjugated. In particular, mixture models and linear Gaussian models are two classes of LVM for which inference and learning are broadly solvable in closed-form. We show that both of these LVMs are forms of conjugated harmonium, and how to generalize them in novel ways. We next develop a theory of conjugation for LVMs with likelihoods based on products of Poisson distributions, that generalizes previous results on probabilistic computation in biological neural circuits (Ma et al., 2006; Beck et al., 2011). We conclude by showing how Bayesian estimation of exponential family parameters is a special case of our inference framework.

3.2.1 Mixture Models

A mixture model describes the statistics of observations with a weighted sum of component models. We can interpret this sum as the observable density q_X of the LVM $q_{XK} = q_{X|K} \cdot q_X$, by equating the likelihood $q_{X|K=k}$ at each k with one of the mixture components, and the prior q_K with the weights. The posterior $q_{K|X=x}$ of q_{XK} is then a set of weights that provides a soft classification of the observation x.

The d_K-dimensional categorical exponential family $\mathcal{M}_K$ contains all densities over indices from 0 to d_K. It is defined by the base measure $\mu_K(k) = 1$, and the sufficient statistic given by a "one-hot" vector, where $\mathbf{s}_K(0) = \mathbf{0}$, and for $k > 0$, $s_{K,i}(k) = 1$ if $i = k$, and 0 otherwise. Because the posterior of a mixture model $q_{K|X}$ is a density over indices, it must be in $\mathcal{M}_K$. If then we assume that the likelihood $q_{X|K} \in \mathcal{M}_X$ for a chosen exponential family $\mathcal{M}_X$ over X, then q_{XK} is an element of the harmonium $\mathcal{H}_{XK}$ defined by $\mathcal{M}_X$ and the categorical family $\mathcal{M}_K$ (Thm. 1). Moreover, the harmonium $\mathcal{H}_{XK}$ is conjugated.

Theorem 12 *Let $\mathcal{H}_{XK}$ be the harmonium defined by $\mathcal{M}_X$ and $\mathcal{M}_K$, where $\mathcal{M}_K$ is the categorical family of dimension d_K. Then $\mathcal{H}_{XK}$ is conjugated, and the conjugation parameters of any $q_{XK} \in \mathcal{H}_{XK}$ are given by*

$$\begin{aligned} \rho_i &= \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\theta}_{XK,i}) - \psi_X(\boldsymbol{\theta}_X), \\ \chi &= \psi_X(\boldsymbol{\theta}_X), \end{aligned} \tag{18}$$

where $(\boldsymbol{\theta}_X, \boldsymbol{\theta}_K, \boldsymbol{\Theta}_{XK})$ are the parameters of q_{XK}, and $\boldsymbol{\theta}_{XK,i}$ is the ith column of $\boldsymbol{\Theta}_{XK}$.

Proof *We must show that for any $k \in \mathcal{K}$, Equation 10 is satisfied for some χ and $\boldsymbol{\rho}_K$. For $k = 0$,*

$$\begin{aligned} &\psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XK} \cdot \mathbf{s}_K(k)) = \mathbf{s}_K(k) \cdot \boldsymbol{\rho}_K + \chi \\ \iff &\chi = \psi_X(\boldsymbol{\theta}_X). \end{aligned}$$

For $k > 0$,

$$\begin{aligned} &\psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XK} \cdot \mathbf{s}_K(k)) = \mathbf{s}_K(k) \cdot \boldsymbol{\rho}_K + \chi \\ \iff &\psi_X(\boldsymbol{\theta}_X + \boldsymbol{\theta}_{XK,i}) = \rho_{K,i} + \chi \\ \iff &\rho_{K,i} = \psi_X(\boldsymbol{\theta}_X + \boldsymbol{\theta}_{XK,i}) - \chi. \end{aligned}$$

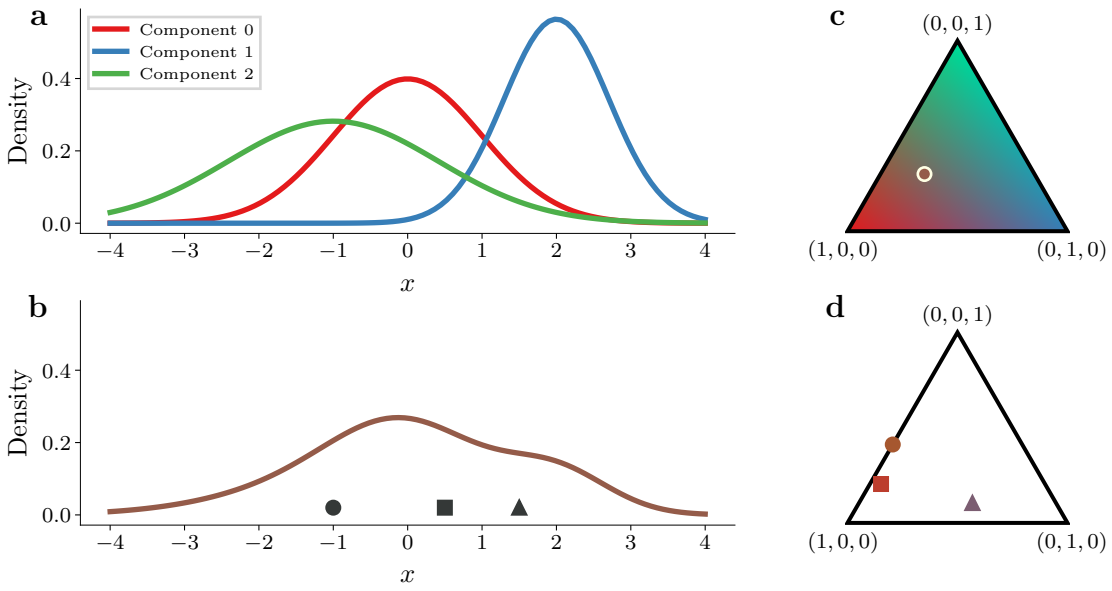

Figure 2: *A mixture of normal densities.* **a:** The three component densities $q_{X|K=0}$ (red line), $q_{X|K=1}$ (blue line), and $q_{X|K=2}$ (green line) of a mixture of normal densities q_{XK}. **b:** The mixture density q_X of q_{XK} (brown line), and three observations $x^{(1)}$, (circle) $x^{(2)}$ (square), and $x^{(3)}$ (diamond). **c:** The exponential family $\mathcal{M}_K$ (simplex) of categorical densities over 3 states (coloured points), and the weights $(0.5, 0.2, 0.3)$ of q_K (brown dot). **d:** The weights (dot colour) of the posterior densities $q_{K|X=x^{(1)}}$ (circle), $q_{K|X=x^{(2)}}$ (square), and $q_{K|X=x^{(3)}}$ (diamond).

The conjugation parameters defined by Equations 18 thus satisfy the Conjugation Equation (Eq. 10) for any $k \in \mathcal{K}$. ■

In Figure 2 we demonstrate a simple mixture of normal densities q_{XK}. Although such a mixture model is far from novel, it exemplifies key features of conjugated harmoniums that we will generalize to more complex models. In particular, the likelihood $q_{X|K} \in \mathcal{M}_X$ for each $k \in \{0, 1, 2\}$ (Fig. 2**a**), yet the observable density q_X is not in $\mathcal{M}_X$ (Fig. 2**b**). On the other hand, both the prior (Fig. 2**c**) and the posterior (Fig. 2**d**) are in $\mathcal{M}_K$. In Section 3.4.4 we use a mixture model as a component of a graphical model.

3.2.2 Linear (Gaussian) Models

A linear Gaussian model (LGM) $\mathcal{G}_{XZ}$ is a set of $n + m$ dimensional multivariate normal densities over observable and latent variables X and Z. Where $\mathcal{M}_X$ and $\mathcal{M}_Z$ are the n and m dimensional multivariate normal families, respectively, the analytic properties of normal densities ensures that for any $q_{XZ} \in \mathcal{G}_{XZ}$, both the likelihood $q_{X|Z}$ and observable density

q_X are in $\mathcal{M}_X$, and the posterior $q_{Z|X}$ and prior q_Z are in $\mathcal{M}_Z$ (Bishop, 2006). Because $q_X \in \mathcal{M}_X$, the $\mathcal{G}_{XZ}$ has no additional representational power over $\mathcal{M}_X$, and so we typically assume additional constraints on the parameters of $\mathcal{G}_{XZ}$. For example, Factor analysis (FA) and principal component analysis (PCA) are forms of LGM that assume that $\mathcal{M}_X$ is the family of multivariate normals with diagonal or isotropic covariance matrices, respectively — consequently, the observable densities for FA and PCA may have non-diagonal covariance matrices and lay outside of $\mathcal{M}_X$. These assumptions allow FA and PCA to effectively model high dimensional observations X, where an unconstrained multivariate normal model would suffer from the curse of dimensionality.

The exponential family of n-dimensional multivariate normal densities $\mathcal{M}_X$ is defined by the base measure $\mu_X(\mathbf{x}) = (2\pi)^{-\frac{n}{2}}$ and sufficient statistic $\mathbf{s}_X(\mathbf{x}) = (\mathbf{x}, \text{low}(\mathbf{x} \otimes \mathbf{x}))$, where $\text{low}(\mathbf{A})$ is the lower triangular part of the given matrix $\mathbf{A}$ (this construction ensures the minimality of $\mathbf{s}_X$). We can partition the natural parameters $\boldsymbol{\theta}_X$ of any $q_X \in \mathcal{M}_X$ into $\boldsymbol{\theta}_X = (\boldsymbol{\theta}^m_X, \boldsymbol{\theta}^\sigma_X)$, such that $\boldsymbol{\theta}^m_X$ and $\boldsymbol{\theta}^\sigma_X$ weight the statistics $\mathbf{x}$ and $\text{low}(\mathbf{x} \otimes \mathbf{x})$ in the exponential family form of q_X (Eq. 1). We can also define the so-called precision matrix $\boldsymbol{\Theta}^\sigma_X$ such that $\boldsymbol{\theta}_X \cdot \mathbf{s}_X(\mathbf{x}) = \mathbf{x} \cdot \boldsymbol{\theta}^m_X + \mathbf{x} \cdot \boldsymbol{\Theta}^\sigma_X \cdot \mathbf{x}$ by converting $\boldsymbol{\theta}^\sigma_X$ into a lower triangular matrix $\boldsymbol{\Theta}^L_X$, and letting $\boldsymbol{\Theta}^\sigma_X = \frac{1}{2}(\boldsymbol{\Theta}^L_X + \boldsymbol{\Theta}^U_X)$, where $\boldsymbol{\Theta}^U_X$ is the transpose of $\boldsymbol{\Theta}^L_X$.

For any $q_{XZ} \in \mathcal{G}_{XZ}$, both $q_{X|Z} \in \mathcal{M}_X$ and $q_{Z|X} \in \mathcal{M}_Z$, and LGMs are therefore subsets of the harmonium $\mathcal{H}_{XZ}$ defined by $\mathcal{M}_X$ and $\mathcal{M}_Z$ (Thm. 1). However, the harmonium $\mathcal{H}_{XZ}$ also contains densities with interactions between the second-order statistics of X and Z, whereas the likelihoods and posteriors of a linear Gaussian model $\mathcal{G}_{XZ}$ model only homoscedastic (first-order) interactions. $\mathcal{G}_{XZ}$ is thus the subset of $\mathcal{H}_{XZ}$ for which all second-order interactions $\theta_{XZ,ij}$ are 0. Moreover, the densities in this subset are conjugated.

Theorem 13 *Let $\mathcal{H}_{XZ}$ be a harmonium defined by the exponential families $\mathcal{M}_X$ and $\mathcal{M}_Z$, where $\mathcal{M}_X$ is the multivariate normal family, and the sufficient statistic of $\mathcal{M}_Z$ is given by $\mathbf{s}_Z(\mathbf{z}) = (\mathbf{z}, \text{low}(\mathbf{z} \otimes \mathbf{z}))$. Then $q_{XZ} \in \mathcal{H}_{XZ}$ is conjugated if*

$$\boldsymbol{\Theta}_{XZ} = \begin{pmatrix} \boldsymbol{\Theta}^m_{XZ} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{19}$$

with conjugation parameters χ, , $\boldsymbol{\rho}^m_Z$, and $\mathbf{P}^\sigma_Z$ given by

$$\begin{aligned} \chi &= -\frac{1}{4}\boldsymbol{\theta}^m_X \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot \boldsymbol{\theta}^m_X - \frac{1}{2}\log|-2\boldsymbol{\Theta}^\sigma_X|, \\ \boldsymbol{\rho}^m_Z &= -\frac{1}{2}\boldsymbol{\Theta}^m_{ZX} \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot \boldsymbol{\theta}^m_X, \\ \mathbf{P}^\sigma_Z &= -\frac{1}{4}\boldsymbol{\Theta}^m_{ZX} \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot \boldsymbol{\Theta}^m_{XZ}, \end{aligned} \tag{20}$$

where $\boldsymbol{\Theta}^m_{ZX}$ is the transpose of $\boldsymbol{\Theta}^m_{XZ}$, and $\boldsymbol{\rho}^m_Z$ $\mathbf{P}^\sigma_Z$ are the conjugation parameters of the precision-weighted means $\boldsymbol{\theta}^m_X$ and precision matrix $\boldsymbol{\Theta}^\sigma_Z$, respectively.

Proof *The log-partition function of a multivariate normal family $\mathcal{M}_X$ is given by*

$$\psi_X(\boldsymbol{\theta}^m_X, \boldsymbol{\Theta}^\sigma_X) = -\frac{1}{4}\boldsymbol{\theta}^m_X \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot \boldsymbol{\theta}^m_X - \frac{1}{2}\log|-2\boldsymbol{\Theta}^\sigma_X|.$$

Assuming Equation 19 holds, we may express the LHS of Equation 10 as

$$\begin{aligned}\psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(\mathbf{z})) &= \psi_X(\boldsymbol{\theta}^m_X + \boldsymbol{\Theta}^m_{XZ} \cdot \mathbf{z}, \boldsymbol{\Theta}^\sigma_X) \\ &= -\frac{1}{4}(\boldsymbol{\theta}^m_X + \boldsymbol{\Theta}^m_{XZ} \cdot \mathbf{z}) \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot (\boldsymbol{\theta}^m_X + \boldsymbol{\Theta}^m_{XZ} \cdot \mathbf{z}) - \frac{1}{2}\log|-2\boldsymbol{\Theta}^\sigma_X|,\end{aligned}$$

and the RHS as

$$\mathbf{s}_Z(\mathbf{z}) \cdot \boldsymbol{\rho} + \chi = \mathbf{z} \cdot \boldsymbol{\rho}^m_Z + \mathbf{z} \cdot \mathbf{P}^\sigma_Z \cdot \mathbf{z} + \chi,$$

so that

$$\begin{aligned}\mathbf{z} \cdot \boldsymbol{\rho}^m_Z + \mathbf{z} \cdot \mathbf{P}^\sigma_Z \cdot \mathbf{z} + \chi = &-\frac{1}{2}\mathbf{z} \cdot {\boldsymbol{\Theta}^{m,\top}_{XZ}} \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot \boldsymbol{\theta}^m_X \\ &-\frac{1}{4}\mathbf{z} \cdot {\boldsymbol{\Theta}^{m,\top}_{XZ}} \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot \boldsymbol{\Theta}^m_{XZ} \cdot \mathbf{z} \\ &-\frac{1}{4}\boldsymbol{\theta}^m_X \cdot {\boldsymbol{\Theta}^\sigma_X}^{-1} \cdot \boldsymbol{\theta}^m_X - \frac{1}{2}\log|-2\boldsymbol{\Theta}^\sigma_X|,\end{aligned}$$

which is clearly solved by Equations 20. ■

Theorem 13 also be applies in cases where $\mathcal{M}_Z$ is not a multivariate normal family, as long as its sufficient statistic is given by $\mathbf{s}_Z(\mathbf{z}) = (\mathbf{z}, \text{low}(\mathbf{z} \otimes \mathbf{z}))$. One example of such families are Boltzmann machines (Ackley et al., 1985), which model the second-order statistics between binary random variables, typically called binary neurons. Gaussian-Boltzmann machines have been previously explored for modelling the joint density of continuous stimuli and neural recordings (Gerwinn et al., 2009), and here we demonstrate how they can serve as tractable latent variable models (Fig. 3).

We consider a Gaussian-Boltzmann harmonium $\mathcal{H}_{XZ}$ defined by the bivariate normal family $\mathcal{M}_X$ and the family of Boltzmann distributions $\mathcal{M}_Z$ with 6 neurons. We learn $q_{XZ} \in \mathcal{H}_{XZ}$ by fitting the model to a synthetic dataset generated from two, noisy concentric circles (Fig. 3**a**). Analogous to a mixture model, we consider a set of "component" densities by evaluating the likelihood $q_{X|Z=\mathbf{z}^{(i)}}$ at the one-hot vectors $\mathbf{z}^{(i)}$ for each neuron i, such that $z^{(i)}_j = 1$ if $i = j$, and 0 otherwise (Fig. 3**a**). Although each $q_{X|Z=\mathbf{z}^{(i)}}$ has the same bivariate normal shape, the observable density q_X is not in the family of bivariate normals, and successfully captures the concentric circle structure (Fig. 3**b**). The correlation matrix of the prior $q_Z \in \mathcal{M}_Z$ reveals how the population of neurons encodes the concentric circles (Fig. 3**c**). Similarly, the moment matrices $\mathbb{E}_Q[Z \otimes Z \mid X = \mathbf{x}^{(i)}]$ of the posterior $q_{Z|X=\mathbf{x}^{(i)}} \in \mathcal{M}_Z$ at 4 example observations $\mathbf{x}^{(i)}$ demonstrates how q_{XZ} encodes each point within the concentric circles by activating and correlating certain subsets of neurons, which mixes and distorts the components $q_{X|Z=\mathbf{z}^{(i)}}$ (Fig. 3**d**). We cover how we trained this model in Section 3.3.

3.2.3 Probabilistic Population Codes

Neuroscientists often model the spiking activity of populations of neurons as random vectors of counts $N = (N_1, \ldots, N_{d_N})$, where each N_i is the spike count of a single neuron. These spike counts may also depend on a stimulus or environmental variable Z, and theories of probabilistic population coding use the framework of Bayesian inference to explain how the

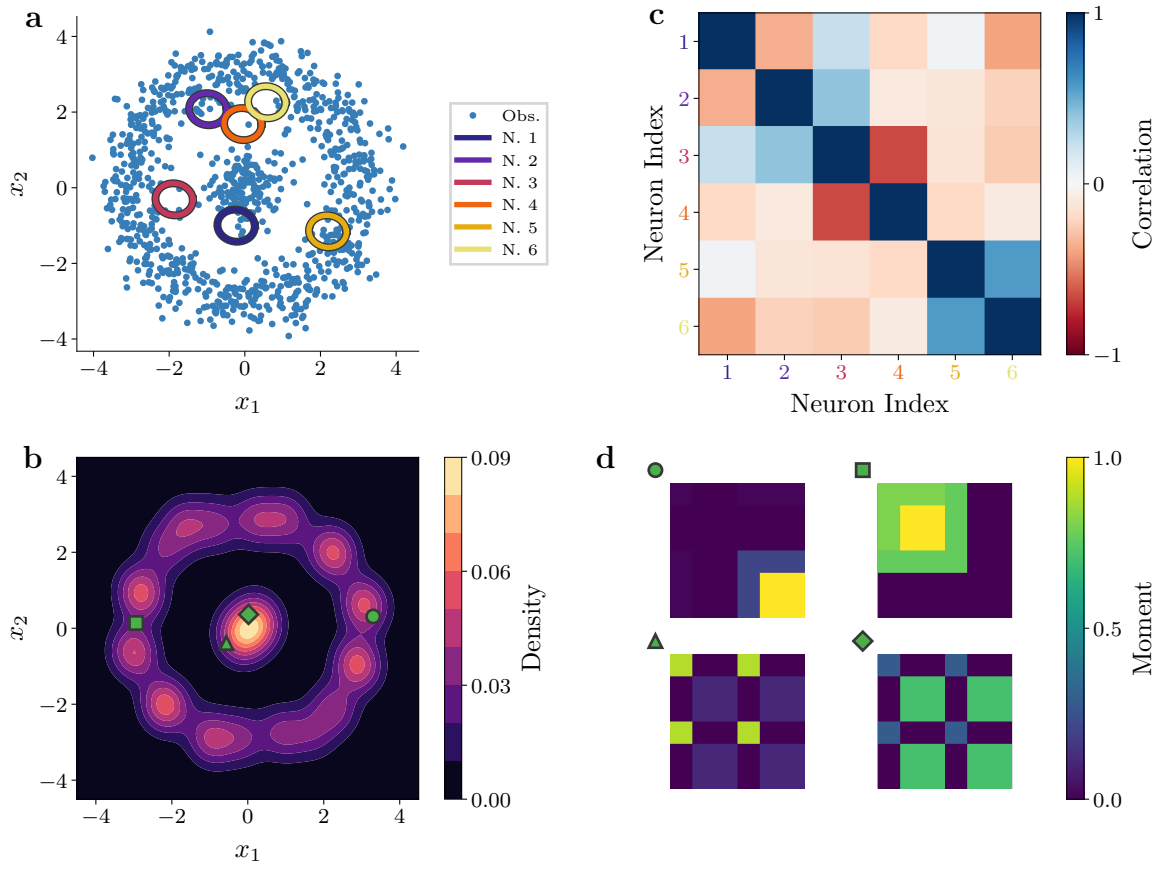

Figure 3: *A Gaussian-Boltzmann model.* **a:** Training data (blue dots) used to learn a Gaussian-Boltzmann density q_{XZ}, and confidence ellipses (coloured circles) of the likelihoods $q_{X|Z=\mathbf{z}^{(i)}}$ given one-hot vectors $\mathbf{z}^{(i)}$ (neuron index indicated by "N. i"). **b:** Observable density q_Z, and example observations $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(4)}$ (green shapes). **c:** Correlation matrix of the prior q_Z. **d:** Moment matrices $\mathbb{E}_Q[Z \otimes Z \mid X = \mathbf{x}^{(i)}]$ of the posteriors for each example observation (labelled by green shape).

neural activity N encodes and processes information about Z. In particular, researchers have found that, under certain conditions, stimulus-dependent models with Poisson-distributed spike counts can support optimal Bayesian inference (Ma et al., 2006; Beck et al., 2011). As we next show, these conditions are special cases of the Conjugation Lemma (Thm. 4).

The family of Poisson distributions $\mathcal{P}_{N_i}$ is defined by the sufficient statistic $s_{N_i}(n) = n$ and base measure $\mu_{N_i}(n) = \frac{1}{n!}$. By extension, the family of independent Poisson product distributions $\mathcal{M}_N$ is defined by the sufficient statistic $\mathbf{s}_N(\mathbf{n}) = \mathbf{n}$ and base measure $\mu_N(\mathbf{n}) = \left(\prod_{i=1}^{d_N} n_i!\right)^{-1}$. Every density $q_N \in \mathcal{M}_N$ can be factored into $q_N = q_{N_1} \cdots q_{N_{d_N}}$ where each $q_{N_i} \in \mathcal{P}_{N_i}$ is a Poisson distribution. In the context of neuroscience we define a population code for a stimulus Z as a joint density q_{NZ}. We refer to the mean spike-count $\mathbb{E}_Q[N_i]$

of N_i under Q as the firing rate, and its stimulus dependent firing-rate $\mathbb{E}_Q[N_i \mid Z]$ as its tuning curve. Population codes are typically specified so that the likelihood $q_{N|Z}$ of a population code is in the Poisson product family $\mathcal{M}_N$, and the posterior $q_{Z|N}$ is in some chosen exponential family $\mathcal{M}_Z$ over the stimulus. Such a population code q_{XZ} is thus an element of the harmonium $\mathcal{H}_{NZ}$ defined by $\mathcal{M}_N$ and $\mathcal{M}_Z$ (Thm. 1). Moreover, under certain conditions on the sum of tuning curves, such population codes are approximately conjugated.

Theorem 14 *Let $\mathcal{H}_{NZ}$ be a harmonium defined by $\mathcal{M}_N$ and $\mathcal{M}_Z$, where $\mathcal{M}_N$ is the Poisson product family. Then $q_{NZ} \in \mathcal{H}_{NZ}$ is conjugated with conjugation parameters $\boldsymbol{\rho}$ and χ if it satisfies*

$$\sum_{i=1}^{d_N} \mathbb{E}_Q[N_i \mid Z = z] = \boldsymbol{\rho} \cdot \mathbf{s}_Z(z) + \chi. \tag{21}$$

Proof *Firstly, note that the log-partition function of a Poisson exponential family $\mathcal{M}_{N_i}$ is given by $\psi_{N_i}(\theta_{N_i}) = e^{\theta_{N_i}}$, which implies $\psi_{N_i}(\theta_{N_i}) = \partial_{\theta_{N_i}} \psi_{N_i}(\theta_{N_i}) = \mathbb{E}_Q[N_i]$. Moreover, the mutual independence of the random counts N_i implies that $\psi_N(\boldsymbol{\theta}_N) = \sum_{i=1}^{d_N} \psi_{N_i}(\theta_{N,i})$, and therefore that $\psi_N(\boldsymbol{\theta}_N) = \sum_{i=1}^{d_N} \mathbb{E}_Q[N_i]$.*

Now, suppose $q_{NZ} \in \mathcal{H}_{NZ}$ has natural parameters $(\boldsymbol{\theta}_N, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{NZ})$. Then the likelihood $q_{N|Z=z} \in \mathcal{M}_N$ has natural parameters $\boldsymbol{\theta}_N + \boldsymbol{\Theta}_{NZ} \cdot \mathbf{s}_Z(z)$, so that $\psi_N(\boldsymbol{\theta}_N + \boldsymbol{\Theta}_{NZ} \cdot \mathbf{s}_Z(z)) = \sum_{i=1}^{d_N} \mathbb{E}_P[N_i \mid Z = z]$. Therefore, if Equation 21 holds,

$$\boldsymbol{\rho}_N \cdot \mathbf{s}_N(\mathbf{n}) + \chi = \sum_{i=1}^{d_N} \mathbb{E}_P[N_i \mid Z = z] = \psi_N(\boldsymbol{\theta}_N + \boldsymbol{\Theta}_{NZ} \cdot \mathbf{s}_Z(z)),$$

and therefore by the Conjugation Lemma (Thm. 4), q_{NZ} is conjugated with conjugation parameters $\boldsymbol{\rho}$ and χ. ■

The canonical neural population code that supports exact Bayesian inference has Gaussian tuning curves that sum to a constant (Ma et al., 2006), and Equation 10 generalizes this constraint by allowing the sum to depend on the sufficient statistics of z. Unlike Theorems 12 and 13, the result of Theorem 14 does not specify a manifold of probability densities that exactly satisfy Equation 10. Nevertheless, depending on the choice of $\mathcal{M}_Z$, tuning curves often have simple (e.g. von Mises or Gaussian) shapes, and serve well as basis functions, so that ensuring that the sum of the tuning curves is an affine function of $\mathbf{s}_Z(z)$ can be easily achieved with a sufficient numbers of model neurons.

We demonstrate conjugated population codes with a model of how the spike counts $N_1, \ldots, N_8$ of 8 neurons can encode the location of an oriented stimulus (Fig. 4). We model the joint density of N and Z with the harmonium $\mathcal{M}_{NZ}$ defined by the Poisson product family $\mathcal{M}_N$ with $d_N = 8$ Poisson neurons, and the von Mises family $\mathcal{M}_Z$, which is defined by the sufficient statistic $\mathbf{s}_Z = (\cos z, \sin z)$, and base measure $\mu_Z(z) = \frac{1}{2\pi}$. We choose a population code $q_{NZ} \in \mathcal{H}_{NZ}$, and demonstrate that it is approximately conjugated by fitting a function of the form $\rho_1 \cos z + \rho_2 \sin z + \chi$ to the sum of its tuning curves (Fig. 4**a**). We also find that the observable density q_N is not an element of $\mathcal{M}_N$, and rather describes a regular pattern of correlations between the neurons (Fig. 4**b**). Nevertheless, the prior q_Z is

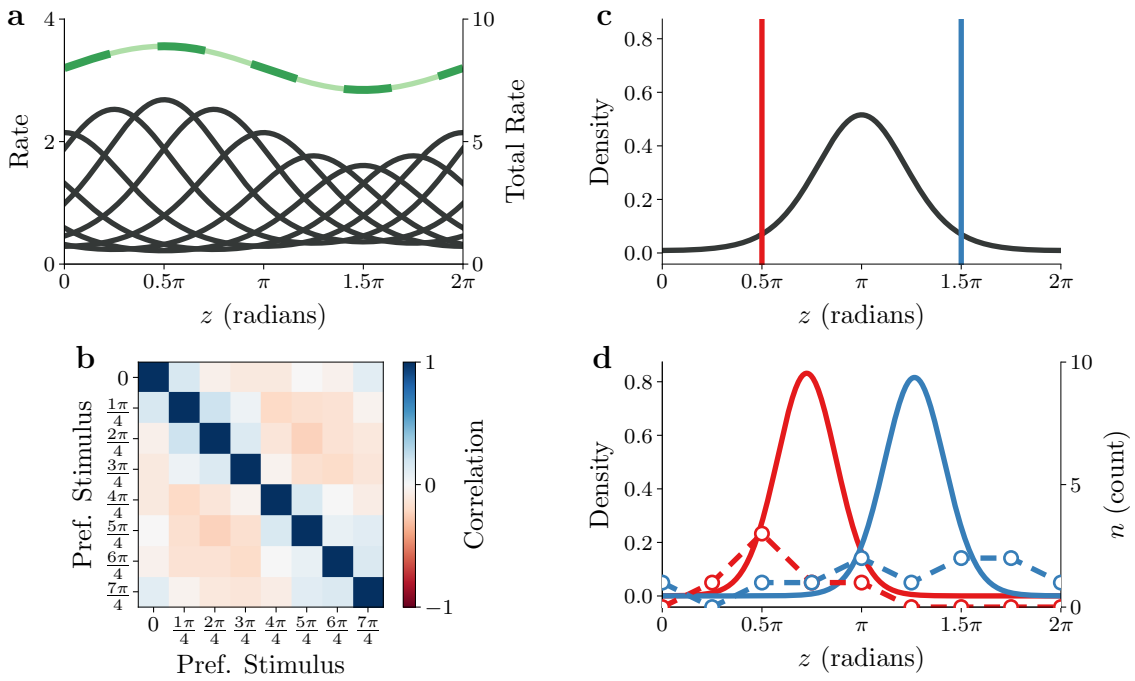

Figure 4: *Conjugation with population codes.* The population code q_{NZ} is a density over 8 spike counts N_i and an oriented stimulus Z. **a:** Tuning curves $\mathbb{E}_Q[N_i \mid Z]$ (black lines), their sum (light green line), and fit of the conjugation parameters (dashed green line) to the sum. **b:** Correlation matrix of q_N, with neuron identified by tuning curve peak (preferred stimulus). **c:** Population code prior q_Z (black line), and two example stimuli $z^{(1)}$ and $z^{(2)}$ (red and blue lines). **d:** Example spike count vectors $N^{(i)} \sim q_{N|Z=z^{(i)}}$ (red and blue line-dots), and posteriors $q_{Z|N=N^{(i)}}$ (red and blue lines).

in $\mathcal{M}_Z$ (Fig. 4**c**). Finally, given spike counts generated in response to a stimulus, we use the posterior $q_{Z|N}$ to decode accurate von Mises densities over stimulus orientation, while accounting for prior beliefs about stimulus orientation (Fig. 4**d**).

3.2.4 Bayesian Parameter Estimation

The final harmonium we consider returns us to the conceptual origins of conjugate priors, in which we infer posteriors over the parameters of exponential family models. In particular, we show how such models are special cases of conjugated harmonium, and thereby clarify the relationship between our theory and the well-established theory of Bayesian parameter estimation.

Bayesian inference over exponential family natural parameters begins by simply treating the parameters of an exponential family as a latent random variable Z (Diaconis and Ylvisaker, 1979; Arnold et al., 1993). A classic result in Bayesian parameter estimation is

that any exponential family $\mathcal{M}_X$ has a conjugate prior family $\mathcal{M}_Z$ with sufficient statistic $s_Z(z) = (z, \psi_X(z))$, where ψ_X is the log-partition function of $\mathcal{M}_X$.

Theorem 15 *Suppose $\mathcal{M}_X$ is a d_X-dimensional exponential family, and let $\mathcal{M}_Z$ be the $d_X + 1$-dimensional exponential family with sample space $\mathcal{Z} = \Theta_X$, and sufficient statistic given by $\mathbf{s}_Z(z) = (z, \psi_X(z))$ for any $z \in \Theta_X$. Then $q_{XZ} \in \mathcal{H}_{XZ}$ is conjugated if $\boldsymbol{\theta}_X = \mathbf{0}$, and*

$$\boldsymbol{\Theta}_{XZ} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{22}$$

Proof *Suppose $\boldsymbol{\theta}_X = \mathbf{0}$ and $\boldsymbol{\Theta}_{XZ}$ is given by Eq. 22. Then*

$$\psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}_Z(z)) = \psi_X(z) = \mathbf{s}_Z(z) \cdot \boldsymbol{\rho} + \chi,$$

where $\boldsymbol{\rho} = (0, 0, \ldots, 0, 1)$ and $\chi = 0$. Therefore q_{XZ} is conjugated by Lemma 4. ■

Although this provides a general solution to Bayesian parameter estimation, in practice the resulting exponential families are often intractable, and many exponential families have alternative conjugate priors with more tractable forms. We explore one example of such a conjugate prior in Section 3.4.1.

3.3 Training Harmoniums

Harmoniums affords simple, yet general expressions for LVM training algorithms within the framework of log-likelihood maximization, or equivalently, cross-entropy minimization — in this paper we favour the terminology of cross-entropy minimization to avoid any confusion of our training objective with the likelihood $q_{X|Z}$ of an LVM density q_{XZ}. In general, where q_X is a parametric density over observable variable X with parameters $\boldsymbol{\theta}$, the cross-entropy objective is

$$\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(X^{(i)}, \boldsymbol{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \log q_X(X^{(i)}), \tag{23}$$

where $X^{(1)}, \ldots, X^{(n)}$ is a sample from P_X, and $\mathcal{L}(X^{(i)}, \boldsymbol{\theta}) = -\log q_X(X^{(i)})$ is the pointwise cross-entropy.

When q_X is an element of some exponential family $\mathcal{M}_X$ with parameters $\boldsymbol{\theta}_X$, then $\mathcal{L}(X^{(i)}, \boldsymbol{\theta}) = -\mathbf{s}_X(X^{(i)}) \cdot \boldsymbol{\theta}_X + \psi_X(\boldsymbol{\theta}_X)$. Moreover, $\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_X$ is a convex function on the natural parameter space Θ_X of $\mathcal{M}_X$, and its optimum is given by $\boldsymbol{\tau}_X^{-1}(-\frac{1}{n} \sum_{i=1}^{n} \mathbf{s}_X(X^{(i)}))$, where $\boldsymbol{\tau}_X^{-1}$ is the backward mapping of $\mathcal{M}_X$ (Wainwright and Jordan, 2008). However, for a harmonium density $q_{XZ} \in \mathcal{H}_{XZ}$, the observable density q_X is not generally in $\mathcal{M}_X$, and accounting for the latent variable Z in the optimization necessitates additional analysis.

3.3.1 Expectation-Maximization

EM is arguably the standard algorithm for minimizing the cross-entropy of an LVM. For an arbitrary exponential family $\mathcal{M}_{XZ}$ over a X and Z, defined by the sufficient statistic $\mathbf{s}_{XZ}$ and base measure μ_{XZ}, the E-Step of EM is to calculate the conditional sufficient statistics

$$\boldsymbol{\eta}_{XZ}^{(i)} = \mathbb{E}_Q[\mathbf{s}_{XZ}(X, Z) \mid X = X^{(i)}], \tag{24}$$

for a given $q_{XZ} \in \mathcal{M}_{XZ}$ (Wainwright and Jordan, 2008). If the exponential family is a harmonium $\mathcal{H}_{XZ}$ defined by $\mathcal{M}_X$ and $\mathcal{M}_Z$, and $q_{XZ} \in \mathcal{H}_{XZ}$ with parameters $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_Z$, and $\boldsymbol{\Theta}_{XZ}$, then the conditional sufficient statistics $\boldsymbol{\eta}_{XZ}^{(i)} = (\boldsymbol{\eta}_X^{(i)}, \boldsymbol{\eta}_Z^{(i)}, \mathbf{H}_{XZ}^{(i)})$, are given by

$$\begin{aligned} \boldsymbol{\eta}_X^{(i)} &= \mathbf{s}_X(X^{(i)}), \\ \boldsymbol{\eta}_Z^{(i)} &= \boldsymbol{\tau}_Z(\boldsymbol{\theta}_Z + \mathbf{s}_X(X^{(i)}) \cdot \boldsymbol{\Theta}_{XZ}), \\ \mathbf{H}_{XZ}^{(i)} &= \mathbf{s}_X(X^{(i)}) \otimes \boldsymbol{\eta}_Z^{(i)}, \end{aligned} \tag{25}$$

for every i, where $\boldsymbol{\tau}_Z$ is the forward mapping of $\mathcal{M}_Z$.

The objective of the M-Step of EM for an exponential family LVM is to minimize the cross-entropy loss (Eq. 23) of the joint density q_{XZ}, where we use the conditional sufficient statistics of Equation 24 to fill in the missing data. The M-Step objective function is thus

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\eta}_{XZ}^{(i)}, \boldsymbol{\theta}_{XZ}) = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\eta}_{XZ}^{(i)} \cdot \boldsymbol{\theta}_{XZ} - \psi_{XZ}(\boldsymbol{\theta}_{XZ}). \tag{26}$$

This is again a convex optimization problem, and its solution is given by the backward mapping, so that

$$\arg\min_{\boldsymbol{\theta}_{XZ}} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\eta}_{XZ}^{(i)}, \boldsymbol{\theta}_{XZ}) = \boldsymbol{\tau}_{XZ}^{-1}(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\eta}_{XZ}^{(i)}), \tag{27}$$

where $\boldsymbol{\tau}_{XZ}^{-1}$ is the backward mapping of $\mathcal{M}_{XZ}$ (or $\mathcal{H}_{XZ}$ in the harmonium case). The EM algorithm thus minimizes the cross-entropy by iteratively updating the model parameters with Equations 24 and 27. As such, we can implement exact EM for a harmonium $\mathcal{H}_{XZ}$ as long as we can evaluate $\boldsymbol{\tau}_Z$ and $\boldsymbol{\tau}_{XZ}^{-1}$. This is in fact the case for mixture models and linear Gaussian models, but certainly not in general, and so we must derive additional approaches for when the harmonium is less analytically tractable.

3.3.2 Gradient Descent

The tractability of exact EM for a harmonium $\mathcal{H}_{XZ}$ reduces to the tractability of the forward mapping $\boldsymbol{\tau}_Z$ and the backward mapping $\boldsymbol{\tau}_{XZ}^{-1}$, and it is often the case that while $\boldsymbol{\tau}_Z$ is tractable, $\boldsymbol{\tau}_{XZ}^{-1}$ is not. An alternative training algorithm that avoids computing $\boldsymbol{\tau}_{XZ}^{-1}$ is cross-entropy gradient descent (CE-GD). To implement CE-GD for $q_{XZ} \in \mathcal{H}_{XZ}$ with natural parameters $(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})$, we first compute the conditional expectations (Eq. 25), and then the model expectations $(\boldsymbol{\eta}_X, \boldsymbol{\eta}_Z, \mathbf{H}_{XZ}) = \boldsymbol{\tau}_{XZ}(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XZ})$ using the forward mapping

Algorithm	**EM**	**CE-GD**	**EM-GD**	**CE-MCGD**	**EM-MCGD**
Condition	$\boldsymbol{\tau}_Z$, $\boldsymbol{\tau}_{XZ}^{-1}$	$\boldsymbol{\tau}_Z$, $\boldsymbol{\tau}_{XZ}$	$\boldsymbol{\tau}_Z$, $\boldsymbol{\tau}_{XZ}$	$Z \sim q_Z$, $X \sim q_{X\|Z}$	$Z \sim q_Z$, $X \sim q_{X\|Z}$

Table 1: *Necessary computations for proposed algorithms*

$\boldsymbol{\tau}_{XZ}$. The gradients of $\mathcal{L}(X^{(i)}, \boldsymbol{\theta}_{XZ})$ with respect to the natural parameters of q_{XZ} are then given by

$$\begin{aligned} \partial_{\boldsymbol{\theta}_X} \mathcal{L}(X_{XZ}^{(i)}, \boldsymbol{\theta}_{XZ}) &= \boldsymbol{\eta}_X - \boldsymbol{\eta}_X^{(i)}, \\ \partial_{\boldsymbol{\theta}_Z} \mathcal{L}(X_{XZ}^{(i)}, \boldsymbol{\theta}_{XZ}) &= \boldsymbol{\eta}_Z - \boldsymbol{\eta}_Z^{(i)}, \\ \partial_{\boldsymbol{\Theta}_{XZ}} \mathcal{L}(X_{XZ}^{(i)}, \boldsymbol{\theta}_{XZ}) &= \mathbf{H}_{XZ} - \mathbf{H}_{XZ}^{(i)}. \end{aligned} \tag{28}$$

Using these gradients we may then iteratively update the natural parameters of $\boldsymbol{\theta}_X$, $\boldsymbol{\theta}_Z$, and $\boldsymbol{\Theta}_{XZ}$, using any standard gradient pursuit algorithm such as Adam (Kingma and Ba, 2014), and by following the average gradient over using a batch strategy.

Both EM and CE-GD rely on Equations 24, and differ in either solving the M-Step with the backward mapping (Eq. 27) or following the cross-entropy gradients (Eq. 28). Yet the M-Step is a convex optimization problem, and we could also approximate its solution with gradient descent. We thus define expectation-maximization gradient descent (EM-GD) as the algorithm that approximates the M-Step by pursuing the cross-entropy gradients in Equations 28 and recomputing $\boldsymbol{\eta}_{XZ}$ at every step, while holding $\boldsymbol{\eta}_{XZ}^{(i)}$ fixed. After convergence to the M-Step optimum, we recompute $\boldsymbol{\eta}_{XZ}^{(i)}$ and begin another iteration of EM-GD. In Table 1 we summarize the requisite computations for the algorithms we propose.

Both CE-GD and EM-GD can effectively train LVMs while avoiding the evaluation of the backward mapping $\boldsymbol{\tau}_{XZ}^{-1}$. On one hand, CE-GD descends the cross-entropy objective directly, and avoids excessive computations that might arise for EM-GD while near the optimum of an M-Step. On the other hand, EM-GD can reuse the conditional expectations $\boldsymbol{\eta}_{XZ}^{(i)}$ over multiple gradient steps, which may be desirable if they are computationally expensive to evaluate. In addition, EM and EM-GD solve a sequence of convex optimization problems, rather than solving a single non-linear optimization problem like CE-GD, which can in practice improve the stability of the learning algorithm.

3.3.3 Monte Carlo Estimation

For cases where we cannot compute the conditional or model expectations of the sufficient statistics analytically, we can often estimate them through sampling and Monte Carlo methods. For many LVMs, Markov chain algorithms like Gibbs sampling can generate approximate samples that we can then use to estimate the necessary expectations. Gibbs sampling, however, can be prohibitively slow, and algorithms such as contrastive divergence (Hinton, 2002; Welling et al., 2005) were developed to approximate the cross-entropy gradient while avoiding full Gibbs sampling. Nevertheless, even contrastive divergence can still necessitate large numbers of sampling cycles, and its convergence properties can be difficult to characterize (Bengio and Delalleau, 2009; Sutskever and Tieleman, 2010).

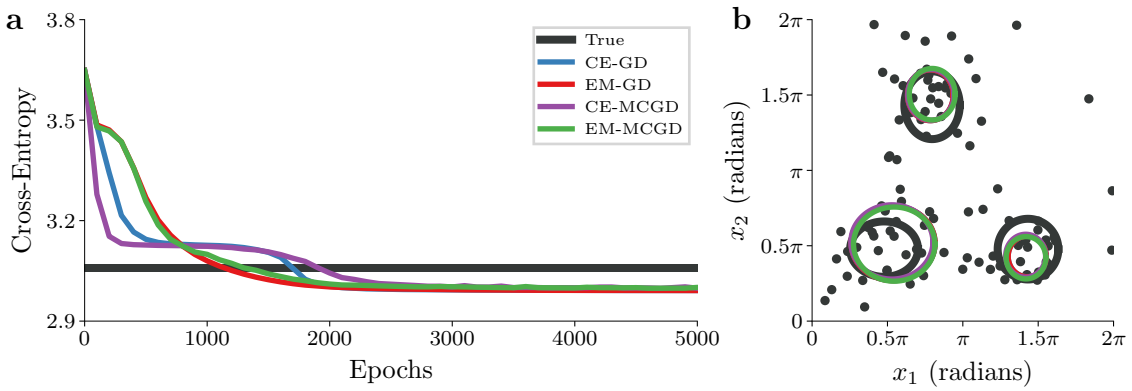

Figure 5: *Training strategies for mixtures of von Mises densities.* **b:** True cross-entropy of a sample (black line) and model cross-entropies (coloured lines) over training epochs. **b:** Sample (black dots, $n = 100$) and precision ellipses (black lines) from the ground-truth von Mises mixture, and learned precision ellipses (coloured lines) for models trained with CE-GD (red), EM-GD (green), CE-MCGD (yellow), and EM-MCGD (blue).

In contrast, we can directly generate exact samples from conjugated harmoniums. Consider a harmonium $\mathcal{H}_{XZ}$ defined by $\mathcal{M}_X$ and $\mathcal{M}_Z$, and suppose that there are efficient algorithms for sampling from the densities in $\mathcal{M}_X$ and $\mathcal{M}_Z$. If $q_{XZ} \in \mathcal{H}_{XZ}$ is a conjugated harmonium density, then its prior $q_Z \in \mathcal{M}_Z$, and we can generate an exact sample point $(X, Z) \sim q_{XZ}$ by sampling $Z \sim q_Z$ followed by $X \sim q_{X|Z=Z}$. We thus propose cross-entropy Monte Carlo gradient descent (CE-MCGD) and EM Monte Carlo gradient descent (EM-MCGD), which estimates the gradients in Equations 28 by using the sample estimators $\tilde{\boldsymbol{\eta}}_{XZ} = \sum_{j=1}^{m} \mathbf{s}_{XZ}(X^{(j)}, Z^{(j)})$ with m exact model samples $(X^{(j)}, Z^{(j)}) \sim q_{XZ}$, and estimating the conditional expectations $\boldsymbol{\eta}_{XZ}^{(i)} = \sum_{j=1}^{l} \mathbf{s}_{XZ}(X^{(i)}, Z^{(j)})$ over l conditional samples $Z^{(j)} \sim q_{Z|X=X^{(i)}}$ for every $X^{(i)}$.

We demonstrate the CE-GD, EM-GD, CE-MCGD, and EM-MCGD algorithms by training them on 100 sample points from a ground-truth mixture of 2D von Mises densities, where each component is a product of two von Mises distributions (Fig. 5). For CE-GD and EM-GD we compute the average gradient over the entire sample, so that each gradient step corresponds to one epoch. For CE-MCGD and EM-MCGD, we generated $m = 10$ model samples, and $l = 1$ conditional samples per training point $X^{(i)}$, and we use a batch size of 10 so that 10 gradient steps corresponds to one epoch. Finally, for the EM-based algorithms, we hold the conditional sufficient statistics $\boldsymbol{\eta}_{XZ}^{(i)}$ fixed for 100 epochs before recomputing them. We find that, overall, all algorithms can perform well, and the learned density converges to a good approximation of the ground-truth model.

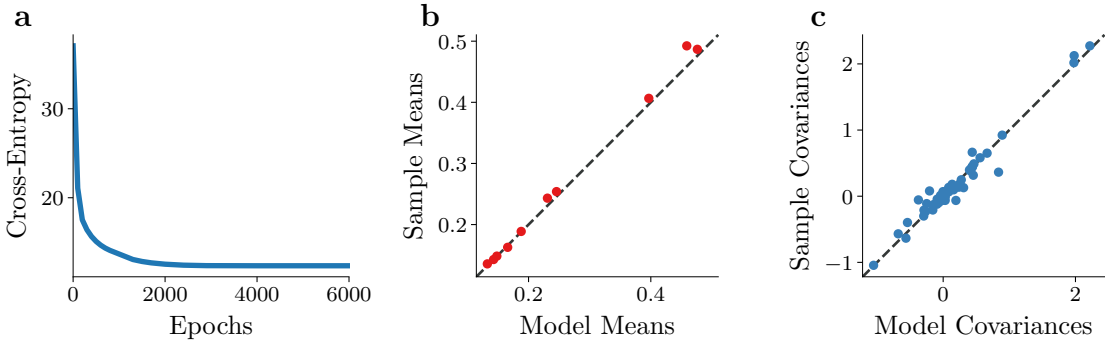

Figure 6: *A factor analysis analogue for count-data.* **a:** Cross-entropy descent of a CoM-based mixture model trained on synthetic multivariate count data. **b-c:** Means (**b**) and covariances (**c**) learned by the model compared to the sample estimators.

3.3.4 Linear Subspaces

We highlight one final algorithmic trick for training a harmonium while restricting the solution to a linear subspace. Consider the harmonium $\mathcal{H}_{XZ}$ defined by the exponential families $\mathcal{M}_X$ and $\mathcal{M}_Z$, and suppose $\mathcal{H}'_{XZ}$ is the harmonium defined by the sufficient statistic $\mathbf{s}'_{XZ} = \mathbf{A} \cdot \mathbf{s}_{XZ}$ for some matrix $\mathbf{A}$ — to ensure that $\mathcal{H}_{XZ}$ is minimal, $\mathbf{A}$ should have full rank and at least as many columns as rows. In this case if $q'_{XZ} \in \mathcal{H}'_{XZ}$ with parameters $\boldsymbol{\theta}'_{XZ}$, then $q'_{XZ} \in \mathcal{M}_{XZ}$ with parameters $\boldsymbol{\theta}'_{XZ} \cdot \mathbf{A}$. We can thus compute the mean parameters $\boldsymbol{\eta}'_{XZ}$ of q'_{XZ} by first evaluating (or estimating) $\boldsymbol{\eta}_{XZ} = \boldsymbol{\tau}(\boldsymbol{\theta}'_{XZ} \cdot \mathbf{A})$, and then using the linearity of expectations to compute $\boldsymbol{\eta}'_{XZ} = \boldsymbol{\tau}'_{XZ}(\boldsymbol{\theta}'_{XZ}) = \mathbf{A} \cdot \boldsymbol{\eta}_{XZ}$, where $\boldsymbol{\tau}_{XZ}$ and $\boldsymbol{\tau}'_{XZ}$ are the forward mappings of $\mathcal{H}_{XZ}$ and $\mathcal{H}'_{XZ}$, respectively. Consequently, for any algorithm in Table 1 except EM, if we can use it to train $\mathcal{H}_{XZ}$, we can use it to train $\mathcal{H}'_{XZ}$. This technique is obviously not relevant when we can compute the expectations in the subspace directly — for example, the expectations of FA can be computed much more efficiently than the expectations of LGMs. Yet we could use this technique to train a Boltzmann machine while restricting the interactions to e.g. neighbouring neurons in a lattice.

Another example for this technique is the so called CoM-based mixture model for modelling multivariate count-data (Sokoloski et al., 2021). A Conway-Maxwell (CoM) Poisson distribution is an exponential family count distribution with distinct location and shape parameters (Shmueli et al., 2005; Stevenson, 2016), and CoM-based mixture models mix products of CoM-Poisson distributions. The parameters of the model are restricted in a manner similar to FA, so that the latent category K only modulated the location, and not the shape of the CoM-Poisson distributions. In contrast with FA, there is no efficient method for computing the expectations on this restricted parameter space, and so they are instead computed in the unrestricted parameter space. Here we train a CoM-based mixture model on synthetic multivariate count data using the EM-GD algorithm (Fig. 6**a**), and show that it captures the sample means (Fig. 6**b**) and covariances (Fig. 6**c**) of the count data.

3.4 Conjugation and Harmonium Graphical Models

In this section we extend the theory of conjugated harmonium to harmonium graphical models. We begin with the simple case of multiple independent observations of a latent variable (Fig. 7**a**), and derive a general algorithm for recursive Bayesian inference with conjugating likelihoods. We then extend the Conjugation Lemma (Thm. 4) to account for conditional independences between variables of an HGM (Fig. 7**b**), and derive a general training algorithm for appropriately conjugated HGMs. We conclude with two examples of HGMs with a hierarchical structure.

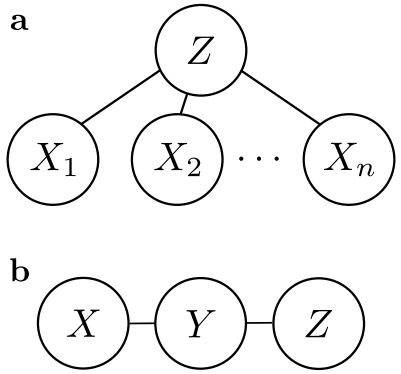

Figure 7: *Prototype graphs*

3.4.1 Conjugation and Recursive Inference

Consider a sequence of observable variables $X_1, \ldots, X_n$ that are conditionally independent given the latent variable Z, such that the conditional density of the observable variables given the latent variable is $p_{X_1,\ldots,X_n|Z} = \prod_{i=1}^n p_{X_i|Z}$. Then $p_{X_1,\ldots,X_n|Z}$ can be expressed in the recursive form

$$p_{Z|X_1,\ldots,X_n}(z \mid x_1, \ldots, x_n) \propto p_{X_n|Z}(x_n \mid z) p_{Z|X_1,\ldots,X_{n-1}}(z \mid x_1, \ldots, x_{n-1}), \tag{29}$$

where for the base-case $n = 1$, we define $p_{Z|X_1,\ldots,X_{n-1}}$ as the prior p_Z. Observe that this recursive relation is a single application of Bayes' rule (Eq. 15) given the prior $p_{Z|X_1,\ldots,X_{n-1}}$, so that Equation 29 reduce sequential inference to iterative applications of Bayes' rule.

In general, however, each application of Bayes' rule may increase the complexity of the beliefs, and ultimately produce an intractable posterior. We may avoid this by assuming assume that there is a single exponential family $\mathcal{M}_Z$ that is a conjugate prior family for each likelihood $p_{X_i|Z}$, which ultimately ensures that the posterior over Z given all the observations is also an element of $\mathcal{M}_Z$.

Theorem 16 *Suppose that the sequence of observable variables $X_1, \ldots, X_n$ are conditionally independent given the latent variable Z, and that each $p_{X_i|Z} \in \mathcal{M}_X$ is a conjugating likelihood in the exponential family $\mathcal{M}_X$ with natural parameters $\boldsymbol{\theta}_{X_i}$ and $\boldsymbol{\Theta}_{X_iZ}$, and conjugation parameters $\boldsymbol{\rho}_i$ and χ_i. Moreover, suppose $\mathcal{M}_Z$ is a conjugate prior family or each $p_{X_i|Z}$, and that $p_Z \in \mathcal{M}_Z$ with parameters $\boldsymbol{\theta}^*_Z$. Then the parameters $\boldsymbol{\theta}_{Z|X_1,\ldots,X_n}$ of the posterior $p_{Z|X_1,\ldots,X_n} \in \mathcal{M}_Z$ are given by*

$$\boldsymbol{\theta}_{Z|X_1,\ldots,X_n}(x_1, \ldots, x_n) = \boldsymbol{\theta}^*_Z + \sum_{i=1}^n \mathbf{s}_{X_i}(x_i) \cdot \boldsymbol{\Theta}_{X_iZ} - \boldsymbol{\rho}_i. \tag{30}$$

Proof *For $n = 1$, the prior is p_Z and its parameters are $\boldsymbol{\theta}^*_Z$. Given the likelihood $p_{X_1|Z}$ with parameters $\boldsymbol{\theta}_{X_1}$ and $\boldsymbol{\Theta}_{X_1Z}$, Corollary 10 implies that the parameters of $p_{Z|X_1}(x_1)$ are $\boldsymbol{\theta}_{Z|X_1} = \boldsymbol{\theta}^*_Z + \mathbf{s}_{X_1}(x_1) \cdot \boldsymbol{\Theta}_{X_1Z} - \boldsymbol{\rho}_1$.*

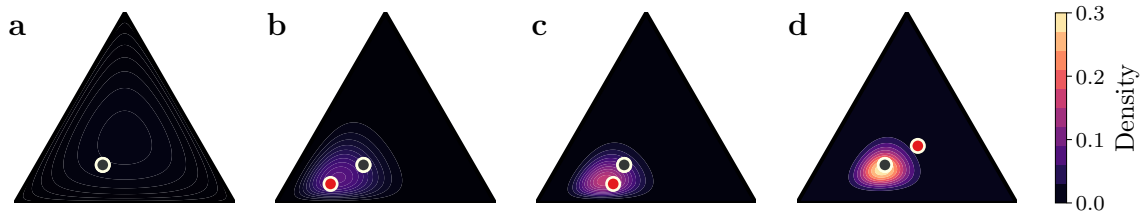

Figure 8: *Inference over a categorical distribution.* **a:** The true weights of a categorical distribution (black dot) and the Dirichlet prior density (contours). **b-d:** The true weights of a categorical distribution (black dot) and the Dirichlet posterior density (contours) after 10 (**b**), 20 (**c**), and 30 (**d**) observations, as well as the centre of mass of each set of 10 observations (red dots).

For n assume that the parameters $\boldsymbol{\theta}_{Z|X_1,\dots,X_n}$ of $p_{Z|X_1,\dots,X_n}$ are given by Equation 30. Where the likelihood $p_{X_{n+1}|Z}$ has parameters $\boldsymbol{\theta}_{X_{n+1}}$ and $\boldsymbol{\Theta}_{X_{n+1}Z}$, Corollary 10 implies that the parameters of $p_{Z|X_1,\dots,X_{n+1}}(x_1,\dots,x_{n+1})$ are

$$\begin{aligned}\boldsymbol{\theta}_{Z|X_1,\dots,X_{n+1}} &= \boldsymbol{\theta}_{Z|X_1,\dots,X_n}(x_1,\dots,x_n) + \mathbf{s}_{X_{n+1}}(x_{n+1}) \cdot \boldsymbol{\Theta}_{X_{n+1}Z} - \boldsymbol{\rho}_{n+1} \\ &= \boldsymbol{\theta}^*_Z + \sum_{i=1}^{n+1} \mathbf{s}_{X_i}(x_i) \cdot \boldsymbol{\Theta}_{X_iZ} - \boldsymbol{\rho}_i.\end{aligned}$$

Therefore, by induction, Equation 30 holds for any n. ■

By combining Theorems 15 and 16 we may conclude that the posterior for Bayesian parameter estimation has parameters $\boldsymbol{\theta}_{Z|X_1,\dots,X_n}(x_1,\dots,x_n) = \boldsymbol{\theta}_Z + \sum_{i=1}^n (\mathbf{s}_{X_i}(x_i), 1)$, which is indeed the formula for the parameters of a conjugate exponential family posterior (Murphy, 2023). As a more practical example, suppose that $\mathcal{H}_{XZ}$ is the harmonium defined by the categorical family $\mathcal{M}_X$ and the Dirichlet family $\mathcal{M}_Z$. It is easy to check that $q_{XZ} \in \mathcal{H}_{XZ}$ is conjugated if $\boldsymbol{\theta}_X = \mathbf{0}$ and

$$\boldsymbol{\Theta}_{XZ} = \begin{bmatrix} -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & \cdots & 1 \end{bmatrix} \tag{31}$$

with conjugation parameters $\boldsymbol{\rho} = (-1, 0, \dots, 0)$ and $\chi = 0$. By implementing this model, we find that the posterior Dirichlet density concentrates around the true weights of categorical distribution given an increasingly large sequence of observations (Fig. 8).

3.4.2 Conjugation at the Latent Variable Boundary

In this section we extend the theory of conjugated harmoniums to HGMs. Our strategy is to reformulate them as bivariate harmoniums over observable and latent variables, so that we can directly apply our theorems from Section 3.1. Extending the theory of conjugated

harmoniums will then prove straightforward, but reformulating HGMs as bivariate harmoniums without losing the graphical model structure necessitates redeveloping some notation and definitions.

To begin, consider the HGM $\mathcal{H}_Y^G$ on the random variables $Y_1, \ldots, Y_{n+m}$ defined by the exponential families $\mathcal{M}_{Y_1}, \ldots, \mathcal{M}_{Y_{n+m}}$, and suppose its graph $G = ((V, W), E)$ has vertices (V, W) such that $|V| = n$ and $|W| = m$. Let us refer to the variables index by V and W as the observable variables and latent variables, and denote them by $X_v = Y_v$ and $Z_w = Y_w$ for any $v \in V$ and $w \in W$, respectively (Fig. 9). Moreover, let

$$B = \{v \in V : w \in W \text{ and } \{v, w\} \in E\} \cup \{w \in W : v \in V \text{ and } \{v, w\} \in E\} \quad (32)$$

denote all vertices in the boundary between X and Z. Finally, where $G[A] = (A, \{\{v, w\} \in E : v, w \in A\})$ indicates the subgraph of G induced by A, let $G_X = G[V]$, $G_Z = G[W]$, and $G_B = G[B]$ be the subgraphs of G induced by V, W, and B, respectively.

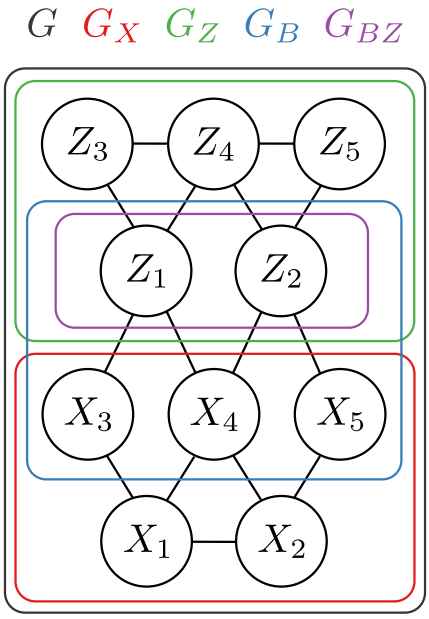

Figure 9: *HGM subgraphs*

We next define, for every clique $C \in \mathcal{C}(G)$, the clique-wise sufficient statistics of the observable variables $\mathbf{s}_X^C = (\bigotimes_{v \in (C \cap V)} \mathbf{s}_{X_v})$ and the latent variables $\mathbf{s}_Z^C = (\bigotimes_{w \in (C \cap W)} \mathbf{s}_{Z_w})$, and assume that v and w are always drawn in ascending order. We define $\mathbf{s}_X^G = (\mathbf{s}_X^C)_{C \in \mathcal{C}(G_X)}$ and $\mathbf{s}_Z^G = (\mathbf{s}_Z^C)_{C \in \mathcal{C}(G_Z)}$ as the graph-wide sufficient statistics of the observable and latent variables, respectively, and assume that cliques C are drawn in lexicographic order. Given these definitions we may re-express any HGM density $q_Y \in \mathcal{H}_Y^G$ as a latent variable HGM density

$$q_{XZ}(\mathbf{x}, \mathbf{z}) \propto e^{\boldsymbol{\theta}_X^G \cdot \mathbf{s}_X^G(\mathbf{x}) + \boldsymbol{\theta}_Z^G \cdot \mathbf{s}_Z^G(\mathbf{z}) + \sum_{C \in \mathcal{C}(G_B)} \mathbf{s}_X^C(\mathbf{x}_C) \cdot \boldsymbol{\Theta}_{XZ}^C \cdot \mathbf{s}_Z^C(\mathbf{z}_C)}, \quad (33)$$

where $\mathbf{x}_C$ and $\mathbf{z}_C$ are the visible and latent variables indexed by clique C, respectively, $\boldsymbol{\theta}_X^G$ and $\boldsymbol{\theta}_Z^G$ are the observable and latent biases, respectively, and each $\boldsymbol{\Theta}_{XZ}^C$ is a clique-wise interaction matrix. We have thus reorganized all the multilinear interactions in the definition of the HGM density (Eq. 8) into vectors and matrices.

We may then express the likelihood of q_{XZ} by

$$q_{X|Z}(\mathbf{x} \mid \mathbf{z}) = q_{X|Z_B}(\mathbf{x} \mid \mathbf{z}_B) \propto e^{\mathbf{s}_X^G(\mathbf{x}) \cdot (\boldsymbol{\theta}_X^G + \sum_{C \in \mathcal{C}(G_B)} \mathbf{I}_X^C \cdot \boldsymbol{\Theta}_{XZ}^C \cdot \mathbf{s}_Z^C(\mathbf{z}_C))}, \quad (34)$$

and its posterior by

$$q_{Z|X}(\mathbf{z} \mid \mathbf{x}) = q_{Z|X_B}(\mathbf{z} \mid \mathbf{x}_B) \propto e^{\mathbf{s}_Z^G(\mathbf{z}) \cdot (\boldsymbol{\theta}_Z^G + \sum_{C \in \mathcal{C}(G_B)} \mathbf{s}_X^C(\mathbf{x}_C) \cdot \boldsymbol{\Theta}_{XZ}^C \cdot \mathbf{I}_Z^C)}, \quad (35)$$

where $\mathbf{x}_B$ and $\mathbf{z}_B$ are the observable and latent variables indexed by the boundary B, and $\mathbf{I}_X^C$ and $\mathbf{I}_Z^C$ are wide and tall binary matrices that realign the indices of the natural parameters $\boldsymbol{\Theta}_{XZ}^C \cdot \mathbf{s}_Z^C(\mathbf{z})$ and $\mathbf{s}_X^C(\mathbf{x}) \cdot \boldsymbol{\Theta}_{XZ}^C$ with their corresponding elements in $\boldsymbol{\theta}_X^G$ and $\boldsymbol{\theta}_Z^G$, respectively. With these definitions in place, let us formally define a latent variable HGM.

Definition 17 (Latent Variable HGM) *Let G be a graph with vertices (V, W) that indexes the observable and latent variables $X = (X_1, \ldots, X_n)$ and $Z = (Z_1, \ldots, Z_m)$, and suppose $\mathcal{H}^G_X$ and $\mathcal{H}^G_Z$ are HGMs over X and Z defined by the sufficient statistics $\mathbf{s}^G_X$ and $\mathbf{s}^G_Z$, respectively. Then the latent variable HGM $\mathcal{H}^G_{XZ}$ defined by $\mathcal{H}^G_X$ and $\mathcal{H}^G_Z$ is the set of all densities with the form of Equation 33.*

Observe that we can also define the bivariate harmonium $\mathcal{H}_{XZ}$ with component exponential families $\mathcal{H}^G_X$ and $\mathcal{H}^G_Z$, that models all interactions $\boldsymbol{\Theta}_{XZ}$ between $\mathbf{s}^G_X$ and $\mathbf{s}^G_Z$. The latent variable HGM $\mathcal{H}^G_{XZ}$ is therefore a submanifold of $\mathcal{H}_{XZ}$, where the complete interaction matrix is decomposed into a sum of clique-wise interaction matrices, such that $\boldsymbol{\Theta}_{XZ} = \sum_{C \in \mathcal{C}(G_B)} \mathbf{I}^C_X \cdot \boldsymbol{\Theta}^C_{XZ} \cdot \mathbf{I}^C_Z$. Through this equivalence between bivariate harmoniums and latent variable HGMs, we may now extend the Conjugation Lemma (Thm. 4).

Theorem 18 *Suppose that $\mathcal{H}^G_{XZ}$ is a latent variable HGM defined by the graph G, and the minimal HGMs $\mathcal{M}^G_X$ and $\mathcal{M}^G_Z$. Then the density $q_{XZ} \in \mathcal{H}^G_{XZ}$ with parameters $\boldsymbol{\theta}^G_X$, $\boldsymbol{\theta}^G_Z$, and $(\boldsymbol{\Theta}^C_{XZ})_{C \in \mathcal{C}(G_B)}$ is conjugated if and only if there exists vectors $\{\boldsymbol{\rho}^C\}_{C \in \mathcal{C}(G_{BZ})}$ and a scalar χ such that*

$$\psi_X(\boldsymbol{\theta}^G_X + \sum_{C \in \mathcal{C}(G_B)} \mathbf{I}^C_X \cdot \boldsymbol{\Theta}^C_{XZ} \cdot \mathbf{s}^C_Z(\mathbf{z}_C)) = \chi + \sum_{C \in \mathcal{C}(G_{BZ})} \mathbf{s}^C_Z(\mathbf{z}_C) \cdot \boldsymbol{\rho}^C, \tag{36}$$

for any $\mathbf{z} \in \mathcal{Z}$, where $G_{BZ} = G[B \cap W]$ is the subgraph of G induced by the latent variables on the boundary (Fig. 9).

Proof *Firstly, let $q_{XZ} \in \mathcal{H}^G_{XZ}$, and suppose $\mathcal{H}_{XZ}$ is the harmonium defined by $\mathcal{H}^G_X$ and $\mathcal{H}^G_Z$. Then $q_{XZ} \in \mathcal{H}_{XZ}$ with parameters $\boldsymbol{\theta}^G_X$, $\boldsymbol{\theta}^G_Z$, and $\boldsymbol{\Theta}_{XZ} = \sum_{C \in \mathcal{C}(G_B)} \mathbf{I}^C_X \cdot \boldsymbol{\Theta}^C_{XZ} \cdot \mathbf{I}^C_Z$. By the Conjugation Lemma (Thm. 4), q_{XZ} is conjugated if an only if there exists $\boldsymbol{\rho}^G$ such that*

$$\psi_X(\boldsymbol{\theta}^G_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}^G_Z(\mathbf{z})) = \chi + \mathbf{s}^G_Z(\mathbf{z}) \cdot \boldsymbol{\rho}^G. \tag{37}$$

By substitution we can rewrite the LHS of this equation as

$$\psi_X(\boldsymbol{\theta}^G_X + \boldsymbol{\Theta}_{XZ} \cdot \mathbf{s}^G_Z(\mathbf{z})) = \psi_X(\boldsymbol{\theta}^G_X + \sum_{C \in \mathcal{C}(G_B)} \mathbf{I}^C_X \cdot \boldsymbol{\Theta}^C_{XZ} \cdot \mathbf{s}^C_Z(\mathbf{z}_C)). \tag{38}$$

On the other hand, we can rewrite the dot product on the RHS of Equation 37 as

$$\mathbf{s}^G_Z(\mathbf{z}) \cdot \boldsymbol{\rho}^G = \sum_{C \in \mathcal{C}(G_Z)} \mathbf{s}^C_Z(\mathbf{z}_C) \cdot \boldsymbol{\rho}^C = \sum_{C \in \mathcal{C}(G_{BZ})} \mathbf{s}^C_Z(\mathbf{z}_C) \cdot \boldsymbol{\rho}^C + \sum_{C \in (\mathcal{C}(G_Z) \setminus \mathcal{C}(G_{BZ}))} \mathbf{s}^C_Z(\mathbf{z}_C) \cdot \boldsymbol{\rho}^C. \tag{39}$$

Now, if $C \notin \mathcal{C}(G_{BZ})$, then $C \notin \mathcal{C}(G_B)$, and $\mathbf{s}^C_Z$ does not appear on the RHS of Equation 38. Therefore $\mathbf{s}^C_Z \cdot \boldsymbol{\rho}^C$ is constant, and $\boldsymbol{\rho}^C = \mathbf{0}$ due to the minimality of $\mathbf{s}^C_Z$. Since it only sums over $C \notin \mathcal{C}(G_B)$, the second term on the RHS of Equation 39 is 0. Therefore by substituting

$$\mathbf{s}^G_Z(\mathbf{z}) \cdot \boldsymbol{\rho}^G = \sum_{C \in \mathcal{C}(G_{BZ})} \mathbf{s}^C_Z(\mathbf{z}_C) \cdot \boldsymbol{\rho}^C, \tag{40}$$

and the RHS of Equation 38 into Equation 37, we obtain Equation 36. ■

The key feature of this theorem is that conjugation depends only on the latent variables that border the observable variables. In practice this means that we can combine conjugating likelihoods with hierarchical priors of arbitrary depth. In particular, sequences of conjugating likelihoods can be composed into complex, latent variable HGMs as long as they locally satisfy Theorem 18.

3.4.3 Hierarchical Conjugated Harmoniums

In this section we develop recursive algorithms for latent variable HGMs with hierarchical structure. For conceptual and notational simplicity we develop our theory for a prototypical, three-level hierarchical model with observable variables X, and latent variables Y and Z, such that $q_{XYZ} = q_{X|Y} \cdot q_{Y|Z} \cdot q_Z$ (Fig. 7**b**). Nevertheless, the results we develop here can be trivially extended to deeper hierarchical models.

Let us first develop a coarse-grained representation of a hierarchical HGM. Consider the observable variables $X = (X_1, \ldots, X_n)$, and the latent variables $Y = (Y_1, \ldots, Y_m)$ and $Z = (Z_1, \ldots, Z_l)$. Suppose $G = ((U, V, W), E)$ is a graph with vertices (U, V, W) such that $|U| = n$, $|V| = m$, and $|W| = l$. Let $G_X = G[U]$, $G_Y = G[V]$, and $G_Z = G[W]$ denote the subgraphs of G induced by the vertices U, V, and W, respectively, and let B_{XY}, B_{XZ}, and B_{YZ} denote the boundaries between X and Y, X and Z, and Y and Z, respectively (Eq. 32). Let H^G_{YZ} be the HGM defined by $\mathcal{H}^G_Y$ and $\mathcal{H}^G_Z$, and let $H^G_{XYZ} = \mathcal{H}^G_{X(YZ)}$ be the HGM defined by $\mathcal{H}^G_X$ and $\mathcal{H}^G_{YZ}$. We say that $\mathcal{H}^G_{XYZ}$ is a hierarchical HGM if $B_{XZ} = \emptyset$. Where we suppress the notation indicating the graph G, we can express any density $q_{XYZ} \in \mathcal{H}^G_{XYZ}$ by

$$q_{XYZ}(\mathbf{x}, \mathbf{y}, \mathbf{z}) \propto e^{\boldsymbol{\theta}_X \cdot \mathbf{s}_X(\mathbf{x}) + \boldsymbol{\theta}_Y \cdot \mathbf{s}_Y(\mathbf{y}) + \boldsymbol{\theta}_Z \cdot \mathbf{s}_Z(\mathbf{z}) + \mathbf{s}_X(\mathbf{x}) \cdot \boldsymbol{\Theta}_{XY} \cdot \mathbf{s}_Y(\mathbf{y}) + \mathbf{s}_Y(\mathbf{y}) \cdot \boldsymbol{\Theta}_{YZ} \cdot \mathbf{s}_Z(\mathbf{z})}, \tag{41}$$

where $\mathbf{s}_X$, $\mathbf{s}_Y$, and $\mathbf{s}_Z$ are the sufficient statistics of $\mathcal{H}^G_X$, $\mathcal{H}^G_Y$, and $\mathcal{H}^G_Z$, respectively; $\boldsymbol{\theta}_X \in \mathcal{H}^G_X$, $\boldsymbol{\theta}_Y \in \mathcal{H}^G_Y$, and $\boldsymbol{\theta}_Z \in \mathcal{H}^G_Z$ are the biases over the subgraphs G_X, G_Y, and G_Z, respectively; and where $\boldsymbol{\Theta}_{XY} = \sum_{C \in \mathcal{C}(B_{XY})} \mathbf{I}^C_X \cdot \boldsymbol{\Theta}^C_{XY} \cdot \mathbf{I}^C_Y$ and $\boldsymbol{\Theta}_{YZ} = \sum_{C \in \mathcal{C}(B_{YZ})} \mathbf{I}^C_Y \cdot \boldsymbol{\Theta}^C_{YZ} \cdot \mathbf{I}^C_Z$ are the interaction matrices across the boundaries B_{XY} and B_{YZ}, respectively.

Now let us assume that $q_{XYZ} \in \mathcal{H}^G_{XYZ}$ is conjugated such that q_{YZ} is in the latent variable HGM $\mathcal{M}^G_{YZ}$ defined by $\mathcal{M}^G_Y$ and $\mathcal{M}^G_Z$. Then by Theorem 18, there exists conjugation parameters $\boldsymbol{\rho}_Y$ and χ_Y such that $\psi_X(\boldsymbol{\theta}_X + \boldsymbol{\Theta}_{XY} \cdot \mathbf{s}_Y(\mathbf{y})) = \chi_Y + \mathbf{s}_Y(\mathbf{y}) \cdot \boldsymbol{\rho}_Y$. We can put this equation in the form of the Conjugation Lemma (Thm. 4) by simply padding $\boldsymbol{\rho}_Y$ with zeroes, such that $\mathbf{s}_Y(y) \cdot \boldsymbol{\rho}_Y = \mathbf{s}_{YZ}(\mathbf{y}, \mathbf{z}) \cdot (\boldsymbol{\rho}_Y, \mathbf{0})$. The simple equation is the key to extending the results in Section 3.1 to hierarchical HGMs.

Corollary 19 *Suppose $\mathcal{H}^G_{XYZ}$ is a hierarchical HGM defined by the graph G and HGMs $\mathcal{H}^G_X$, $\mathcal{H}^G_Y$, and $\mathcal{H}^G_Z$, and that $q_{XYZ} \in \mathcal{H}^G_{XYZ}$ with parameters $(\boldsymbol{\theta}_X, \boldsymbol{\theta}_Y, \boldsymbol{\theta}_Z, \boldsymbol{\Theta}_{XY}, \boldsymbol{\Theta}_{YZ})$ is conjugated with conjugation parameters $\boldsymbol{\rho}_Y$ and χ_Y. Then the parameters of $q_{YZ} \in \mathcal{M}^G_{YZ}$ are given by*

$$\boldsymbol{\theta}^*_{YZ} = (\boldsymbol{\theta}_Y + \boldsymbol{\rho}_Y, \boldsymbol{\Theta}_{YZ}, \boldsymbol{\theta}_Z). \tag{42}$$

Marginalizing out the observable variables from q_{XYZ} can thus be reduced to translating the biases $\boldsymbol{\theta}_Y$ of the low-level latent variables Y of the hierarchical prior q_{YZ}. Of course, this does not ensure the tractability of hierarchical HGMs, because q_{YZ} may itself be intractable. However, when q_{YZ} is also conjugated such that $q_Z \in \mathcal{H}^G_Z$, we can derive recursive algorithms for sampling, density evaluation, and learning for hierarchical HGMs.

Definition 20 (Iteratively Conjugated Harmonium) *The density $q_{XYZ} \in \mathcal{H}^G_{XYZ}$ is iteratively conjugated if $q_{YZ} \in \mathcal{H}^G_{YZ}$ and $q_Z \in \mathcal{H}^G_Z$. The hierarchical HGM $\mathcal{H}^G_{XYZ}$ is iteratively conjugated if every $q_{XYZ} \in \mathcal{H}^G_{XYZ}$ is iteratively conjugated.*

Firstly, with regards to sampling, suppose $q_{XYZ} \in \mathcal{H}^G_{XYZ}$ is iteratively conjugated such that $\boldsymbol{\rho}_Y$ and χ_Y are the conjugation parameters of q_{XYZ}, and χ^*_Z and $\boldsymbol{\rho}^*_Z$ are the conjugation parameters of $q_{YZ} \in \mathcal{H}^G_{YZ}$. Then according to Corollary 5, q_Z has parameters $\boldsymbol{\theta}^*_Z = \boldsymbol{\theta}_Z + \boldsymbol{\rho}^*_Z$. Therefore, as long as we can tractably sample from $\mathcal{H}^G_X$, $\mathcal{H}^G_Y$, and $\mathcal{H}^G_Z$, we may sample $(X^{(i)}, Y^{(i)}, Z^{(i)})$ from q_{XYZ} by sampling $Z^{(i)} \sim q_Z$, $Y^{(i)} \sim q_{Y|Z=Z^{(i)}}$, and $X^{(i)} \sim q_{X|Y=Y^{(i)}}$.

To express the observable density q_X at $\mathbf{x}$ of a harmonium density q_{XYZ} in terms of conjugation parameters, we must assume that both the prior $q_{YZ} \in \mathcal{H}^G_{YZ}$ and posterior $q_{YZ|X} \in \mathcal{H}^G_{YZ}$ are conjugated harmonium densities. In this case, by extending the latent variable Z in Equation 14 to the hierarchical variables (Y, Z), and re-expressing the log-partition function ψ_{YZ} with Corollary 6, we may express the observable density q_X by

$$\log q_X(\mathbf{x}) = \mathbf{s}_X(\mathbf{x}) \cdot \boldsymbol{\theta}_X + \psi_Z(\boldsymbol{\theta}_Z + \boldsymbol{\rho}_Z(\mathbf{x})) - \psi_Z(\boldsymbol{\theta}_Z + \boldsymbol{\rho}^*_Z) + \chi_Z(\mathbf{x}) - \chi^*_Z - \chi_Y, \qquad (43)$$

where χ_Z and $\boldsymbol{\rho}_Z(\mathbf{x})$ are the conjugation parameters of $q_{YZ|X=\mathbf{x}}$. Finally, assuming we can sample from both q_{XYZ} and $q_{YZ|X}$, we can fit an iteratively conjugated harmonium to data by applying either the CE-MCGD or EM-MCGD training algorithms (Tbl. 1). We may thus implement tractable sampling, density evaluation, and learning in iteratively conjugated harmoniums as long as the component harmoniums $\mathcal{H}^G_{XY}$ and $\mathcal{H}^G_{YZ}$ are conjugated.

3.4.4 Examples of Hierarchical Harmoniums

We conclude with two examples of hierarchical HGM. In the first we demonstrate how to cluster high-dimensional data, and in the second we demonstrate how to embed and learn low-dimensional subspaces, or "neural manifolds", in high dimensional neural activity.

Many clustering algorithms, such as mixtures of Gaussians, suffer from the so-called "curse of dimensionality", and exhibit limited performance when classifying high-dimensional data (Beyer et al., 1999; Assent, 2012). A common approach for addressing this is to first apply dimensionality reduction techniques such as PCA or FA to project the data into a lower dimensional space, and then cluster the projected data. Such two-stage algorithms are widely applied in domains ranging from image processing (Houdard et al., 2018; Hertrich et al., 2022), to neuroscience (Lewicki, 1998; Baden et al., 2016), and to bioinformatics (Witten and Tibshirani, 2010; Duò et al., 2020). Nevertheless, suboptimal clusterings can arise when the dimensionality-reduction is implemented only as a preprocessing step. If we consider PCA as our dimensionality reduction technique, for example, the directions in the data that best separate the clusters might not be the directions of maximum variance that define the PCA projection (Chang, 1983; McLachlan et al., 2019).

Let us develop a hierarchical HGM for high-dimensional clustering that jointly defines the dimensionality-reduction and clustering models. In particular, consider a hierarchical HGM $\mathcal{H}_{XYZ}$ defined by the graph $G = (\{1\}, \{2\}, \{3\}, \{\{1, 2\}, \{2, 3\}\})$ (Fig. 7**b**), and the HGMs $\mathcal{H}^G_X$, $\mathcal{H}^G_Y$, and $\mathcal{H}^G_K$, where $\mathcal{H}^G_X$ is the family of n-variate multivariate-normals with diagonal covariance matrices, $\mathcal{H}^G_Y$ is the family of m-variate normals with full covariance matrices, and $\mathcal{H}^G_K$ is the categorical family of dimension d_K (Fig. 10). This is in fact a hierarchical form of

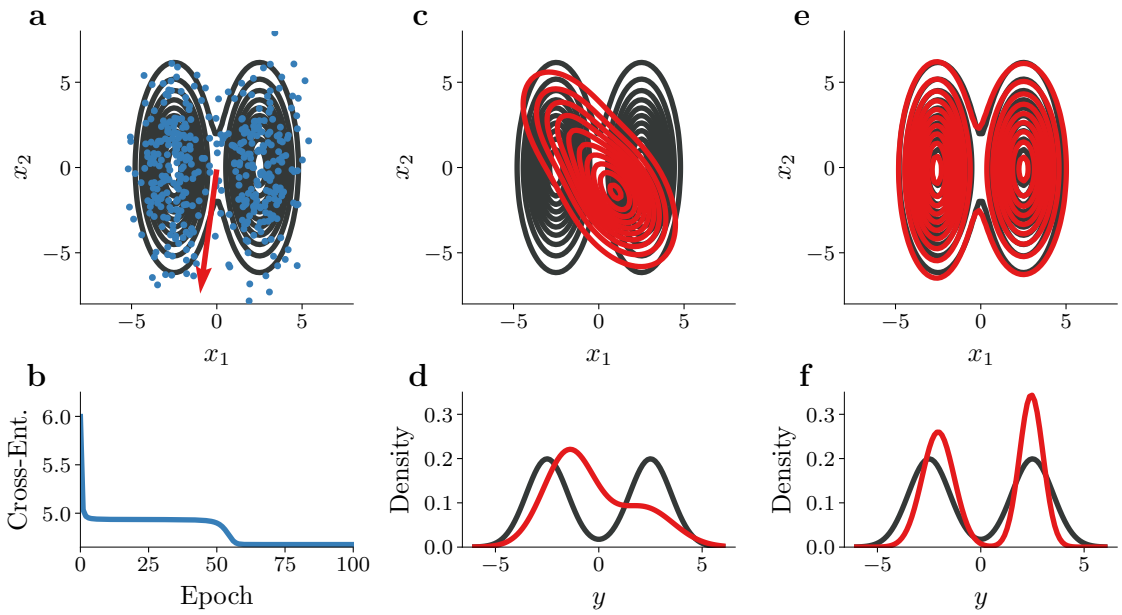

Figure 10: *A hierarchical mixture of Gaussian distributions.* **a:** Contours (black lines) of a ground-truth observable density p_X, a sample (blue dots) from p_X, and the direction of maximum variance (red arrow) of the sample. **b:** Cross-entropy (blue line) of the model density q_X over epochs. **c:** Contours of p_X (black lines) and the initial model density q_X. **d:** Ground-truth latent density p_Y (black line) and the latent density q_Y of the initial model. **e:** Contours of the ground-truth observable density p_X (black lines) and the learned q_X (red lines). **f:** Ground-truth latent density p_Y (black line) and the latent density q_Y of the learned model.

a mixture of factor analyzers (Ghahramani and Hinton, 1996; McNicholas and Murphy, 2008; Baek et al., 2010), where the latent space $\mathcal{Y}$ is shared across the factor analysis components.

Let us consider a toy example for which $n = 2$, $m = 1$, and $d_k = 1$ (Fig. 10). We choose a ground-truth density $p_{XYK} \in \mathcal{H}_{XYK}$ so that the direction of maximum variance is orthogonal to the direction along which class membership changes, and sample from the observable density p_X (Fig .10**a**). By applying EM we minimize the cross-entropy (Fig. 10**b**) of an initial density q_{XYK} with observable density q_X (Fig. 10**c**), and a mixture density in the latent space q_Y that does not match the ground-truth mixture density p_Y (Fig. 10**d**). After training we find that we recover both the ground-truth observable density (Fig. 10**e**) and latent mixture density (Fig. 10**f**).

In neuroscience, low-dimensional subspaces within high-dimensional neural activity are referred to as neural manifolds, and latent variables ranging from head-direction to spatial location have been found encoded as neural manifolds in recordings of key brain areas from various animals (Churchland et al., 2012; Cunningham and Yu, 2014; Langdon et al., 2023). Let us develop a model of neural population activity that encodes regions of interest in

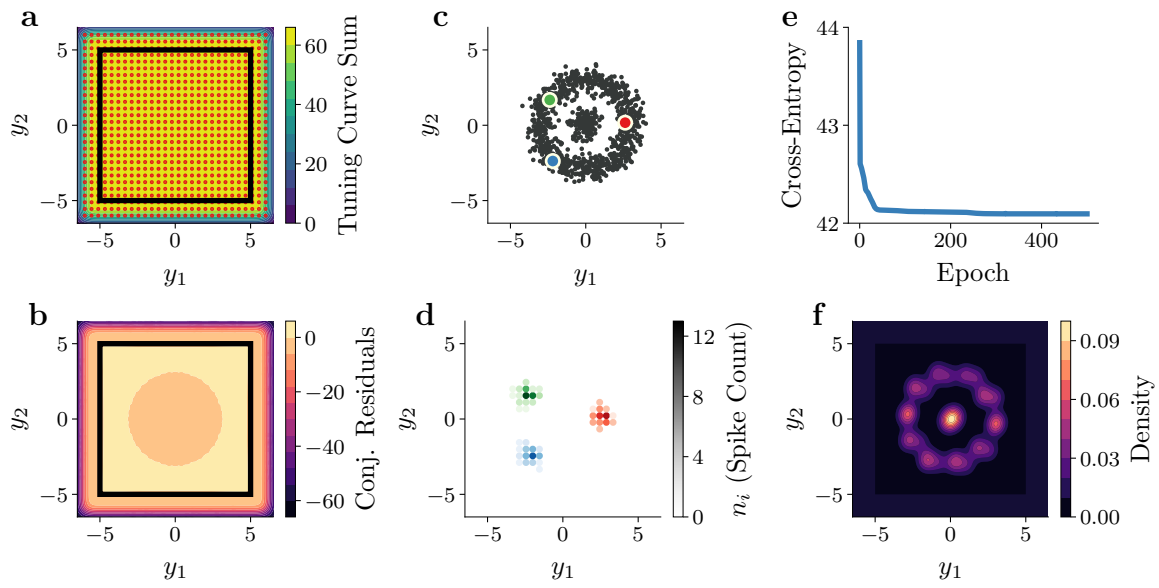

Figure 11: *A probabilistic population code with a Gaussian-Boltzmann machine prior.* **a:** Preferred stimuli (red dots) and the sum of the tuning curves (filled contour) of a 2D Gaussian-tuned Poisson population. **b:** Residuals of the linear regression of the conjugation parameters of the Poisson population. **c:** Sample $Y^{(1)}, \ldots, Y^{(1000)}$ (black dots) from noisy concentric circles, with three points highlighted (red, green, and blue dots). **d:** Spike response (coloured points) of the Poisson population to the example points in (**c**). **e:** Cross-entropy (blue line) of the model density q_N given sample $N^{(1)}, \ldots, N^{(1000)}$ over training epochs. **f:** Filled contours of the learned stimulus density q_Y.

a 2-dimensional continuous space. In particular, consider the hierarchical HGM $\mathcal{H}^G_{XYZ}$ specified by the prototypical hierarchical graph G (Fig. 7**b**), where $\mathcal{H}^G_X$ is a product 28×28 Poisson neurons, and $\mathcal{H}^G_{YZ}$ is a Gaussian-Boltzmann over a 2D surface with 6 hidden neurons (as depicted in Fig. 3). We choose $q_{N|Y}$ so that preferred stimuli of the tuning curves of the Poisson neurons tile tile a 2D surface, and thereby ensure that the sum of tuning curves over a square region is approximately constant (Fig. 11**a**). To verify that Equation 21 is approximately satisfied, we fit conjugation parameters to $\sum_{i=1}^{d_N} \mathbb{E}_Q[N_i \mid Y = y]$ with linear regression, and find that the residuals are nearly zero over the region (Fig. 11**b**).

To demonstrate how to encode a particular submanifold in the neural activity, we generate a noisy sample $Y^{(1)}, \ldots, Y^{(1000)}$ from two concentric rings in the stimulus space (Fig. 11**c**), and use those points to generate spiking observations $N^{(i)} \sim q_{N|Y=Y^{(i)}}$ in the 28×28 dimensional space of spike counts $\mathcal{N}$ (Fig. 11**d**). We then used EM-GD to minimize the cross-entropy of q_N given $N^{(1)}, \ldots, N^{(1000)}$ while keeping the parameters of $q_{N|Y}$ (i.e. the tuning curve parameters) fixed (Fig. 11**e**), and found that the model learns an accurate latent representation of the concentric circles (Fig. 11**f**).

4 Discussion

Bayesian inference is the most theoretically well-founded approach to inferring information about unknown quantities given observations (Jaynes, 2003). Similarly, learning a model through the method of maximum likelihood is grounded in an axiomatic characterization of how to minimize the divergence of a model from a target (i.e. empirical) distribution (Shore and Johnson, 1980). In this paper we developed a theory of exponential family latent variable models we call conjugated harmoniums, that afford both exact Bayesian inference and general algorithms for maximum-likelihood learning. In the best case we can implement learning with exact EM, yet even in the "worst" case we can generate exact model samples from conjugated harmoniums, and thereby maximize the likelihood of the model parameters with an estimation error that can be made arbitrary small.

Our theory thus illuminates the class of models that allow researchers to stay close to the theoretical foundations of LVMs. Moreover, the intrinsic, exponential family characterization of this class allows researchers to easily identify whether a model of interest lies within it. Yet in spite of our goal of avoiding approximation schemes, we see this work as complementary to methods for approximate inference and learning. Variational autoencoders, for example, typically approximate the intractable posterior with an exponential family (Shekhovtsov et al., 2021), and this exponential family need not be the Gaussian or categorical families, as is often assumed (Vahdat et al., 2020). Our theory thus provides numerous candidate models with sufficient tractability to model the latent space of a variational autoencoder.

LVMs also play a large role in computational neuroscience, where they are used to model how behavioural and environmental variables are encoded in low-dimensional submanifolds of neural activity (Cunningham and Yu, 2014; Schneidman, 2016; Langdon et al., 2023). Boltzmann machines and products of Poisson distributions are both examples of exponential families that are widely applied to modelling neural activity, and we showed how they both serve as effective components of conjugated harmoniums. Moreover, we showed how to assemble them into a hierarchical model for identifying neural manifolds. Yet the example we provided can be extended in numerous ways: (i) by restricting the correlation structure between the Boltzmann neurons to capture specific forms of neural manifold; (ii) by learning the likelihood of the model rather than tiling the stimulus with model neurons; and (iii) by replacing Gaussian tuning curves with von Mises tuning curves to capture periodic structure (Kutschireiter et al., 2023). We hope to explore these extensions in future work.

Our theory could also help answer a fundamental question in computational neuroscience: do neural circuits implement approximate inference and learning, or is neural connectivity constrained so that exact inference and learning are possible (Shivkumar et al., 2018; Lange et al., 2023)? By expanding the scope of neural models that exhibit exact inference and learning, we should be better able to isolate those neural circuits that can indeed avoid approximation schemes.

Another topic we aim to explore is the relationship between the framework of conjugated harmoniums and sequential data models such as Hidden Markov Models (HMMs) and linear state space models (LSSMs). Both HMMs and LSSMs afford exact inference and learning (Särkkä, 2013). Moreover, the component distributions that define HMMs and LSSMs are defined in terms of categorical and normal distributions respectively, both of which are core families for constructing conjugated harmoniums. In future work we therefore aim

to subsume inference and learning algorithms for HMMs and LSSMs within the framework of conjugated harmoniums, and explore how best to approximate nonlinear dynamics within this framework (Sokoloski, 2017).

Acknowledgements

This paper has been shaped by countless conversations I have had over the years with colleagues and friends. Nevertheless, I would like to thank Philipp Berens in particular for additional conversations and support that allowed me to finalize this manuscript.

This research was supported by the Hertie Foundation (Gemeinnützige Hertie-Stiftung); the Deutsche Forschungsgesellschaft (DFG) Sonderforschungsbereich (SFB) 1233, "Robust Vision: Inference Principles and Neural Mechanisms", Teilprojekt (TP) 13, project number: 276693517; and the DFG Cluster of Excellence "Machine Learning — New Perspectives for Science", EXC 2064.

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

S.-i. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191. American Mathematical Soc., 2007.

B. C. Arnold and S. J. Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

B. C. Arnold, E. Castillo, and J. M. Sarabia. Conjugate Exponential Family Priors For Exponential Family Likelihoods. *Statistics*, 25(1):71–77, Jan. 1993. ISSN 0233-1888. doi: 10.1080/02331889308802432.

B. C. Arnold, E. Castillo, and J. M. Sarabia. Conditionally Specified Distributions: An Introduction (with comments and a rejoinder by the authors). *Statistical Science*, 16(3): 249–274, Aug. 2001. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1009213728.

I. Assent. Clustering high dimensional data. *WIREs Data Mining and Knowledge Discovery*, 2(4):340–350, 2012. ISSN 1942-4795. doi: 10.1002/widm.1062.

T. Baden, P. Berens, K. Franke, M. Román Rosón, M. Bethge, and T. Euler. The functional diversity of retinal ganglion cells in the mouse. *Nature*, 529(7586):345–350, Jan. 2016. ISSN 1476-4687. doi: 10.1038/nature16468.

J. Baek, G. J. McLachlan, and L. K. Flack. Mixtures of Factor Analyzers with Common Factor Loadings: Applications to the Clustering and Visualization of High-Dimensional Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1298–1309, July 2010. ISSN 1939-3539. doi: 10.1109/TPAMI.2009.149.

J. Beck, P. Latham, and A. Pouget. Marginalization in Neural Circuits with Divisive Normalization. *The Journal of Neuroscience*, 31(43):15310–15319, 2011.

J. Beck, A. Pouget, and K. A. Heller. Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 3059–3067. Curran Associates, Inc., 2012.

Y. Bengio and O. Delalleau. Justifying and generalizing contrastive divergence. *Neural Computation*, 21(6):1601–1621, 2009.

J. Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974. ISSN 0035-9246.

K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When Is "Nearest Neighbor" Meaningful? In C. Beeri and P. Buneman, editors, *Database Theory — ICDT'99*, Lecture Notes in Computer Science, pages 217–235, Berlin, Heidelberg, 1999. Springer. ISBN 978-3-540-49257-3. doi: 10.1007/3-540-49257-7_15.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.

N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. Modeling Temporal Dependencies in High-dimensional Sequences: Application to Polyphonic Music Generation and Transcription. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1881–1888, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1.

W.-C. Chang. On Using Principal Components before Separating a Mixture of Two Multivariate Normal Distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 32(3):267–275, 1983. ISSN 1467-9876. doi: 10.2307/2347949.

M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405): 51–56, July 2012. ISSN 1476-4687. doi: 10.1038/nature11129.

J. P. Cunningham and B. M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, Nov. 2014. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.3776.

P. Diaconis and D. Ylvisaker. Conjugate Priors for Exponential Families. *The Annals of Statistics*, 7(2):269–281, 1979. ISSN 0090-5364.

A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, Nov. 2020. ISSN 2046-1402. doi: 10.12688/f1000research.15666.3.

D. Durstewitz. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. *PLOS Computational Biology*, 13(6):e1005542, June 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005542.

S. Gerwinn, P. Berens, and M. Bethge. A joint maximum-entropy model for binary neural population patterns and continuous signals. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

J. Hertrich, D.-P.-L. Nguyen, J.-F. Aujol, D. Bernard, Y. Berthoumieu, A. Saadaldin, and G. Steidl. PCA reduced Gaussian mixture models with applications in superresolution. *Inverse Problems and Imaging*, 16(2):341, 2022. doi: 10.3934/ipi.2021053.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

A. Houdard, C. Bouveyron, and J. Delon. High-Dimensional Mixture Models for Unsupervised Image Denoising (HDMI). *SIAM Journal on Imaging Sciences*, 11(4):2815–2846, Jan. 2018. doi: 10.1137/17M1135694.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge university press, 2003.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

A. Kutschireiter, M. A. Basnak, R. I. Wilson, and J. Drugowitsch. Bayesian inference in ring attractor networks. *Proceedings of the National Academy of Sciences*, 120(9):e2210622120, Feb. 2023. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2210622120.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, Dec. 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aab3050.

C. Langdon, M. Genkin, and T. A. Engel. A unifying perspective on neural manifolds and circuits for cognition. *Nature Reviews Neuroscience*, 24(6):363–377, June 2023. ISSN 1471-0048. doi: 10.1038/s41583-023-00693-x.

R. D. Lange, S. Shivkumar, A. Chattoraj, and R. M. Haefner. Bayesian encoding and decoding as distinct perspectives on neural coding. *Nature Neuroscience*, 26(12):2063–2072, Dec. 2023. ISSN 1546-1726. doi: 10.1038/s41593-023-01458-6.

M. S. Lewicki. A review of methods for spike sorting: The detection and classification of neural action potentials. *Network: Computation in Neural Systems*, 9(4):R53–R78, Jan. 1998. ISSN 0954-898X. doi: 10.1088/0954-898X/9/4/001.

W. J. Ma, J. Beck, P. Latham, and A. Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438, Oct. 2006. ISSN 1097-6256. doi: 10.1038/nn1790.

G. J. McLachlan, S. X. Lee, and S. I. Rathnayake. Finite Mixture Models. *Annual Review of Statistics and Its Application*, 6(1):355–378, 2019. doi: 10.1146/annurev-statistics-031017-100325.

P. D. McNicholas and T. B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, Sept. 2008. ISSN 1573-1375. doi: 10.1007/s11222-008-9056-0.

K. P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT press, 2023.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Springer, 1998.

A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*, Nov. 2015.

S. Roweis and Z. Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11(2):305–345, Feb. 1999. ISSN 0899-7667. doi: 10.1162/089976699300016674.

R. Salakhutdinov and G. Hinton. An Efficient Learning Procedure for Deep Boltzmann Machines. *Neural Computation*, 24(8):1967–2006, Apr. 2012. ISSN 0899-7667. doi: 10.1162/NECO_a_00311.

R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba. Learning with Hierarchical-Deep Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1958–1971, Aug. 2013. ISSN 1939-3539. doi: 10.1109/TPAMI.2012.269.

S. Särkkä. *Bayesian Filtering and Smoothing*. Number 3. Cambridge University Press, 2013.

E. Schneidman. Towards the design principles of neural population codes. *Current Opinion in Neurobiology*, 37:133–140, Apr. 2016. ISSN 0959-4388. doi: 10.1016/j.conb.2016.03.001.

A. Shekhovtsov, D. Schlesinger, and B. Flach. VAE Approximation Error: ELBO and Exponential Families. In *International Conference on Learning Representations*, Oct. 2021.

S. Shivkumar, R. Lange, A. Chattoraj, and R. Haefner. A probabilistic population code based on neural samples. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

G. Shmueli, T. P. Minka, J. B. Kadane, S. Borle, and P. Boatwright. A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):127–142, 2005. ISSN 1467-9876. doi: 10.1111/j.1467-9876.2005.00474.x.

J. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *Information Theory, IEEE Transactions on*, 26(1):26–37, 1980.

P. Smolensky. Information Processing in Dynamical Systems: Foundations of Harmony Theory. Technical report, COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, Feb. 1986.

S. Sokoloski. Implementing a Bayes Filter in a Neural Circuit: The Case of Unknown Stimulus Dynamics. *Neural Computation*, 29(9):2450–2490, June 2017. ISSN 0899-7667. doi: 10.1162/neco_a_00991.

S. Sokoloski, A. Aschner, and R. Coen-Cagli. Modelling the neural code in large populations of correlated neurons. *eLife*, 10:e64615, Oct. 2021. ISSN 2050-084X. doi: 10.7554/eLife.64615.

I. H. Stevenson. Flexible models for spike count data with both over- and under- dispersion. *Journal of Computational Neuroscience*, 41(1):29–43, Aug. 2016. ISSN 0929-5313, 1573-6873. doi: 10.1007/s10827-016-0603-y.

I. Sutskever and T. Tieleman. On the convergence properties of contrastive divergence. In *International Conference on Artificial Intelligence and Statistics*, pages 789–795, 2010.

W. Tansey, O. H. M. Padilla, A. S. Suggala, and P. Ravikumar. Vector-Space Markov Random Fields via Exponential Families. In *PMLR*, pages 684–692, June 2015.

G. W. Taylor, G. E. Hinton, and S. T. Roweis. Two distributed-state models for generating high-dimensional time series. *Journal of Machine Learning Research*, 12(Mar):1025–1068, 2011.

A. Vahdat, E. Andriyash, and W. Macready. Undirected Graphical Models as Approximate Posteriors. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9680–9689. PMLR, Nov. 2020.

E. Vértes and M. Sahani. Flexible and accurate inference and learning for deep generative models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4166–4175. Curran Associates, Inc., 2018.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

M. Welling, M. Rosen-zvi, and G. E. Hinton. Exponential Family Harmoniums with an Application to Information Retrieval. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1481–1488. MIT Press, 2005.

D. M. Witten and R. Tibshirani. A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association*, 105(490):713–726, June 2010. ISSN 0162-1459. doi: 10.1198/jasa.2010.tm09415.

E. Yang, G. Allen, Z. Liu, and P. Ravikumar. Graphical Models via Generalized Linear Models. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847, 2015.