

Generic Unsupervised Optimization for a Latent Variable Model With Exponential Family Observables

Hamid Mousavi

HAMID.MOUSAVI@UOL.DE

Jakob Drefs

JAKOB.DREFS@UOL.DE

Florian Hirschberger

FLORIAN.HIRSCHBERGER@UOL.DE

Jörg Lücke

JOERG.LUECKE@UOL.DE

*Machine Learning Lab
University of Oldenburg
26129 Oldenburg, Germany*

Editor: Aapo Hyvärinen

Abstract

Latent variable models (LVMs) represent observed variables by parameterized functions of latent variables. Prominent examples of LVMs for unsupervised learning are probabilistic PCA or probabilistic sparse coding which both assume a weighted linear summation of the latents to determine the mean of a Gaussian distribution for the observables. In many cases, however, observables do not follow a Gaussian distribution. For unsupervised learning, LVMs which assume specific non-Gaussian observables (e.g., Bernoulli or Poisson) have therefore been considered. Already for specific choices of distributions, parameter optimization is challenging and only a few previous contributions considered LVMs with more generally defined observable distributions. In this contribution, we do consider LVMs that are defined for a range of different distributions, i.e., observables can follow any (regular) distribution of the exponential family. Furthermore, the novel class of LVMs presented here is defined for binary latents, and it uses maximization in place of summation to link the latents to observables. In order to derive an optimization procedure, we follow an expectation maximization approach for maximum likelihood parameter estimation. We then show, as our main result, that a set of very concise parameter update equations can be derived which feature the same functional form for all exponential family distributions. The derived generic optimization can consequently be applied (without further derivations) to different types of metric data (Gaussian and non-Gaussian) as well as to different types of discrete data. Moreover, the derived optimization equations can be combined with a recently suggested variational acceleration which is likewise generically applicable to the LVMs considered here. Thus, the combination maintains generic and direct applicability of the derived optimization procedure, but, crucially, enables efficient scalability. We numerically verify our analytical results using different observable distributions, and, furthermore, discuss some potential applications such as learning of variance structure, noise type estimation and denoising.

Keywords: Latent variable models, unsupervised learning, exponential family distributions, expectation maximization, variational optimization.

1. Introduction

Latent Variable Models (LVMs) represent a probabilistic approach in which observed variables are assumed to depend on latent variables through specific parameterized functions. A large variety of unsupervised Machine Learning approaches is based on LVMs, and prominent examples are proba-

bilistic Principal Component Analysis (p-PCA; Tipping and Bishop, 1999; Roweis, 1998), Factor Analysis (FA; Everitt, 1984; Bartholomew, 1987) or probabilistic Sparse Coding (SC; Olshausen and Field, 1996). The data models of these three and many of the related approaches consist of latent variables whose weighted linear sum determines the mean of a Gaussian distribution. In probabilistic SC, for instance, the set of D observed variables y_d is modelled to depend on a set of H latent variables s_h (we assume $\vec{y} = (y_1, \dots, y_D)^T$ and $\vec{s} = (s_1, \dots, s_H)^T$, where superscript T denotes the transpose operator), according to the following generative model:

$$p(\vec{s} | \Theta) = \prod_{h=1}^H p_{\text{sparse}}(s_h; \Lambda), \quad (1)$$

$$p(\vec{y} | \vec{s}, \Theta) = \prod_{d=1}^D \mathcal{N}(y_d; \sum_{h=1}^H W_{dh} s_h, \sigma^2), \quad (2)$$

where $\vec{W}_h = (W_{1h}, \dots, W_{Dh})^T$ is commonly referred to as the *generative field* of unit h , and the matrix containing all generative fields, $W = (\vec{W}_1, \dots, \vec{W}_H)$, is often known as the model's *dictionary*. Here, $\Theta = (\Lambda, W, \sigma^2)$ denotes all parameters of the model (including prior parameter Λ , dictionary W and a variance σ^2), and the term ‘ p_{sparse} ’ refers to a sparse distribution that is used as prior distribution $p(\vec{s} | \Theta)$. The canonical choice for such a distribution is the Laplace distribution (see, e.g., Olshausen and Field, 1996; Tibshirani, 1996). Other choices include the Cauchy distribution (which was used by Olshausen and Field, 1996, alongside Laplace), Student-t (Berkes et al., 2009), Bernoulli (Haft et al., 2004; Henniges et al., 2010), Categorical (Exarchakis and Lücke, 2017), or spike-and-slab (Titsias and Lázaro-Gredilla, 2011; Goodfellow et al., 2012; Sheikh et al., 2014).

All LVMs just referred to (and presumably the great majority of LVMs) assume a Gaussian distribution for the observables. Most LVMs for Gaussian data then linearly link the set of latent variables to the set of observables using a weighted sum. But a Gaussian distribution cannot be assumed for all data. Consequently, other observable distributions have been investigated. For instance, the Poisson distribution has been of interest for count data, and likewise, the Bernoulli distribution for binary data. The aforementioned LVMs (p-PCA, SC etc.) represent well-known examples for Gaussian observables. For the case of non-Gaussian observables, examples include FA with Poisson observables (Zhou et al., 2012), FA with binary (Boolean) observables (BFA; Frolov et al., 2014) or PCA with Poisson observables (Salmon et al., 2014; Chiquet et al., 2018). Also well-known models such as noisy-OR Bayes nets (Šingliar and Hauskrecht, 2006; Jernite et al., 2013) or shallow Sigmoid Belief Networks (SBNs; Saul et al., 1996; Gan et al., 2015) can be considered as LVMs with Bernoulli distributed observables (also compare van der Linden and Hambleton, 2013). Moreover, types of non-negative matrix factorization (where non-Gaussian distributions are also used) can be regarded as a form of LVM (see, e.g., Hoffman, 2012; Basbug and Engelhardt, 2016, and references therein).

Unlike such models with specific observable distributions, some previous contributions have also considered more generally defined LVMs for unsupervised learning. For instance, Collins et al. (2002), Mohamed et al. (2009) or Li and Tao (2010) considered a PCA generalization to include observables of the exponential family (referred to as EF-PCA), and Lee et al. (2009) considered SC with exponential family observables (EF-SC). Parameter optimization for specific non-Gaussian distributions is already considered more challenging than for Gaussian distribution. If LVMs are defined more generally, optimization becomes still more challenging as any method has to be defined without exploiting properties that only apply for a specific distribution.

Besides the approaches mentioned above, another class of models which is very general by considering exponential family distributions for observables is represented by Generalized Linear Models (GLMs; Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). GLMs can be related to the aforementioned approaches (and also to our work here) since they likewise exploit other noise distributions besides the Gaussian and do provide generic parameter update equations also for the case of non-Gaussian distributions (e.g. McCullagh and Nelder, 1989). Nonetheless, GLMs essentially rely on supervised learning, i.e. regression, which is not the subject of this study (see Collins et al., 2002, for a discussion).

In this study, we consider a class of LVMs for unsupervised learning which is like EF-PCA or EF-SC more generally defined to include observable distributions of the exponential family. Unlike EF-PCA or EF-SC, however, we will seek an optimization procedure that is generally applicable to any regular exponential family distribution. In order to achieve our goal of a generic parameter optimization for any observable distribution of the exponential family, we replace the common linear summation of latents by a maximization. More precisely, for the LVMs considered here, the latents determine the *mean* of the observable distribution through maximization in contrast to the conventional Gaussian-based LVMs such as p-PCA or SC (and also GLM-related approaches) that determine the mean using summation. In addition, our approach maintains the same influence of latents on the observable mean (using a maximum non-linearity) independently of the specific choice for the observable distribution. For approaches such as EF-SC, a weighted sum of latents is used to set the natural parameter(s) of the (usually single parameter) exponential family distribution. Doing so implies a general non-linear function for the distribution mean, and this function changes depending on the used exponential family distribution. Also in the context of regression, GLMs similarly use a weighted summation followed by a non-linear function to set the distribution mean.

Models with a non-linear superposition of the latents have been of interest for a variety of tasks, and were actively researched in previous works. For instance, the combination of noisy-OR Bayes nets can be considered as such a non-linear link as well as point-wise sigmoidal functions after summation as used in SBNs (Saul et al., 1996; Gan et al., 2015). Still further forms of models with a specific latent variable structure have been investigated in the context of non-linear Independent Component Analysis (ICA; Hyvärinen and Pajunen, 1999) and deep generative models such as Variational Autoencoders (VAEs; Kingma and Welling, 2013), generative adversarial nets, flow nets etc. (e.g. Bond-Taylor et al., 2022). Standard (linear) ICA approaches (Hyvärinen and Oja, 2000; Comon, 1994) investigate continuous non-noisy observables while (similar to standard SC) non-Gaussian latents are considered. Later developments have significantly generalized earlier restrictions with recent work allowing, e.g., for binary observations (Hyttinen et al., 2022) and non-linear influences of the latents on the observables (Hyvärinen and Pajunen, 1999). For non-linear ICA approaches, non-linearities can take either the form of a post-linear non-linearity (i.e., a linear superposition of generative fields followed by an invertible non-linear function; Taleb and Jutten (1999)), or, more generally, the form of a general smooth and invertible function that can be used in place of the sum (Sprekeler et al., 2014; Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019). Non-linear ICA and variational autoencoders as presumably most relevant types of deep models will be discussed further in Sec. 4, i.e. after the definition of the proposed LVMs.

Like the above types of non-linearities, the here considered fixed non-linearity in the form of a (weighted) maximization, has also been investigated previously. For instance, Roweis (2003), Shelton et al. (2012), Shelton et al. (2017) or Sheikh et al. (2019) used maximization for LVMs with observables distributed according to standard Gaussian distributions, and Lücke and Sahani

(2008) used the maximum for an LVM with Poisson observables. Moreover, similar fixed non-linear superpositions have been used to model mutual occlusions of objects in images (Lücke et al., 2009; Henniges et al., 2014). The approach named Occlusive Component Analysis (OCA) likewise uses Gaussian observables (also see Dai and Lücke (2014) for another relevant study in this direction). Like the work by Lücke and Sahani (2008) or Sheikh et al. (2019), we will use the fixed maximum non-linearity in this study. But unlike such previous works, we will not use the non-linearity together with only one specific observable distribution. Instead, we will consider LVMs which encompass any (regular) member of the exponential family as observable distribution.

After defining the general class of LVMs, we will address the challenge of parameter optimization by following a maximum likelihood approach using Expectation Maximization (EM; Dempster et al., 1977). Our central analytical result will then be a set of concise parameter update equations (M-steps) that are generally and directly applicable to any regular member of the exponential family. Furthermore, the here derived M-steps can be combined with likewise generally applicable and efficient variational E-steps using a recently suggested variational approach (Drefs et al., 2022a). Neither the generic M-steps derived in this contribution nor the variational E-steps of Drefs et al. (2022a) require any analytical derivations before they can be applied. If generic M-steps and variational E-steps are combined, we then obtain an efficient EM optimization algorithm that is, without any derivations, applicable to any observables that follow a regular exponential family distribution¹. Finally, we provide numerical evaluations for one- as well as two-parameter distributions of the exponential family and point to some potential applications.

2. Preliminary Definitions

Before defining the family of LVM models, we provide details on exponential family distributions which will be used throughout this paper. We use observable distributions $p(y; \vec{\eta})$ that take the form of an exponential family distribution given by:

$$p(y; \vec{\eta}) = h(y) \exp(\vec{\eta}^T \vec{T}(y) - A(\vec{\eta})), \quad y \in \mathcal{Y}, \quad (3)$$

where \mathcal{Y} is the domain of a scalar observable y , $h(y)$ is the base measure, $\vec{T}(y)$ represents the *sufficient statistics* of the data, and $\vec{\eta}$ and $A(\vec{\eta})$ are the *natural parameters* and *log-partition* function, respectively. The vectors $\vec{T}(y) = (T_1(y), \dots, T_L(y))^T$ and $\vec{\eta} = (\eta_1, \dots, \eta_L)^T$ have L elements each when the distribution $p(y; \vec{\eta})$ is an L -parameter distribution. We remark that a notational alternative is provided by using $p(y | \vec{\eta})$ in (3) instead of $p(y; \vec{\eta})$. However, we here use the latter notation (which describes $p(\cdot)$ as a function of y with given parameters $\vec{\eta}$) for the sake of a better readability. We follow the same argumentation for the other probability density functions used in the paper (including those of Eqns. (1) and (2)).

1. Generic applicability does by itself not mean that the EM algorithm derived here will necessarily be based entirely on closed-form equations. Even estimating the parameters of many exponential family distributions is itself often not possible in closed-form (e.g. for Gamma or Beta distributions). But if the parameters of the exponential family distribution can be estimated in closed-form, then EM for the corresponding LVM will be closed-form. If the parameters of the exponential family distribution can *not* be estimated in closed-form, then this will turn out to be the only part requiring approximations (which are usually readily available). See end of Sec. 6 for details.

The log-partition $A(\vec{\eta})$ plays the role of a normalizer. Following from (3), it is for a given $\vec{\eta}$ defined as:

$$A(\vec{\eta}) = \log \left(\int h(y) \exp(\vec{\eta}^T \vec{T}(y)) dy \right). \quad (4)$$

Formally, the domain of A is usually defined to be the set of all natural parameters $\vec{\eta}$ for which $A(\vec{\eta}) < \infty$. We will only consider exponential families that are *regular*, i.e., families for which the set of all natural parameters $\vec{\eta}$ is a non-empty and open set (e.g. Bickel and Doksum, 2015; Efron, 2023). Furthermore, all integrals over the domain \mathcal{Y} can be general Lebesgue integrals but we will not use Lebesgue measures explicitly. For the purposes of this paper, the integrals can be considered as using the ordinary Lebesgue measure on \mathbb{R} in the continuous case, while in the discrete case integrals become a summation over finite probabilities (compare, e.g., Wainwright and Jordan, 2008). For any regular exponential family, and for any $l, k \in \{1, \dots, L\}$, we have (e.g. Bickel and Doksum, 2015; Efron, 2023):

$$\frac{\partial A(\vec{\eta})}{\partial \eta_l} = \mathbb{E}_{p(y; \vec{\eta})}[T_l(y)], \quad \frac{\partial^2 A(\vec{\eta})}{\partial \eta_l \partial \eta_k} = \text{cov}(T_l(y), T_k(y)), \quad (5)$$

where $\mathbb{E}_{p(\cdot)}[f(y)]$ denotes the expectation value of $f(y)$ w.r.t. the distribution $p(\cdot)$ (evaluated per entry if f is a vector) and $\text{cov}(\cdot, \cdot)$ represents the covariance. Also note that an alternative notation for the expectation value is $\langle f(y) \rangle_{p(\cdot)}$ which has been used in previous works (e.g., Lücke and Sahani, 2008; Sheikh et al., 2014).

As an important tool to later derive parameter update equations, we will use the *mean value parametrization* (a.k.a. the *mean parametrization*) of the exponential family (e.g. Wainwright and Jordan, 2008; Cox, 2006; Efron, 2023). That is, we consider parameters \vec{w} (as the model's parameters) defined by:

$$\vec{w} := \mathbb{E}_{p(y; \vec{\eta})}[\vec{T}(y)], \quad \vec{w} = (w_1, \dots, w_L)^T. \quad (6)$$

The mapping from $\vec{\eta}$ to \vec{w} is bijective, i.e. it is invertible, if the distribution has a *minimal representation*² property (compare Wainwright and Jordan, 2008; Efron, 2023, and many others). Assuming minimality from now on, there exists a function $\vec{\Phi}$ which maps the mean value parameters to the natural parameters as follows:

$$\vec{\eta} = \vec{\Phi}(\vec{w}) \quad \text{such that} \quad \vec{w} = \mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[\vec{T}(y)]. \quad (7)$$

For some distributions of the exponential family the function $\vec{\Phi}$ can be obtained in closed-form (e.g., for Poisson, Bernoulli and Gaussian). For others, however, a closed-form expression may in general not be available. Still, the existence of the inverse function $\vec{\Phi}$ is guaranteed as long as the distribution has a minimal representation property. In the next section, we will show how the function $\vec{\Phi}$ along with other properties presented above will be exploited to define a specific family of generative models for exponential family observables.

2. A distribution of the exponential family is said to be minimal if there are no linear dependencies among components of the sufficient statistics, i.e., there is no coefficient $\vec{a} \in \mathbb{R}^L$ with $\vec{a} \neq \vec{0}$ such that

$$\sum_{l=1}^L a_l T_l(y) = \text{constant} \quad \text{for all } y \in \mathcal{Y}.$$

3. Proposed Family of Latent Variable Models

We hereafter consider LVMs with the same notation as in Eqns. (1)-(2). However, we assume the latents to be binary (i.e. each element $s_h \in \{0, 1\}$), and for the sake of consistency with previous work, we assume each s_h to follow a Bernoulli distribution. Using small values of Bernoulli parameters, the model can thus realize a sparse encoding of the latents. We remark, nonetheless, that the analytical results derived in the following only require binary latents such that the results will also apply for non-sparse encodings and priors more complex than H independent Bernoulli distributions. As a further assumption we assume, similar to the standard SC, that all observables y_d are conditionally independent given the latents. Unlike standard SC, however, the conditional distribution for the observables can follow any regular distribution of the exponential family (3) with natural parameters depending on \vec{s} . Finally, and importantly, the observables in our model class depend on the latents through maximization.

3.1 Model Definition

We investigate LVMs according to the following generative model:

$$p(\vec{s} | \Theta) = \prod_{h=1}^H (\pi_h^{s_h} (1 - \pi_h)^{1-s_h}), \quad (8)$$

$$p(\vec{y} | \vec{s}, \Theta) = \prod_{d=1}^D p(y_d; \vec{\eta}_d(\vec{s}, \Theta)), \text{ where} \quad (9)$$

$$p(y; \vec{\eta}) = h(y) \exp(\vec{\eta}^T \vec{T}(y) - A(\vec{\eta})), \quad (10)$$

and where $\pi_h \in (0, 1)$ parameterizes the prior distribution on s_h ; we let $\vec{\pi} = (\pi_1, \dots, \pi_H)^T$. For model (8)-(10), the latent variables \vec{s} couple to the observed variable y_d through the function $\vec{\eta}_d(\vec{s}, \Theta)$ which is defined further below (the function can be considered as a link function in a broader sense). The observable distribution (10) will in the following also be referred to as *noise distribution*.

Note that $\vec{\eta}_d(\vec{s}, \Theta)$ is a vector with L entries for each d . For one-parameter distributions of the exponential family (Poisson, Exponential, Pareto etc.) $L = 1$, for two-parameter distributions (Gamma, Beta etc.) $L = 2$, and so forth (e.g., Categorical and Multinomial distributions can be mentioned for the case of arbitrary L). For standard SC, a Gaussian with known variance is commonly considered (just the mean is parameterized). Standard SC then sets the *mean* of the observables using a matrix $W \in \mathbb{R}^{D \times H}$, see Eqns. (1) and (2). Here, we seek to couple latents and observables in an analog but more generally applicable way.

In order to facilitate our notation, let us from now on assume two-parameter distributions ($L = 2$) as it applies, e.g., for Gamma, Beta and Gaussian distributions. Arbitrary L will be treated later on. For the generative model presented above, we will for each observable y_d parameterize the distribution (9) in terms of the mean value parameters $\vec{w} = (w_1, w_2)^T$ (Eqn. 6). For this purpose and for the $L = 2$ case, let $w_1 = \bar{W}_d$ and $w_2 = \bar{V}_d$ for each d , where \bar{W}_d and \bar{V}_d are functions depending on the latents \vec{s} and *two* matrices W and V with $D \times H$ entries. Linear and non-linear combination of latents to determine the observable distributions will be possible through different choices for the functions \bar{W}_d and \bar{V}_d (see below). The notation using $L = 2$ serves for gaining some intuition as (e.g., for Gaussian or Gamma) \bar{W}_d may denote the distribution mean of observable y_d and \bar{V}_d the distribution variance (or the second moment of the distribution). Figure 1 illustrates the structure of

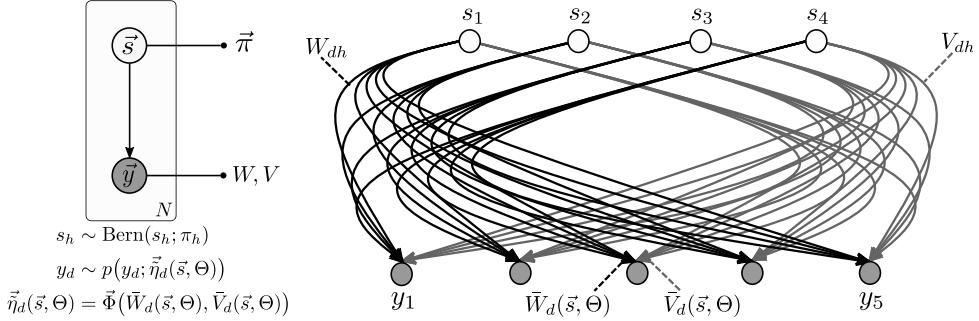


Figure 1: Graphical representation of the proposed generative model with (as an example) $H = 4$ latent variables and $D = 5$ observed variables. Given a binary vector \vec{s} , the value y_d of an observable follows an exponential family distribution $p(y_d; \vec{\eta}_d(\vec{s}, \Theta))$ with natural parameters determined by the functions $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ for the distribution's mean value parameters. The functions depend on the latents and on the matrices $W \in \mathbb{R}^{D \times H}$ and $V \in \mathbb{R}^{D \times H}$ visualized at the right-hand-side. Here, the set of all model parameters is $\Theta = (\vec{\pi}, W, V)$ and N denotes the number of samples.

the proposed model where, in the case of $L = 2$, two matrices connect the latents to the observables. Now, using the function $\vec{\Phi}(\vec{w})$ given by (7), we define the link from latents to observables as follows:

$$\vec{\eta}_d(\vec{s}, \Theta) := \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)), \quad d = 1, \dots, D. \quad (11)$$

The mapping $\vec{\Phi}$, as mentioned before, is specific to the choice of observable distribution (10) which is a regular exponential family distribution. Although a closed-form formulation of the function is only available for some instances of the family, the function is well-defined as long as the noise distribution (10) has a minimal representation property (the assumption that we consider here). Thus, such a general link function is well-defined for the proposed data models using $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ (also compare Fig. 1). For the case of arbitrary L , we will define this link using L parameters and subsequently L matrices (see Sec. 6 further below). To complete the definition of our generative models, it now remains to define the functions $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ which determines the influence of latents on the observable distributions. Before that, let us elaborate the notation above using a preeminent example of the exponential family distributions.

Example 1 *To provide an intuition for the link defined by (11), consider Gaussian observation noise. In that case, sufficient statistics and natural parameters are given by $\vec{T}(y) = (y, y^2)^T$ and $\vec{\eta} = (\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2})^T$ (where μ and σ^2 are the mean and variance of the Gaussian distribution, respectively). By expressing the mean value parameters \vec{w} in terms of the natural parameters $\vec{\eta}$, we get:*

$$\vec{w} = \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix} = \frac{1}{4\eta_2^2} \begin{pmatrix} -2\eta_1\eta_2 \\ \eta_1^2 - 2\eta_2 \end{pmatrix}. \quad (12)$$

Given the equation above, the inverse mapping $\vec{\Phi}$ can be computed in closed-form (in this case) and is given by:

$$\vec{\Phi}(\vec{w}) = \frac{1}{2(w_2 - w_1^2)} \begin{pmatrix} 2w_1 \\ -1 \end{pmatrix}. \quad (13)$$

Using the definition of the link function in (11), the coupling of latents to observables is consequently given by:

$$\begin{aligned}\vec{\eta}_d(\vec{s}, \Theta) &= \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)) \\ &= \frac{1}{2(\bar{V}_d(\vec{s}, \Theta) - \bar{W}_d^2(\vec{s}, \Theta))} \begin{pmatrix} 2\bar{W}_d(\vec{s}, \Theta) \\ -1 \end{pmatrix}.\end{aligned}\quad (14)$$

In order to recover standard linear SC, we can complete the definition of $\vec{\eta}_d(\vec{s}, \Theta)$ by using parameters (σ^2, W) with a global, scalar $\sigma^2 \in \mathbb{R}^+$ and $W \in \mathbb{R}^{D \times H}$ and further setting:

$$\bar{W}_d(\vec{s}, \Theta) = \sum_h W_{dh} s_h \quad \text{and} \quad \bar{V}_d(\vec{s}, \Theta) = \sigma^2 + \bar{W}_d^2(\vec{s}, \Theta), \quad (15)$$

which then results in:

$$\vec{\eta}_d(\vec{s}, \Theta) = \begin{pmatrix} (W\vec{s})_d / \sigma^2 \\ -1/2\sigma^2 \end{pmatrix}. \quad (16)$$

That is, we recover the standard SC parametrization with $\mu_d = (W\vec{s})_d$ as the mean of observable d and σ^2 as its variance (with σ^2 being the same for all d).

3.2 Definition of a Maximum Superposition Model

In Example 1, given a matrix $W \in \mathbb{R}^{D \times H}$, the mean μ_d of observable d was given by $(W\vec{s})_d$. Our definition of the link (11) from latents to observables is sufficiently flexible to also allow for types of latent combinations other than the weighted linear sum that is very commonly used (e.g., for p-PCA, sparse coding etc). In this respect, let us first ask the question what aspects of a given observable distribution the two functions \bar{W}_d and \bar{V}_d shall determine. Most commonly, the latents determine the *mean* of observable d using summation (again as, e.g., for p-PCA, sparse coding etc). For many members of the exponential family, the first mean value parameter coincides with the mean of the corresponding exponential family distribution (that is we have $T_1(y) = y$). Note that if there exists an l with $T_l(y) = y$ (which is not always the case), we will assume this sufficient statistics' component to be listed at the first position ($l = 1$). If $T_1(y) = y$, then (following from our definition) the function $\bar{W}_d(\vec{s}, \Theta)$ alone directly determines the observable mean:

$$\begin{aligned}\mu_d &= \mathbb{E}_{p(y; \vec{\eta}_d(\vec{s}, \Theta))}[y] = \mathbb{E}_{p(y; \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)))}[y] \\ &= \mathbb{E}_{p(y; \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)))}[T_1(y)] = \bar{W}_d(\vec{s}, \Theta),\end{aligned}\quad (17)$$

where the last step follows from (7) and the fact that $w_1 = \bar{W}_d(\vec{s}, \Theta)$. The function $\bar{W}_d(\vec{s}, \Theta)$ can then be either a weighted summation $\bar{W}_d(\vec{s}, \Theta) = (W\vec{s})_d$ as in Example 1 or a non-linear function of the latents. For our purposes, we will follow previous work (Lücke and Sahani, 2008; Bornschein et al., 2013; Sheikh et al., 2019) and demand that the mean of an observable distribution be given by a weighted maximum instead of a weighted sum. More concretely, we will replace (e.g., for the Gaussian or Gamma distribution):

$$\bar{W}_d(\vec{s}, \Theta) = (W\vec{s})_d = \sum_h W_{dh} s_h \quad \text{by} \quad \bar{W}_d(\vec{s}, \Theta) = \max_h \{W_{dh} s_h\}. \quad (18)$$

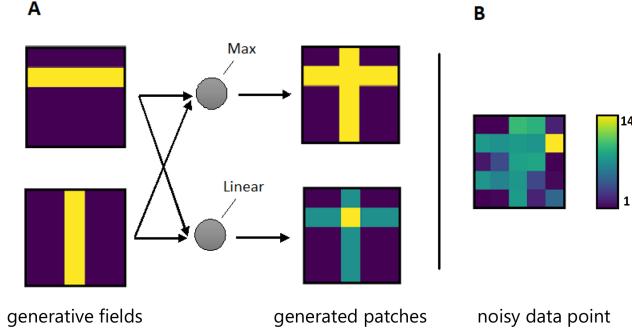


Figure 2: Illustration of non-linear vs. linear superposition of generative fields. **A** Two generative fields in the form of bars are combined through maximization and summation to generate data. **B** Example of a noisy data point which is generated using the Exponential-MCA model (i.e., we used the patches generated by the maximum superposition in **A** together with Exponential noise).

Figure 2 illustrates the differences between weighted summation and weighted maximization of the latents using artificial generative fields in the form of bars. Maximization instead of summation has been suggested before for acoustic data (Roweis, 2003; Lücke and Sahani, 2008; Sheikh et al., 2019) and for visual data (Bornstein et al., 2013) where it aligns more closely with the actual data generating process. As summation of generative fields is also referred to as a *linear superposition* of generative fields, maximization can be considered as a *non-linear superposition*. In the following, summation and maximization are, respectively, referred to linear or maximum *superposition models*. Importantly for the purposes of this study, the specific properties of maximization will enable us to derive a generic parameter optimization procedure for the family of data models (8)-(10).

Equation (18) would cover members of the exponential family such as Gaussian and Gamma (as $T_1(y) = y$ holds for these distributions). However, to be applicable to the whole exponential family, the definition of the maximum superposition has to be further generalized. The challenge is that, in general, the sufficient statistics $\vec{T}(y)$ may not have an element that is proportional to y , and also a definition for the function $\bar{V}_d(\vec{s}, \Theta)$ is required. As an example which makes a generalization necessary consider the Beta distribution where $T_1(y) = \log(y)$ and $T_2(y) = \log(1 - y)$. As a consequence, the definition of the maximum superposition has to involve both functions $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$. Nevertheless, it is possible to define $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ such that the mean of the observables is determined by a maximum superposition. For this purpose, consider a weight matrix $M(\Theta)$ with $D \times H$ entries which can potentially be a non-trivial function of the parameters Θ . Given a latent vector \vec{s} , we now require the mean μ_d of an observable d to be given by:

$$\mu_d = \mathbb{E}_{p(y; \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)))}[y] \stackrel{!}{=} \max_h \{M_{dh}(\Theta) s_h\}. \quad (19)$$

Clearly, if $T_1(y) = y$, we can satisfy the demand by choosing $M_{dh}(\Theta) = W_{dh}$ and further setting $\bar{W}_d(\vec{s}, \Theta) = \max_h \{W_{dh} s_h\}$. From (17), we can consequently obtain:

$$\mu_d = \mathbb{E}_{p(y; \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)))}[y] = \bar{W}_d(\vec{s}, \Theta) = \max_h \{W_{dh} s_h\} = \max_h \{M_{dh}(\Theta) s_h\}, \quad (20)$$

which fulfills (19). For the general case, nonetheless, the definition of $M(\Theta)$ has to be more elaborate in order to apply for any exponential family distribution. As mentioned before, we consider two

$D \times H$ matrices W and V as part of the model parameters Θ , i.e. $\Theta = (\vec{\pi}, W, V)$. We can then define $M(\Theta)$ as follows:

$$\forall d, h : M_{dh}(\Theta) := F(W_{dh}, V_{dh}) \text{ where } F(w, v) = \mathbb{E}_{p(y; \vec{\Phi}(w, v))}[y]. \quad (21)$$

Using matrix $M(\Theta)$, we now define our mappings $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ as follows:

$$\bar{W}_d(\vec{s}, \Theta) := W_{dh(d, \vec{s}, \Theta)}, \quad \bar{V}_d(\vec{s}, \Theta) := V_{dh(d, \vec{s}, \Theta)}, \quad (22)$$

$$\text{where } h(d, \vec{s}, \Theta) := \operatorname{argmax}_h \{M_{dh}(\Theta) s_h\}.$$

That is, we define the functions $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ via an index $h(d, \vec{s}, \Theta)$ which selects, for a given observable d , the latent with the maximal value of $M_{dh}(\Theta)$. Using the relations in (21) and (22) for the link (11) completes the definition of the non-linear superposition model. The model represents a generalization of the superposition model (20) suitable for the whole exponential family. To see this, we derive:

$$\begin{aligned} \mu_d &= \mathbb{E}_{p(y; \vec{\eta}_d(\vec{s}, \Theta))}[y] = \mathbb{E}_{p(y; \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)))}[y] \\ &= F(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)) = F(W_{dh(d, \vec{s}, \Theta)}, V_{dh(d, \vec{s}, \Theta)}) \\ &= M_{dh(d, \vec{s}, \Theta)}(\Theta) = \max_h \{M_{dh}(\Theta) s_h\}, \end{aligned} \quad (23)$$

which again fulfils (19) but this time for the general case. In virtue of (21) and (22), we do *not* have to require that there exists a sufficient statistic $T_1(y) = y$, i.e., no further restrictions to an exponential family distribution have to be demanded. If $T_1(y) = y$, then definition (21) results in $F(w, v) = w$. Consequently, we drop back to the easier special case of $M_{dh}(\Theta) = W_{dh}$.

The definition of functions $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ are relatively technical but allow for a general definition of generative models using any exponential family (with a minimal representation property). Our LVM defines a link $\vec{\eta}_d(\vec{s}, \Theta)$ from latents to observables for the proposed data models which is a consistent generalization of the maximum non-linearity (20) used, e.g., for Poisson (Lücke and Sahani, 2008) and Gaussian (Lücke and Eggert, 2010) observables. In our case, the proposed link remains consistently defined for all noise distributions of the exponential family. That is, independently of the choice of noise distribution, the link $\vec{\eta}_d(\vec{s}, \Theta)$ defined by (21) and (22) ensures that the latents change the *mean* of the observables always according to the maximum superposition model. This is in contrast to the previous LVMs for exponential family distributions such as EF-PCA (Collins et al., 2002), Bayesian EF-PCA (Mohamed et al., 2009) and EF-SC (Lee et al., 2009) where the latents set the natural parameters of the observables using a linear superposition model (we will elaborate on these studies further below).

To summarize, the definition of the proposed family of generative models is given by (8)-(10) and link (11) with (21) and (22). In analogy to previous models defined using the maximum non-linearity (Lücke and Sahani, 2008), we will refer to the family of models as *Exponential Family Maximal Causes Analysis* (EF-MCA). While the EF-MCA data models are very general, we will later see that the chosen parametrization results in generic equations for parameter updates. Before that, we will illustrate the above technical definitions using again the Gaussian distribution as an example.

Example 2 Consider the Gaussian case of Example 1. For the corresponding EF-MCA model, the matrix $M(\Theta)$ is (because of $T_1(y) = y$) given by $M_{dh}(\Theta) = W_{dh}$ and

$$\bar{W}_d(\vec{s}, \Theta) = W_{dh(d, \vec{s}, \Theta)} = \max_h \{W_{dh} s_h\} \quad (24)$$

$$\bar{V}_d(\vec{s}, \Theta) = V_{dh(d, \vec{s}, \Theta)}. \quad (25)$$

Using Eqn. (7) and the $\vec{\Phi}$ function (13), we again obtain $\vec{\eta}_d(\vec{s}, \Theta)$ similar to (14) but this time with $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ given by (24) and (25), respectively. In terms of the standard Gaussian parametrization, we would thus obtain as noise model:

$$p(y_d; \vec{\eta}_d(\vec{s}, \Theta)) = \mathcal{N}(y_d; \bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta) - \bar{W}_d^2(\vec{s}, \Theta)).$$

The matrix $V \in \Theta$ thus allows for the latents to also parameterize the variance. We will later see how this parametrization can also be changed back to the normal parametrization using a matrix $\sigma_{dh}^2 = V_{dh} - W_{dh}^2$ instead of V_{dh} (using σ_{dh}^2 is more closely aligned with the conventional parametrization). Note, however, that we here use a variance element σ_{dh}^2 for each d and h which is in contrast to a global scalar variance considered in (1)-(2). A generalization assuming a matrix (with elements σ_{dh}^2), to the best of our knowledge, has not been considered before and the data model presented in (8)-(10) together with (11), (21) and (22) is the first non-linear LVM which allows for the training of two matrices (one corresponding to the mean of observables, W , and another which parameterizes the variances, V). This feature makes the data model more flexible compared to previous approaches (e.g. Lücke and Eggert (2010); Puertas et al. (2010)) and provides more information about the given dataset. The generalized Gaussian model has previously been applied to model the statistics of natural image patches (Mousavi et al., 2020) where standard Gabor-like generative fields were obtained together with variance profiles per latent variable. The new statistical properties captured by the novel matrix V can predict richer response properties of simple cells in primary visual cortex (VI) than conventional modeling using one matrix W for the mean responses (we refer the reader to Mousavi et al. (2020) and Sec. 7.2.1 for more information). If we wanted to enforce scalar variances (as conventionally used), we could define $\bar{V}_d(\vec{s}, \Theta)$ as in (15).

4. Relation to Previous Work on Gaussian and Non-Gaussian Observables

It has been the goal of many previous approaches to derive learning algorithms for non-Gaussian data (see, e.g., Collins et al., 2002; Mohamed et al., 2009; Lee et al., 2009; Mohamed et al., 2012; Giryes and Elad, 2014; Lu et al., 2016; Zhou et al., 2012; Salmon et al., 2014; Chiquet et al., 2018; Tolooshams et al., 2020). Often the derivation of an update procedure for just a specific non-Gaussian distribution is aimed at (e.g., by assuming Poisson distributed observables for count data, Salmon et al., 2014). However, a subset of these approaches do aim at more generally deriving updates allowing for a range of different non-Gaussian observable distributions (see, e.g., Collins et al., 2002; Mohamed et al., 2009; Lee et al., 2009; Li and Tao, 2010; Chiquet et al., 2018; Tolooshams et al., 2020).

As an important reference for this work, the approach of exponential family sparse coding (EF-SC; Lee et al., 2009), for instance, defines a SC approach for noise distributions of the exponential family. The work chooses a link function that assumes the natural parameter of an exponential family distribution to be given by a linear superposition of generative fields. Essentially, if $\tilde{\eta}_d(\vec{s}, \Theta)$ is the natural parameter of a one-parameter distribution, then $\tilde{\eta}_d(\vec{s}, \Theta) = (W\vec{s})_d$ (in Lee et al., 2009, W is denoted by B). With a link defined in this way and a standard Laplacian prior, the posteriors of the model maintain a mono-modal shape such that a maximum a-posteriori (MAP) approximation can be applied (Lee et al., 2009). Nevertheless, MAP-based training of SC is usually restricted to the generative fields; like for standard SC, neither prior parameters nor parameters of the noise model are inferred. Cross-validation may be used in addition but is only feasible for very few additional parameters.

Setting the natural parameters of a given exponential family distribution using a linear superposition may (while mathematically convenient) not in general capture the true data generating process well. Which of the modeling choices is preferable (conventional linear superposition or a specific maximum non-linearity) does very much depend on the data, of course. As an example for the conventional linear modeling choice, consider Poisson noise as a one-parameter exponential family distribution. Then the natural parameter is given by $\eta = \log(\lambda)$, where λ is the mean. Using a linear superposition to set the natural parameter (as for EF-SC according to Lee et al. (2009)) would then mean $\lambda_d = \exp(\sum_h W_{dh} s_h)$ for an observable d . The same choice for the link was used for the Poisson PCA approach (Salmon et al., 2014) which focuses on Poisson noise and reported state-of-the-art denoising performance when it was first suggested (also see Giryes and Elad, 2014). The example shows that the conventional modeling choice results in a specific post-linear non-linearity to set the mean of the observable distribution. Other distributions of the exponential family will result in other types of non-linearities if the natural parameter is assumed to be set by a linear sum. Examples of data that may well be described by a non-linear generative model include, e.g., low-intensity images where image pixels may be assumed to approximately follow a Poisson distribution whose mean is determined by a non-linearity that is close to a maximum function (also compare Poisson FA; Zhou et al., 2012). Acoustic waveforms, on the other hand, suggest the use of conventional linear modeling (with an appropriate noise distribution).

Most approaches like Poisson PCA (Salmon et al., 2014; Chiquet et al., 2018) or EF-PCA (Collins et al., 2002; Sajama and Orlitsky, 2004; Li and Tao, 2010), use links from latents to observables very similar to EF-SC (Lee et al., 2009). Concretely, the method for generalizing linear regression to non-Gaussian data as introduced for the GLM (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) was used in definitions of the corresponding models for unsupervised learning. Furthermore, the parameter optimization procedures for unsupervised learning are themselves also often strongly inspired by GLM parameter optimization (Collins et al., 2002; Lee et al., 2009, and others). EF-PCA thus uses (like Poisson PCA) a deterministic optimization for parameter learning.

The original EF-PCA work only considered one-parameter distributions of the exponential family, however, with sufficient statistics proportional to y (see Collins et al., 2002, Eqn. 2); this restriction excludes the Pareto, Laplace and other one-parameter distributions. A fully Bayesian approach to exponential family PCA (Mohamed et al., 2009) also uses the GLM-inspired linear superposition to set the natural parameters but replaces MAP-based training with a Hybrid Monte Carlo approach (which also allows for learning parameters other than the weight matrix). Bayesian EF-PCA is formulated for distributions of the exponential family with one and several parameters, but the concrete algorithm uses a sufficient statistic equal to the identity function (see Mohamed et al., 2009, Eqn. 7). EF-SC (Lee et al., 2009) can be considered as more general in its initial formulation. However, the used generative fields are all defined to set one-parameter distributions (their matrix B sets the natural parameter of an exponential family distribution). Also for the Gaussian case, which is used as an introductory example, a single-parameter Gaussian (with known variance) is chosen. Closely related to such assumptions is the restriction to the natural exponential family as explicitly discussed by some approaches (Chiquet et al., 2018; Tolooshams et al., 2020).

Notably, none of the above mentioned previous approaches, i.e. Poisson PCA (Salmon et al., 2014), Poisson FA (Zhou et al., 2012), EF-PCA (Collins et al., 2002; Sajama and Orlitsky, 2004; Chiquet et al., 2018), Bayesian EF-PCA (Mohamed et al., 2009) and EF-SC (Lee et al., 2009), used any other than one-parameter distributions of the exponential family in their contributions as a concrete example. Moreover, and to the best knowledge of the authors, the numerical experiments of

these contributions (including EF-SC) actually only used the Bernoulli, Exponential or the Poisson distribution (usually two of the three). Although, e.g. EF-SC, may in principle be able to treat other distributions of the exponential family, the meaning of setting the natural parameters of a distribution with more than one parameter using a linear superposition of the latents is arguably unclear. For instance, even for the Gaussian (as a standard two-parameter distribution), the second moment would be determined by two linear sums. Such a case (and more intricate cases resulting from other two-parameter distributions) have not been discussed or treated by EF-SC nor by EF-PCA or similar approaches. In addition, note that approaches such as EF-SC are restricted beyond being focused on one-parameter distributions. As a deterministic MAP-based learning is used, only weight parameters are updated (and neither additional noise nor prior parameters are learned).

In contrast to such previous approaches for exponential family observables, the here studied EF-MCA models define the links from latents to the observables using a maximum superposition model. Moreover, the link function always sets the mean of the observable distribution using the auxiliary weight matrix $M(\Theta)$ which is defined by the two dictionaries W and V (if the distribution has two parameters). This novel definition enables us to also treat more intricate cases such as Gamma and Beta distributions (and, indeed, any two parameter distributions of the exponential family). We will later show how our definitions can be generalized to exponential family distributions with an arbitrary number of parameters L , and how a generic parameter optimization can be derived using the EM algorithm. The derived algorithm will consequently apply for the entire set of (regular) exponential family distributions. We note, however, that the proposed EF-MCA models exploit binary latents (here in the form of Bernoulli distributed priors), which may be restrictive in some settings. Also, the proposed link function defined by (11) with (21) and (22) imposes a specific non-linearity in the form of a (weighted) maximum function. While these assumptions will be advantageous for our focus on different (possibly intricate) noise models, other data models in the literature show other advantages. Examples are, for instance, models with emphasis on learnable non-linearities to link latents to observables; we discuss non-linear ICA (Hyvärinen and Pajunen, 1999) and VAEs (Kingma and Welling, 2013) further below.

One example of studies which did investigate other than Bernoulli and Poisson distributions is the work by Khan et al. (2010) where a generalized mixture of factor analyzers model is presented that allows for addressing both continuous and discrete data (Gaussian and Multinomial distributions have been treated). Moreover, they established a new variational EM algorithm based on a lower bound of the Multinomial likelihood for fitting the model to the considered mixed data. Although the model can, in principle, deal with binary and categorical data along with the continuous case, it is not defined to subsume all distributions of the exponential family. Further examples are LVMs with the focus on estimating the noise distribution for observables of unknown types (Valera and Ghahramani, 2017; Molina et al., 2018; Vergari et al., 2019). Those approaches can typically select between a predefined set of observable distributions to estimate the best fit given a set of (usually) heterogeneous observations. Grounded on sum-product networks with assumptions such as piecewise polynomial leaf distributions, generalized Bayesian inference procedures are developed (e.g. Molina et al., 2018). Such approaches are consequently very different from approaches discussed above or the here investigated approach. They are related to the considered approach here as being relatively generally defined, however, and we will come back to common features when we discuss noise type estimation in Sec. 7.2.2.

Still other lines of research for unsupervised learning follow deep generative modeling approaches with variational autoencoders (VAEs; Kingma and Welling, 2013; Rezende et al., 2014) being among

the most popular such models. In order to relate latents to observables, VAEs make use of non-linear functions which are parameterized using deep neural networks (DNNs). The non-linear functions are used to define an encoding model that, in the standard case, performs amortized inference (the encoder) and a generative model (the decoder). VAEs owe much of their capabilities to the used DNNs, while prior and observation noise distributions for standard VAEs are very elementary (i.e., diagonal Gaussians). Nevertheless, VAEs can in principle be used to also model non-Gaussian observables, and standard stochastic gradient ascent for optimization is not limited to the Gaussian case. However, already optimization of Gaussian VAEs is considered challenging: Major challenges are inherited from DNN optimization (starting with choices for decoder architecture, many hyperparameters etc.) paired with the same choices for encoder DNN(s) and necessarily stochastic optimization. Concrete examples of non-Gaussian VAEs use one specific observation noise, usually in the form of elementary distributions such as Bernoulli or Poisson (e.g., Grønbech et al., 2020; Jang et al., 2017; Baby and Bourlard, 2021). The latents are for standard forms of VAEs continuous and usually Gaussian distributed. But variants with discrete latents have also been investigated (e.g., Jang et al., 2017; Maddison et al., 2017; Rolfe, 2017; Yin et al., 2019; Berliner et al., 2022; Drefs et al., 2022b).

In a related field, recent works on non-linear ICA approaches focus on identifiability which can be guaranteed under certain latent dependency conditions including temporal dependencies (see, e.g., Sprekeler et al., 2014; Hyvärinen and Morioka, 2016, 2017; Hyvärinen et al., 2019). Commonly for these approaches, a contrastive learning (CL) algorithm is applied in order to estimate the non-linear ICA model: Hyvärinen and Morioka (2016) use a time-contrastive learning (TCL), Hyvärinen and Morioka (2017) use a permutation-contrastive learning (PCL), and Hyvärinen et al. (2019) use a generalized form of CL. Identifiability in non-linear ICA is, in principle, possible also for non-Gaussian observation noise. Recent work explicitly specifies relatively general conditions for noise and identifiability in a framework for structured non-linear ICA (Hälvä et al., 2021). At the same time, the concrete realization of an algorithm within this framework uses Gaussian noise (which is explicitly stated as limitation). In another related work, Khemakhem et al. (2020) investigate an identifiable VAE (iVAE) that bridges the gap between VAEs and non-linear ICA. The approach, however, assumes the noise distribution to be known and additive which likewise restricts the applicability of the model to Gaussian distributed observables. A later extension of iVAE (termed pi-VAE), was proposed by Zhou and Wei (2020) which assumes Poisson noise for the observables. In the pi-VAE approach, the mean of the Poisson distribution is set by a link function that is defined in terms of a general invertible function.

Other than non-linear ICA and VAEs, there also have been investigations of more conventional deep models, e.g., on convolutional neural networks (CNNs), that focused on non-Gaussian observables. For example, contributions such as Remez et al. (2018, 2017) proposed an approach for denoising images corrupted with Gaussian and Poisson noises using supervised training. Also, the study by Tolooshams et al. (2020) considered observables of non-Gaussian distributions using convolutional autoencoders trained in both a supervised or an unsupervised manner. The method explicitly investigated the natural exponential family distributions which is a specific subset of the exponential family.

For supervised learning (i.e., regression), GLMs do explicitly consider different noise distributions and specifically distributions of the exponential family with more than one parameter. GLMs can, for instance, use a specific dependency of the observable variances on the mean to treat two-parameter distributions. However, regression approaches for specific distributions (such as the Beta

Table 1: Summary of related previous work as discussed in Secs. 1 and 4.

Acronym	Model Details	Reference
Examples of LVMs with a linear component structure		
EF-PCA	exp. family obser. / dist. with sufficient stat. proportional to y	Collins et al. (2002)
Bayes. EF-PCA	exp. family obser. / sampling-based optim. / dist. with sufficient stat. proportional to y	Mohamed et al. (2009)
EF-SC	exp. family obser. / MAP approx. / only one parameter dist. used	Lee et al. (2009)
Poiss. FA	Poiss. obser. / sampling-based optim.	Zhou et al. (2012)
Linear ICA	contin. obser. / non-Gaussian sources	Hyvärinen and Oja (2000)
Binary ICA	binary obser. / non-stationary sources / linear mixing of sources followed by binarization	Hyttinen et al. (2022)
NLSPCA	Poiss. obser. / non-local (sparse) PCA / designed for Poiss. denoising	Salmon et al. (2014)
SPDA	Poiss. SC / designed for Poiss. denoising	Giry and Elad (2014)
Examples of LVMs with a specific non-linear component structure		
non-linear ICA	non-linear blind source separa. / slow feature analy. / temporal correl.	Sprekeler et al. (2014)
non-linear ICA	exp. family sources / temporal non-stationary structure / TCL	Hyvärinen and Morioka (2016)
non-linear ICA	non-Gaussian and temporally dependent stationary sources / PCL	Hyvärinen and Morioka (2017)
non-linear ICA	conditionally independent sources (given an auxiliary var.) / gen. CL	Hyvärinen et al. (2019)
OCA	Gauss. obser. / occlusion non-linearity / var. to control the depth of objects	Henniges et al. (2014)
MCA	Poiss. obser. / truncated approximation / binary latents	Lücke and Sahani (2008)
ss-MCA	Gauss. obser. / trunc. approx. / non-linear and spike-and-slab SC	Shelton et al. (2012)
G-MCA	Gauss. obser. / truncated variational approximations / binary latents	Sheikh et al. (2019)
EF-MCA	exp. family obser. / generic M-steps and EVO / binary latents / L dictionaries	our contribution
Examples of related DNNs for Gaussian / non-Gaussian observables		
DCEA	natural exp. family obser. / trained both supervised and unsupervised	Tolooshams et al. (2020)
DenoiseNet	Poiss. obser. / designed for Poiss. denoising	Remez et al. (2017)
TVAE	Gauss. obser. / variational training based on EVO / binary latents	Drefs et al. (2022b)
iVAE	noisy (Gauss.) obser. / identifiable VAE (unifies non-linear ICA and VAE)	Khemakhem et al. (2020)
pi-VAE	Poiss. obser. / extends the iVAE model to Poisson noise	Zhou and Wei (2020)

distribution) still require further considerations related to GLM techniques in order to realize concrete regression algorithms (see, e.g., Ferrari and Cribari-Neto, 2004, for a concrete discussion).

In summary, the here considered LVM (while based on a standard bipartite graphical model) is exceptionally general in allowing for any (regular) exponential family distribution to be used to model observation noise. For a general treatment of such observables, the crucial challenge is, of course, the derivation of a generic parameter optimization procedure, which we will address in the following. Tab. 1 presents a short summary of the approaches discussed here.

5. Maximum Likelihood

Having defined the family of EF-MCA models in Sec. 3, we now seek parameters Θ that optimize the models for a given set of N independent and identically distributed (i.i.d.) data points $\vec{y}^{(1)}, \dots, \vec{y}^{(N)}$. We follow a standard maximum likelihood approach and seek parameters Θ that optimize the log-likelihood $\mathcal{L}(\Theta) = \sum_n \log(p(\vec{y}^{(n)} | \Theta))$. We use the EM algorithm (Dempster et al., 1977; Neal and Hinton, 1998) which instead of maximizing the log-likelihood directly, optimizes a lower bound known as the *free energy* (Neal and Hinton, 1998) or the *evidence lower bound* (ELBO; Mc Cauliffe and Blei, 2008). For the family of generative models defined in (8)-(10), the ELBO is given by:

$$\mathcal{F}(q, \Theta) := \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \left\{ \sum_d \log(p(y_d^{(n)}; \vec{\eta}_d(\vec{s}, \Theta))) + \sum_h \log(p(s_h | \Theta)) \right\} + \mathcal{H}(q), \quad (26)$$

where $q^{(n)}(\vec{s})$ for each n is a variational distribution, $q = (q^{(1)}, \dots, q^{(N)})$ and $\mathcal{H}(q)$ is the Shannon entropy term. The ELBO is the objective we seek to optimize, and its relation to the log-likelihood is,

for completeness, given in Appendix B.3. In our learning algorithm, the ELBO is optimized iteratively w.r.t. distributions q (the E-step) and w.r.t. the model parameters Θ (the M-step) as commonly done. It has been shown, e.g. by Neal and Hinton (1998), that the ELBO is maximized w.r.t. distributions q if, for each \vec{s} and n , we equate $q^{(n)}(\vec{s})$ to the posteriors, i.e. $q^{(n)}(\vec{s}) = p(\vec{s} | \vec{y}^{(n)}, \Theta)$. Indeed, the distributions $q^{(n)}(\vec{s})$ can be the exact posteriors for tractable models (we refer to this case as *full EM* or *exact EM*) or they can represent variational approximations (in this case we refer to *variational EM* and consequently *variational E-steps*). The terms (variational) free energy $\mathcal{F}(q, \Theta)$ or ELBO are more common for intractable models, i.e., if the $q^{(n)}$ are approximations of intractable exact posteriors.

For tractable models, the E-steps reduce to the computation of exact posteriors. For our EF-MCA model, a generic form to compute the E-steps is, therefore, already given by Bayes' rule as long as the model is sufficiently small scale (small H). For larger scale EF-MCA models, Sec. 5.2 will address how generic variational E-steps can be used. For all EF-MCA models, the central challenge that we have to address here is, however, the derivation of parameter update equations in the M-step, which is pertinent to the noise distribution. We therefore pursue such a task in the next section and importantly show that the non-linear link function (22) enables us to derive a set of concise and generally applicable updates that maximizes the ELBO (26).

5.1 Parameter Update Equations – The M-step

Following the standard optimization procedure, we will set the derivatives of $\mathcal{F}(q, \Theta)$ w.r.t. all model parameters to zero, and derive parameter update rules from the resulting equation systems. For this, first note that the derivatives of $\mathcal{F}(q, \Theta)$ will contain derivatives of $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$. Then, the following applies for these derivatives:

$$\frac{\partial}{\partial W_{dh}} \bar{W}_d(\vec{s}, \Theta) = \mathcal{A}_{dh}(\vec{s}, \Theta), \quad \frac{\partial}{\partial W_{dh}} \bar{V}_d(\vec{s}, \Theta) = 0, \quad (27)$$

$$\frac{\partial}{\partial V_{dh}} \bar{V}_d(\vec{s}, \Theta) = \mathcal{A}_{dh}(\vec{s}, \Theta), \quad \frac{\partial}{\partial V_{dh}} \bar{W}_d(\vec{s}, \Theta) = 0, \quad (28)$$

where

$$\mathcal{A}_{dh}(\vec{s}, \Theta) := \begin{cases} 1 & h = h(d, \vec{s}, \Theta) \\ 0 & \text{otherwise} \end{cases} \quad (29)$$

which follows simply from considering the cases $h = h(d, \vec{s}, \Theta)$ and $h \neq h(d, \vec{s}, \Theta)$ separately. We will later discuss the properties of the function $\mathcal{A}_{dh}(\vec{s}, \Theta)$ in Appendix A (in Lemma 1). Now, given the equations above, we present the main result of this study: We derive concise equations for W and V which guarantee that all derivatives of $\mathcal{F}(q, \Theta)$ w.r.t. W_{dh} and V_{dh} vanish. These equations can then be used for parameter updates in each M-step. For the sake of readability, however, we only provide a proof sketch of the theorem now and present the full proof in Appendix A.1. Also, note that we here discuss the case for $L = 2$, and the general case will be further presented in Sec. 6 and Appendix A.3.

Theorem 1 Consider an EF-MCA data model (8)-(10) with $p(y; \vec{\eta})$ being a regular exponential family distribution with $L = 2$. Furthermore, let the parameters Θ contain the matrices W and V with $D \times H$ entries and $\vec{\eta}_d(\vec{s}, \Theta)$ to be defined as in Eqns. (11) and (21)-(22). Then, the derivatives

of the ELBO (26) w.r.t. all matrix elements W_{dh} and V_{dh} are zero iff for all d and h :

$$W_{dh} = \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] T_1(y_d^{(n)})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]}, \quad (30)$$

and

$$V_{dh} = \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] T_2(y_d^{(n)})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]}, \quad (31)$$

where $\mathcal{A}_{dh}(\vec{s}, \Theta)$ is given by (29).

Proof Sketch To proof the theorem, we decisively make use of the relations in (5) and the maximum superposition defined in (11) and (21)-(22), and set the derivatives of the ELBO w.r.t. the dictionary elements W_{dh} and V_{dh} to zero. In short, if we abbreviate $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ by \bar{W}_d and \bar{V}_d , then for a single dictionary element W_{dh} we have (using the chain rule):

$$\begin{aligned} \frac{\partial}{\partial W_{dh}} \log(p(y_d; \tilde{\eta}_d(\vec{s}, \Theta))) &= \frac{\partial}{\partial W_{dh}} \log(p(y_d; \Phi(\bar{W}_d, \bar{V}_d))) \\ &= \sum_{l=1}^2 \mathcal{A}_{dh}(\vec{s}, \Theta) \left(\frac{\partial}{\partial w} \Phi_l(w, \bar{V}_d) \Big|_{w=\bar{W}_d} \right) \left(T_l(y_d) - \frac{\partial A(\vec{\eta})}{\partial \eta_l} \right) \Big|_{\vec{\eta}=\Phi(\bar{W}_d, \bar{V}_d)}. \end{aligned} \quad (32)$$

For this expression we now substitute $\frac{\partial A(\vec{\eta})}{\partial \eta_1} = \mathbb{E}_{p(y; \vec{\eta})}[T_1(y)]$ and $\frac{\partial A(\vec{\eta})}{\partial \eta_2} = \mathbb{E}_{p(y; \vec{\eta})}[T_2(y)]$ from Eqn. (5). Furthermore, we use that for any function g of \bar{W}_d and \bar{V}_d applies:

$$\mathcal{A}_{dh}(\vec{s}, \Theta) g(\bar{W}_d, \bar{V}_d) = \mathcal{A}_{dh}(\vec{s}, \Theta) g(W_{dh}, V_{dh}),$$

see Lemma 1 in Appendix A. We then obtain:

$$\begin{aligned} \frac{\partial}{\partial W_{dh}} \log(p(y_d; \tilde{\eta}_d(\vec{s}, \Theta))) &= \sum_{l=1}^2 \mathcal{A}_{dh}(\vec{s}, \Theta) \left(\frac{\partial}{\partial w} \Phi_l(w, V_{dh}) \Big|_{w=W_{dh}} \right) \\ &\quad \times \left(T_l(y_d) - \mathbb{E}_{p(y; \Phi(W_{dh}, V_{dh}))}[T_l(y)] \right). \end{aligned} \quad (33)$$

Now, using Eqn. (26) and the definition of the mean value parameters, $\vec{w} := \mathbb{E}_{p(y; \Phi(\vec{w}))}[\vec{T}(y)]$, we get:

$$\begin{aligned} \frac{\partial}{\partial W_{dh}} \mathcal{F}(q, \Theta) &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \left(\sum_{d'} \frac{\partial}{\partial W_{dh}} \log(p(y_{d'}^{(n)}; \tilde{\eta}_{d'}(\vec{s}, \Theta))) \right) \\ &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \sum_{l=1}^2 \mathcal{A}_{dh}(\vec{s}, \Theta) \left(\frac{\partial}{\partial w} \Phi_l(w, V_{dh}) \Big|_{w=W_{dh}} \right) \\ &\quad \times \left(T_l(y_d^{(n)}) - \mathbb{E}_{p(y; \Phi(W_{dh}, V_{dh}))}[T_l(y)] \right) \\ &= \sum_{l=1}^2 \left(\frac{\partial}{\partial w} \Phi_l(w, V_{dh}) \Big|_{w=W_{dh}} \right) \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \\ &\quad \times \left(T_l(y_d^{(n)}) - \mathbb{E}_{p(y; \Phi(W_{dh}, V_{dh}))}[T_l(y)] \right). \end{aligned} \quad (34)$$

The expression is equal to zero, i.e. $\frac{\partial \mathcal{F}(q, \Theta)}{\partial W_{dh}} = 0$, if the following equations are satisfied:

$$\sum_n \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta)] (T_1(y_d^{(n)}) - W_{dh}) = 0, \quad \text{and} \quad (35)$$

$$\sum_n \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta)] (T_2(y_d^{(n)}) - V_{dh}) = 0. \quad (36)$$

Then, the above equations yield (30) and (31), which completes the proof. ■

Fulfilling Eqns. (30) and (31) guarantees vanishing derivatives and provides a generally applicable approach for updating W and V in each M-step. We do remark, however, that we have not strictly proven that (30) and (31) correspond to a maximum (and not a minimum or a saddle point). In this respect, one can further investigate the second derivatives of the ELBO (26) at these stationary points to observe if Eqns. (30) and (31) do in fact correspond to a maximum. We will later discuss this point in Appendix A.2. Furthermore, we here emphasize that Eqns. (30) and (31) are valid for any regular two-parameter distribution of the exponential family which includes Gaussian, Gamma, Beta and many more, i.e., a large variety of noise models is covered. Importantly, the above theorem reveals that the same functional form is obtained for the parameters of all these distributions under the generative model (8)-(10) (note that we here considered the $L = 2$ case but the general case will be discussed further below). A straightforward outcome of the foregoing theorem is when the distribution does contain a sufficient statistic proportional to y , i.e. $T_1(y) = y$. This specific form yields a further simplification:

Corollary 1 *Prerequisites as for Theorem 1. If the distribution $p(y; \vec{\eta})$ has a sufficient statistic $T_1(y) = y$, then the condition for W_{dh} is given by:*

$$W_{dh} = \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta)] y_d^{(n)}}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta)]}. \quad (37)$$

The corollary is instructive about previous contributions: It explains why the update equation for the original MCA data model (Lücke and Sahani, 2008) (which used Poisson noise) and the update equation for the later MCA models (Lücke and Eggert, 2010; Bornschein et al., 2013; Sheikh et al., 2019) (which used Gaussian noise) are identical and given by (37).

Finally, observe that Eqns. (30) and (31) do not represent closed-form solutions for W and V since the right-hand-sides also depend on W and V through the function $\mathcal{A}_{dh}(\vec{s}, \Theta)$. We thus follow previous work (e.g., Lücke and Sahani, 2008; Lücke and Eggert, 2010), and use Eqns. (30) and (31) in the fixed-point sense and update (for each d and h):

$$W_{dh}^{\text{new}} = \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta^{\text{old}})] T_1(y_d^{(n)})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta^{\text{old}})]}, \quad (38)$$

and

$$V_{dh}^{\text{new}} = \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta^{\text{old}})] T_2(y_d^{(n)})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta^{\text{old}})]}, \quad (39)$$

where also $q^{(n)} = q^{(n)}(\vec{s}; \Theta^{\text{old}})$ depends on the old parameters. If repeated updates result in the values of W and V to converge, then the converged values fulfill (30) and (31), respectively.

To complete the M-step parameter updates, we can also derive updates for the prior parameters $\vec{\pi}$. Those derivations do not involve the specific form of the observables' distribution. We can therefore use previous derivations (e.g., Henniges et al., 2010; Puertas et al., 2010) and update $\vec{\pi}$ as follows:

$$\pi_h^{\text{new}} = \frac{1}{N} \sum_{n=1}^N \mathbb{E}_{q^{(n)}}[s_h], \quad h = 1, \dots, H. \quad (40)$$

Algorithm 1: EM for training EF-MCA data models (for $L = 2$ parameter distributions)

```

initialize model parameters  $\Theta = (\vec{\pi}, W, V)$ ;
repeat
    consider the inverse mapping  $\vec{\Phi}$  defined in (7);
    if the distribution  $p(y; \vec{\eta})$  has a sufficient statistic  $T_1(y) = y$  then
         $M(\Theta) = W$ ;
    else
        consider the function  $F$  in (21) based on the relation of the natural parameters  $\vec{\eta}$  and
        the mean of  $p(y; \vec{\eta})$ ;
         $M(\Theta) = F(W, V)$ ;
    end
    for each vector  $\vec{s}$  of the latent space do
        for  $d = 1 : D$  do
             $h(d, \vec{s}, \Theta) = \text{argmax}_h\{M_{dh}(\Theta)s_h\}$ ;
             $\bar{W}_d = W_{dh(d, \vec{s}, \Theta)}$  and  $\bar{V}_d = V_{dh(d, \vec{s}, \Theta)}$ ;
             $\vec{\eta}_d = \vec{\Phi}(\bar{W}_d, \bar{V}_d)$ ;
        end
        for  $n = 1 : N$  do
             $q^{(n)}(\vec{s}) = p(\vec{s} | \vec{y}^{(n)}, \Theta)$ ;
            for  $h = 1 : H$  and  $d = 1 : D$  do
                compute  $q^{(n)}(\vec{s})s_h$ ;
                compute  $q^{(n)}(\vec{s})\mathcal{A}_{dh}(\vec{s}, \Theta)$  where  $\mathcal{A}_{dh}(\vec{s}, \Theta)$  is defined in (29);
            end
        end
    end
    update parameters  $\Theta$  using (38)-(40);
until parameters  $\Theta$  have sufficiently converged;

```

5.2 Optimization of Variational Distributions – The E-step

The main focus of the previous section was the derivation of parameter updates in the M-step (Eqns. (38)-(40)). These M-step equations contain expectation values w.r.t. distributions $q^{(n)}(\vec{s})$ that can be the exact posteriors $p(\vec{s} | \vec{y}^{(n)}, \Theta)$ for sufficiently small scale models (small H). Larger H require approximate posteriors $q^{(n)}(\vec{s})$ as the computational cost of evaluating expectations

w.r.t. exact posteriors scales exponentially, i.e., with 2^H . To enable efficient approximations in variational E-steps, different approaches can be applied. For our purposes, we here use a variational approximation based on truncated posteriors as a family of variational distributions (Lücke and Eggert, 2010; Lücke and Sahani, 2008). Concretely, we use variational distributions $q^{(n)}$ defined as:

$$q^{(n)}(\vec{s} \mid \mathcal{K}^{(n)}, \Theta) := \frac{p(\vec{s} \mid \vec{y}^{(n)}, \Theta)}{\sum_{\vec{s}' \in \mathcal{K}^{(n)}} p(\vec{s}' \mid \vec{y}^{(n)}, \Theta)} \delta(\vec{s} \in \mathcal{K}^{(n)}), \quad n = 1, \dots, N, \quad (41)$$

where the set $\mathcal{K}^{(n)}$ contains, for each data point $\vec{y}^{(n)}$, a fixed number of hidden states \vec{s} , and where the indicator function $\delta(\vec{s} \in \mathcal{K}^{(n)})$ is equal to 1 if \vec{s} is in the set $\mathcal{K}^{(n)}$ (and 0 otherwise). Distributions (41) approximate full posteriors by truncating sums over the whole latent space to sums over those subsets $\mathcal{K}^{(n)}$ which ideally accumulate most of the posterior mass.

Variational approaches based on truncated posteriors have been shown to result in very efficient optimization for mixture models (Hirschberger et al., 2022; Exarchakis et al., 2022) as well as LVMs (e.g., Sheikh and Lücke, 2016; Shelton et al., 2017; Sheikh et al., 2019). When choosing $q^{(n)}$ as defined by (41) to approximate full posteriors, the expectation values $\mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]$ in (38) and (39) are, for the family of LVMs in (8)-(10), obtained as follows:

$$\begin{aligned} \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] &= \frac{\sum_{\vec{s} \in \mathcal{K}^{(n)}} p(\vec{s}, \vec{y}^{(n)} \mid \Theta) \mathcal{A}_{dh}(\vec{s}, \Theta)}{\sum_{\vec{s}' \in \mathcal{K}^{(n)}} p(\vec{s}', \vec{y}^{(n)} \mid \Theta)} \\ &= \frac{\sum_{\vec{s} \in \mathcal{K}^{(n)}} \exp \left(\sum_d ((\vec{\eta}_d(\vec{s}, \Theta))^T \vec{T}(y_d^{(n)}) - A(\vec{\eta}_d(\vec{s}, \Theta))) \right) \text{Bern}(\vec{s}; \vec{\pi}) \mathcal{A}_{dh}(\vec{s}, \Theta)}{\sum_{\vec{s}' \in \mathcal{K}^{(n)}} \exp \left(\sum_d ((\vec{\eta}_d(\vec{s}', \Theta))^T \vec{T}(y_d^{(n)}) - A(\vec{\eta}_d(\vec{s}', \Theta))) \right) \text{Bern}(\vec{s}'; \vec{\pi})}, \end{aligned} \quad (42)$$

where $\text{Bern}(\vec{s}; \vec{\pi})$ abbreviates the Bernoulli prior in (8). Likewise, the expectation values $\mathbb{E}_{q^{(n)}}[s_h]$ in (40) can be obtained (for each h) accordingly.

5.2.1 EVOLUTIONARY VARIATIONAL OPTIMIZATION (EVO)

Crucial for variational optimization of truncated posteriors (41) are strategies to efficiently and effectively populate the $\mathcal{K}^{(n)}$ sets with latent states capturing high posterior mass. To this end, here we apply a recently suggested strategy based on evolutionary algorithms (Evolutionary Variational Optimization, EVO; Drefs et al. 2022a). The approach treats the sets $\mathcal{K}^{(n)}$ in Eqns. (41) and (42) as variational parameters, which consequently yields the variational lower bound (26) to be reformulated as follows (see Lücke, 2019, for a concrete discussion):

$$\mathcal{F}(\mathcal{K}^{(1, \dots, N)}, \Theta) = \sum_n \log \left(\sum_{\vec{s} \in \mathcal{K}^{(n)}} p(\vec{y}^{(n)}, \vec{s} \mid \Theta) \right).$$

Based on this form, the variational lower bound can be increased by replacing $\vec{s} \in \mathcal{K}^{(n)}$ with $\vec{s}^{\text{new}} \notin \mathcal{K}^{(n)}$ for each n such that $p(\vec{y}^{(n)}, \vec{s}^{\text{new}} \mid \Theta) > p(\vec{y}^{(n)}, \vec{s} \mid \Theta)$. EVO, specifically, treats the bit vectors \vec{s} as genomes and uses them to evolve \vec{s}^{new} through application of elementary genetic

operations including selection, mutation and crossover. A requirement of EVO is that the model's joint $p(\vec{y}^{(n)}, \vec{s} | \Theta)$ is analytically tractable in order to define a fitness function for the selection step. Apart from that, EVO is generically applicable, i.e., the approach does not involve model-specific derivations (in contrast to, e.g., mean-field approaches; Jordan et al. 1999). For details, see (Drefs et al., 2022a,b).

For the family of LVMs we consider here, the required joints are directly defined by the chosen exponential family distributions. Consequently, if the sufficient statistics $\vec{T}(y)$, the log-partition $A(\vec{\eta})$ and the link $\vec{\eta}_d(\vec{s}, \Theta)$ are efficiently computable, then so are the expectation values (42) and the entire variationally accelerated learning algorithm.

6. Parametrization of the Proposed Generative Models – General Case

Let us now finally generalize the definition of the proposed family of generative models presented in (8)-(10) to any arbitrary L -parameter distribution. We remain using the independent Bernoulli distributions of (8). For the noise distribution given by (9) and (10), however, instead of two matrices W and V , we can in general consider L matrices (corresponding to the L parameters of an exponential family distribution). Let us denote these matrices by $W^{(1)}$ to $W^{(L)}$, i.e. $\Theta = (\vec{\pi}, W^{(1)}, \dots, W^{(L)})$. Therefore, we can generalize our definition of $\vec{\eta}_d(\vec{s}, \Theta)$, which is now given by:

$$\vec{\eta}_d(\vec{s}, \Theta) := \vec{\Phi}(\bar{W}_d^{(1)}(\vec{s}, \Theta), \dots, \bar{W}_d^{(L)}(\vec{s}, \Theta)). \quad (43)$$

Moreover, given L matrices $W^{(1)}$ to $W^{(L)}$ with $D \times H$ entries, we have to define the mappings $\bar{W}_d^{(1)}(\vec{s}, \Theta)$ to $\bar{W}_d^{(L)}(\vec{s}, \Theta)$ for each observable d . In analogy to the $L = 2$ case, we do so by defining a matrix $M(\Theta)$ as follows:

$$\forall d, h : M_{dh}(\Theta) := F(W_{dh}^{(1)}, \dots, W_{dh}^{(L)}) \text{ with } F(\vec{w}) = \mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[y], \quad (44)$$

where \vec{w} is assumed a row vector. Using the matrix $M(\Theta)$, we now define our mappings $\bar{W}_d^{(l)}(\vec{s}, \Theta)$ for $l = 1, \dots, L$:

$$\forall d, h : \bar{W}_d^{(l)}(\vec{s}, \Theta) := W_{dh(d, \vec{s}, \Theta)}^{(l)} \text{ with } h(d, \vec{s}, \Theta) := \operatorname{argmax}_h\{M_{dh}(\Theta) s_h\}. \quad (45)$$

Also in the general case, definitions (44) and (45) guarantee that always the *mean* of observable d is given by $\mu_d = \max_h\{M_{dh}(\Theta) s_h\}$. The above equations then finalize the definition of EF-MCA data models where the noise distribution can now be any arbitrary regular distribution of the exponential family. Then, for Eqns. (8)-(10) with (43)-(45), the following theorem applies which is a straightforward generalization of Theorem 1:

Theorem 2 Consider an EF-MCA data model (8)-(10) with $p(y; \vec{\eta})$ being a regular exponential family distribution with sufficient statistics $\vec{T}(y)$ of length $L \in \mathbb{N}$. Moreover, let the parameters Θ contain L matrices $W^{(1)}, \dots, W^{(L)}$ with $D \times H$ entries and let $\vec{\eta}_d(\vec{s}, \Theta)$ be defined as in (43)-(45). Then, the derivatives of the ELBO (26) w.r.t. all $W_{dh}^{(l)}$ are zero if for all d , h , and l :

$$W_{dh}^{(l)} = \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] T_l(y_d^{(n)})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]}, \quad (46)$$

where $\mathcal{A}_{dh}(\vec{s}, \Theta)$ is given by (29).

Proof The proof is a direct generalization of the proof of Theorem 1 and will be further presented in Appendix A.3. \blacksquare

Now, given the relations (46) for the matrices $W^{(l)}$, $l = 1, \dots, L$, the following update equations can be used (likewise in a fixed-point sense):

$$\forall d, h : (W_{dh}^{(l)})^{\text{new}} = \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta^{\text{old}})] T_l(y_d^{(n)})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta^{\text{old}})]}. \quad (47)$$

For the proposed EM algorithm, now in the general case, we initialize dictionaries $W_{dh}^{(l)}$ for $l = 1, \dots, L$ and update their elements in each M-step using (47). After convergence, the converged dictionaries represent the optimal (possibly local optima) parameters of the model. Moreover, note that the prior parameter update remains unchanged and is given by (40). Finally, also note that the application of EVO as variational acceleration is possible in the same way as for $L = 2$.

Having obtained a generic optimization procedure for general EF-MCA, let us finally emphasize again in what sense the procedure is directly applicable. Optimization for the general case iterates M-step equations (47) and (40) and the E-step (which can optionally use full posteriors or the truncated posteriors in (41)). Given a distribution of the exponential family, we require for those equations the following: The sufficient statistics $\vec{T}(y)$, the log-partition $A(\vec{\eta}_d(\vec{s}, \Theta))$, and the function $\mathcal{A}_{dh}(\vec{s}, \Theta)$. For essentially all exponential family distributions conventionally used as noise models in LVMs, the sufficient statistics and log-partition functions $A(\vec{\eta})$ are efficiently computable. Also the mapping $\vec{\eta}_d(\vec{s}, \Theta)$ from the standard parametrization to the natural parameters is usually a closed-form function. Only the function $\mathcal{A}_{dh}(\vec{s}, \Theta)$ of (29) can contain a potential analytical intractability. The function is defined using $h(d, \vec{s}, \Theta)$ of Eqns. (43) to (45) and involves the (L dimensional) mapping $\vec{\Phi}$ from the mean parameters to the natural parameters. For many examples the mapping is known to be closed-form (e.g., Bernoulli, Poisson, Exponential, Gaussian, Categorical etc.) or partly closed-form (e.g., Gamma). If the mapping is not closed-form (e.g., Beta, Gamma), then our optimization inherits precisely the same analytical intractability problem as parameter optimization for the single distribution itself already exhibits. However, the intractability problem is in all common cases low-dimensional (usually $L = 1$ or 2), and the function $\vec{\Phi}$ is smooth and one-to-one, i.e., analytical or numerical solutions are usually available.

The here derived generic optimization thus contrasts with more formal frameworks of EM-like optimization that are discussed in the context of exponential families and/or graphical models. Usually such frameworks rest on assumptions for the variational distributions such as mean-field assumptions (e.g. Saul et al., 1996; Opper and Saad, 2000); or assumptions for the generative model itself (Wainwright and Jordan, 2008). For instance, if we assume the generative model to have an exponential family representation of its *joint* distribution, it is (under mild conditions) straightforward³ to derive a general M-step which takes the form: $\sum_n (\mathbb{E}_{q^{(n)}} [\vec{T}(\vec{y}, \vec{s})] - \mathbb{E}_{p(\vec{y}, \vec{s} | \vec{\zeta}(\Theta))} [\vec{T}(\vec{y}, \vec{s})]) = \vec{0}$, where $\vec{T}(\vec{y}, \vec{s})$ denotes the joint sufficient statistics given an observation \vec{y} and a latent variable \vec{s} , and $\vec{\zeta}(\Theta)$ represents the natural parameter of the joint. Such a general M-step can serve theoretical considerations, or it can serve as a starting point to derive efficient optimization procedure (e.g.,

3. One needs to use the exponential family form of the joint. Then the derivative of $\mathcal{F}(q, \Theta)$ in (81) has to be computed using the equivalent of (5) for the joint distribution. Mild conditions have to apply for the parametrization from standard parameters Θ to natural parameters $\vec{\zeta}$ which are usually fulfilled.

if mean-field approaches are developed, Blei et al., 2017; Wainwright and Jordan, 2008). It does by itself not necessarily represent a directly applicable optimization procedure, however. In the case of the here considered EF-MCA models, a joint exponential family which is *genuine* (i.e., truly non-curved, see Efron, 1975; Bickel and Doksum, 2015; Efron, 2023) is not available, and differentiability of a (joint) partition function is not given because of the used maximum non-linearity. As a further example, energy models (compare Hinton et al., 2006; LeCun et al., 2006; Fischer and Igel, 2014) usually do have a joint exponential family distribution by definition. However, their partition function is usually computationally intractable.

7. Numerical Verification and Example Applications

We now numerically verify the derived optimization procedure using different well-known distributions of the exponential family as examples. Concretely, we use Bernoulli, Poisson and Exponential as one-parameter distributions, and Gaussian, Gamma and Beta as two-parameter distributions of the family, and optimize the corresponding EF-MCA data models with EM presented in Alg. 1. We use the results of Theorem 1 (and Theorem 2 as the general case) for the M-steps. For the E-steps, we use full posteriors when the used models are sufficiently small or apply approximations (41) otherwise⁴. For technical details, we refer to Appendix C, and also refer to Mousavi et al. (2021) for further examples of the approach applied to Beta-distributed observables.

7.1 Artificial Data

First, an example with artificial data and known data-generating parameters Θ^{gen} (the ground-truth parameters) was considered to verify and evaluate the derived optimization procedure. Specifically, using the setup of a bars test (Földiák, 1990; Lücke and Sahani, 2008; Drefs et al., 2022a), we were interested in how well the derived update equations recover ground-truth parameters. We generated data using the generative model (8)-(10) itself using link (11) with (21) and (22), and considered Exponential and Bernoulli distributions for observables to assess the efficacy of the proposed updates (38) and (40) in increasing the log-likelihood of the data. We refer to the corresponding EF-MCA models as Exponential-MCA (E-MCA) and Bernoulli-MCA (B-MCA) models, respectively.

For data generation and for the models, we used $H = 10$ generative fields \vec{W}_h in the form of horizontal and vertical bars each occupying 5 pixels on a $D = 5 \times 5$ grid. The latents were sampled independently (according to the prior) with probability π_h for $h = 1, \dots, H$, and the corresponding generative fields were superimposed non-linearly according to (22). We generated two sets of data points using the given EF-MCA models: One with Exponential distribution as the observation noise and another with Bernoulli distribution. We then trained the corresponding EF-MCA models using each of the generated datasets: E-MCA was trained using the data with Exponential noise and likewise B-MCA using the data with Bernoulli noise. Details of the experimental settings and parameter initializations are presented in Appendix C.1.

We repeated this procedure 100 times and closely examined the behaviour of the log-likelihood functions for each of the runs. The results are depicted in Fig. 3. As it can be seen, the application of the optimization procedure (Theorems 1 and 2) gives rise to a robust and reliable algorithm that in practice monotonically increases the log-likelihood of the data to at least local log-likelihood optima (note that we used full posteriors here). Also, the figure illustrates that the algorithms converge to

4. The codes are available in <https://github.com/tvlearn/evo>.

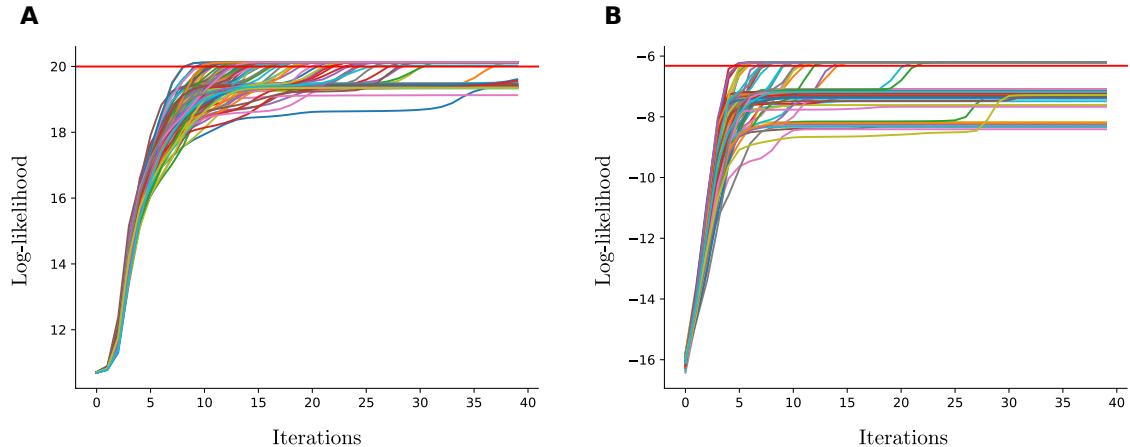


Figure 3: Behaviour of the log-likelihood functions corresponding to (A) E-MCA and (B) B-MCA for 100 runs using artificial bars test. Different colors denote different runs of the algorithms. Both algorithms result in a monotonic increase of the log-likelihood function and demonstrate fast convergence (at least to a local optimum). Out of 100 runs, 71 for E-MCA and 29 for B-MCA, respectively, converged to the global optimum (i.e., to the ground-truth log-likelihood denoted by the red horizontal line). The best runs achieved slightly higher log-likelihoods compared to the ground-truth, which may be attributed to the effect of overfitting due to the finite sample sizes.

ground-truth log-likelihood values (log-likelihood values computed using the generative parameters; red horizontal lines in Fig. 3) in some cases. In these runs, the learned parameters are very close to the generating ones with small differences that could be due to finite sample size effects (we have observed that increasing the number of data points instantly increases the estimation accuracy). Nevertheless, local optima can be observed for some cases where the learned log-likelihood values differ significantly from the ground-truth. This effect appears to be pronounced to varying degrees in different models. For instance, in the P-MCA model, we here observed 83 out of 100 runs that converged to global optimum. In comparison, the P-MCA approach of Lücke and Sahani (2008), when trained using exact EM, converged to global optimum in 92 out of 100 runs, i.e. more often compared to the P-MCA model studied here. Such a difference, however, is expected since the latter model exploits an additional annealing procedure to avoid local optima. While such annealing approaches have been commonly investigated in previous studies, our focus in this work remains on evaluating the monotonically increasing log-likelihood function for each of the EF-MCA data models (we will discuss further details in Appendix C.1). Also Fig. 4 displays the model parameters learned in one such run for the E-MCA algorithm.

Besides the local optima effect, we also found the sparseness (the value of $\sum_h \pi_h$ or simply πH if we use the same π for all h) to be an important factor influencing the outcome of the algorithm: The *reliability* of the algorithm (the probability of recovering all model parameters; compare, e.g., Spratling (1999)) showed to decrease as the sparseness increases (see Appendix C.1 for details).

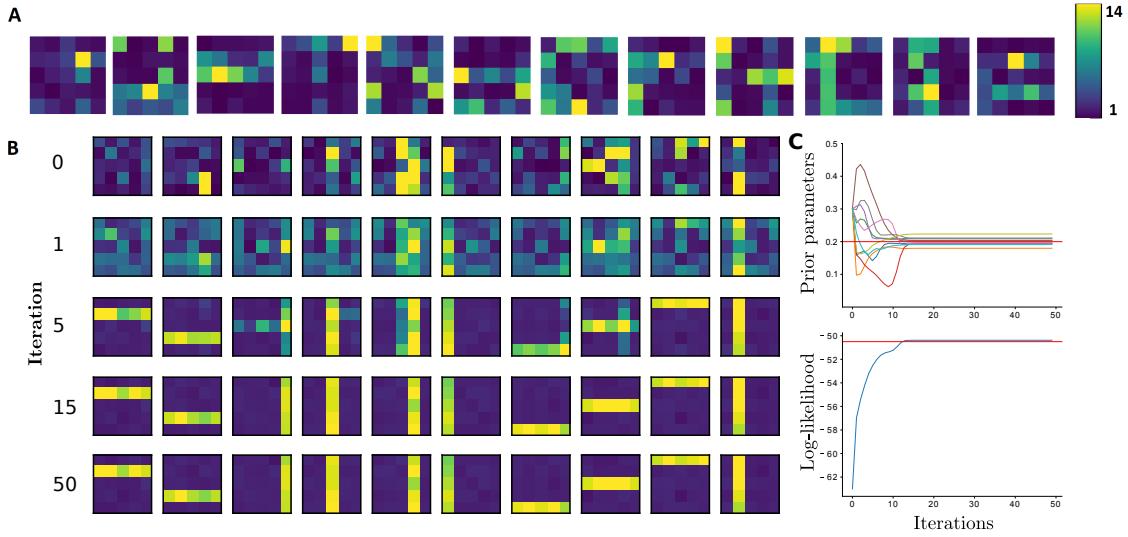


Figure 4: **A** 12 examples of input data. **B** Learned generative fields using the E-MCA model with the number of iterations illustrated on the left hand side. **C** Behaviour of the π_h values for $h = 1, \dots, H$ (top) and the log-likelihood function (bottom) given the 50 EM iterations. Here the generating parameter is $\pi_h^{\text{gen}} = 0.2$ for all h (i.e., the sparseness is 2); see Appendix C.1 for more details.

7.2 More Realistic Data

After the initial verification above, let us now point to some example applications on real data. Applications often require large models such that computing full posteriors becomes infeasible. To scale the algorithms to larger sizes, we applied variational E-steps as described in Sec. 5.2 using $|\mathcal{K}^{(n)}| = 60$ variational states for each n .

7.2.1 FEATURE EXTRACTION – NATURAL IMAGE PATCHES

For the first application to real data, we considered the model of Example 2, i.e., Gaussian-MCA. As discussed, the model has two matrices that we can optimize using Theorem 1: W for the mean and a combination of V and W for the variance of the Gaussian. To facilitate interpretation, we reparameterized the matrices back to the normal Gaussian parameters by setting $\sigma_{dh}^2 = V_{dh} - W_{dh}^2$ (see Appendix B.1 for details). Also, we followed Bornschein et al. (2013) and used a maximum magnitude non-linearity, i.e., we replaced $h(d, \vec{s}, \Theta)$ in (22) by (since $M(\Theta) = W$ for Gaussian):

$$h(d, \vec{s}, \Theta) = \operatorname{argmax}_h \{|W_{dh}s_h|\}, \quad (48)$$

where $|x|$ denotes the absolute value of x . Such choice does not affect our main results from Theorems 1 and 2 as the relations in (27)-(28) and also the proofs (in Appendix A.1) remain valid.

Using a set of $N = 100,000$ ZCA-whitened image patches of size $D = 20 \times 20$ pixels taken from a database of natural scenes (van Hateren and van der Schaaf, 1998), we trained a Gaussian-MCA model with $H = 1,000$ components to learn individual dictionaries for component means and component variances. After training, we observed a large variety of generative fields (GFs) for

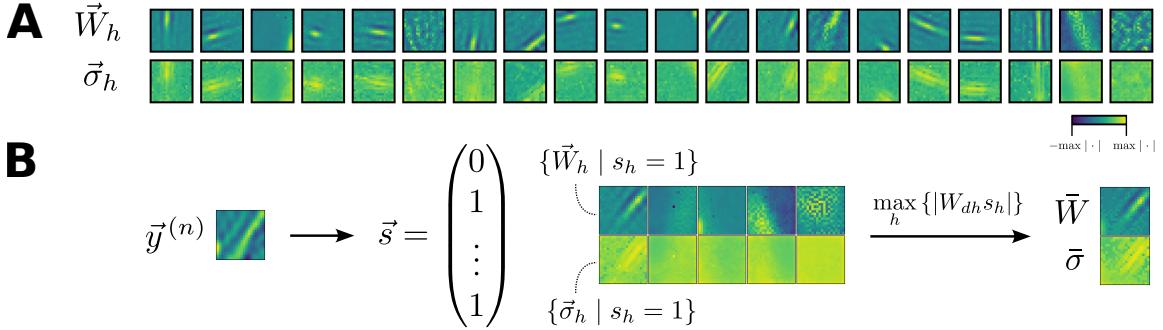


Figure 5: **A** Exemplary dictionary elements \vec{W}_h and $\vec{\sigma}_h$ learned by Gaussian-MCA on natural image patches of $D = 20 \times 20$ pixels (20 generative fields each; we depict standard deviations σ_{dh} instead of V_{dh} to facilitate interpretation). See Appendix C.2 for all $H = 1000$ dictionaries. **B** Illustration of data encoding. The left part depicts an exemplary training data point. The middle part depicts a latent vector \vec{s} with high posterior probability as well as the dictionary elements of the components h associated with the non-zero entries in \vec{s} . These components effectively determine the Gaussian mean \bar{W} and standard deviation $\bar{\sigma}$ through a point-wise maximum magnitude non-linearity as shown on the right part ($\bar{W} = \{W_{dh(d,\vec{s},\Theta)}\}_{d=1}^D$ and $\bar{\sigma} = \{\sigma_{dh(d,\vec{s},\Theta)}\}_{d=1}^D$ with $h(d, \vec{s}, \Theta)$ given by (48)).

the component means (including the familiar Gabor-like and globular fields that have likewise been observed in previous studies using MCA models; e.g., Lücke and Sahani 2008; Bornschein et al. 2013) together with a large variety of GFs for component variances. Figure 5 illustrates 20 examples of such GFs (see Appendix C.2 for the full dictionaries). The observed variety of variance GFs stands in contrast to a uniform variance with equal value for all latents and observables (i.e., $\sigma_{dh} = \sigma, \forall d, h$) as assumed by standard SC or previous MCA versions. Compared to such previous approaches, a generalized encoding capturing component means and component variances may provide a more elaborate model of the combination of first- and second-order statistics in real-world datasets. Such variance GFs may be used, e.g., to model intricate image structures, and/or to further analyze the effect of added noise on different parts of the image patches (also see Mousavi et al., 2020, 2021).

Besides, the data and model dimensionalities (i.e., the values of N, D, H) used here demonstrate the scalability of the proposed EF-MCA data models when accelerated by EVO (Sec. 5.2.1).

7.2.2 NOISE TYPE ESTIMATION

Next, we considered the problem of determining the unknown noise type of a given dataset. This problem is of importance for data analysis in a number of applications and has been actively researched during the last years (see Appendix C.3 for a discussion of related work). Here, the results of Theorems 1 and 2 enabled us to develop approaches for determining the type of the data distribution. As a proof-of-concept in this direction, we investigated how, e.g., Gaussian-, Gamma- and Beta-MCA models can be used to determine which of the three noise distributions is better suited for a given dataset. To quantify suitability, we compared variational lower bounds (as approximations to log-likelihoods) obtained with the three models on a given dataset. Note that the considered models have the same number of parameters such that we can directly use their lower bounds for model selection (i.e., we do not have to consider penalty terms for the number of parameters; see, e.g., MacKay (2003) for a comparison regarding AIC or BIC).

Table 2: Comparison of the obtained ELBO values by the three different EF-MCA data models when trained on images corrupted with three different noises: Gaussian, Gamma and Beta were investigated here. Corrupted images were linearly rescaled to the interval $(0, 1)$ for the sake of a reliable comparison. In addition, we trained all the models with the same settings of hyper-parameters. For the case of Gamma-MCA (and also Beta-MCA for some cases), we used the approach presented in Appendix C.1 to avoid local optima. The results correspond to the highest value of the ELBO for each model among three different runs, and the bold numbers represent the best fit (for each noisy image) among the three models. We refer the reader to Sec. 7.2.2 and Appendix C.3 for more details.

Noise	Model	House	Camera	Peppers	Saturn	Bridge	Flag	Average
Beta	Beta-MCA	39.00	-28.89	37.93	57.82	31.29	78.88	36.00
	Gamma-MCA	15.39	-29.02	25.80	58.96	16.23	97.25	30.76
	Gaussian-MCA	16.89	-42.36	14.73	32.31	7.99	66.00	15.92
Gamma	Beta-MCA	125.16	251.87	215.27	562.55	159.46	138.78	242.18
	Gamma-MCA	125.91	266.36	217.71	554.82	159.47	139.96	244.03
	Gaussian-MCA	116.38	232.57	200.05	534.19	143.00	118.13	224.05
Gaussian	Beta-MCA	123.43	143.77	119.66	150.23	116.20	146.02	133.21
	Gamma-MCA	121.70	143.15	118.50	148.65	114.11	143.97	131.68
	Gaussian-MCA	124.15	144.42	120.63	151.30	117.05	145.67	133.87

We investigated noise type estimation using both visual and acoustic data. We also refer the reader to Appendix C.3 for technical details of the experiments, Examples 1, 2 and Appendices B.1 and B.2 for detailed information regarding the parametrization of the Gaussian-MCA and the Gamma-MCA models, respectively, as well as to (Mousavi et al., 2021) for the details of the Beta-MCA.

Visual Data. We here considered standard benchmark images (depicted in Fig. 11), to which we either added Gaussian, Gamma or Beta noise. We then trained the three models of Gaussian-, Gamma- and Beta-MCA on each of the noisy images. In order to have a reliable comparison, we ran each algorithm three times (each time with a new set of initial parameters and for 100 variational EM iterations), and picked the one with the highest ELBO among these three runs. The results are presented in Tab. 2. As can be observed and expected, the lower bound is, in most cases, highest when the noise distribution assumed by the model matches the actual noise in the data. Factors potentially causing the few discrepancies between the best model (in terms of lower bound) and actual noise type, as well as the small differences between the lower bounds that observed in some cases may include, e.g., local optima effects and/or the noise levels which we have used for the experiments (we used noise level of 50 for the cases of Gaussian or Gamma noise and 0.1, 0.2 or 0.3 for the case of Beta noise).

Acoustic Data. As a further example for data with natural noise sources, we considered the CHiME dataset (Foster et al., 2015) and fitted Gaussian- and Gamma-MCA models to amplitude spectrograms of three selected audio files. For both algorithms, we performed three runs for each audio example, and for comparison then selected the one with the highest ELBO (which can be computed efficiently). The resulting ELBO values are shown in Tab. 3. The higher ELBOs observed for Gamma-MCA suggest the Gamma distribution as a better suited noise model compared to Gaussian for the considered data, which is consistent with amplitude data being positive. While

Table 3: Results of the noise type estimation experiment with natural noise sources (see Sec. 7.2.2 and Appendix C.3 for details). Listed are the highest ELBO values among three different runs given each of the considered audio examples.

	Gaussian-MCA	Gamma-MCA
audio example 1	215.20	298.53
audio example 2	95.61	250.77
audio example 3	47.36	206.14

differences of the ELBO values across runs showed to be comparably small, larger differences across audio examples and algorithms could be observed. These may be attributed to, among other things, differences between the content and noise sources in the individual audio files (see Appendix C.3 for data-related details).

The experiments reveal the efficacy of the updates (38) and (39) and demonstrate applicability of the algorithms for noise type estimation at large scales. Model selection using other types of noise distributions can proceed along the same line but require a more elaborate treatment (compare, e.g., Valera and Ghahramani, 2017) which exceeds the purposes of this study.

7.2.3 DENOISING

Finally, we used the presented algorithms to denoise images corrupted by non-Gaussian noises. This task can be considered more difficult compared to the removal of additive white Gaussian noise which numerous established denoising algorithms are optimized for (compare, e.g., Dabov et al., 2007; Gu et al., 2014; Burger et al., 2012; Chaudhury and Roy, 2017; Chen and Pock, 2017; Zhang et al., 2017; Tai et al., 2017; Zhang et al., 2018). We investigated two different noise types, namely Poisson and Exponential, and used standard test images for benchmarking (see Appendix C.4 and Fig. 11 for details on the images used). Concretely, we applied Poisson-MCA (P-MCA) to denoise images corrupted by Poisson noise and analogously Exponential-MCA (E-MCA) to denoise images corrupted by Exponential noise. The family of EF-MCA models allows for treating a much larger variety of noise distributions, of course, but focusing on P- and E-MCA may illustrate the potential also for other noise types. For denoising images with Beta noise (using the Beta-MCA model), the reader may be referred to Mousavi et al. (2021).

In general, one can see denoising as a (non-)linear inverse problem (compare Tarantola, 2005) where the task is to estimate the underlying non-noisy image using a specific method. Formally, given a set $Y_{\text{corrupted}} = f(Y_{\text{original}})$ of corrupted data points (f denotes a noise operator, in this case Poisson or Exponential), we aim at estimating the uncorrupted data Y_{original} by Y_{estimate} . A variety of different approaches have been investigated in the literature in order to either obtain the corresponding Y_{estimate} using a so-called *denoiser* (like the approaches mentioned above), or to somehow improve the performance of a given denoiser (either for denoising or for other relevant tasks) using established methods like the work by, e.g., Kadkhodaie and Simoncelli (2021) that uses the prior embedded within a denoiser. We here employed the proposed EF-MCA data models to restore the non-noisy images by *only* observing a noisy version of the clean images. That is, P-MCA and E-MCA models were directly applied to the noisy image itself (using overlapping patches) without leveraging external, clean training data. This task is also referred to as *zero-shot* denoising (compare, e.g., Shocher et al., 2018; Imamura et al., 2019). In detail, we followed Drefs et al. (2022a)

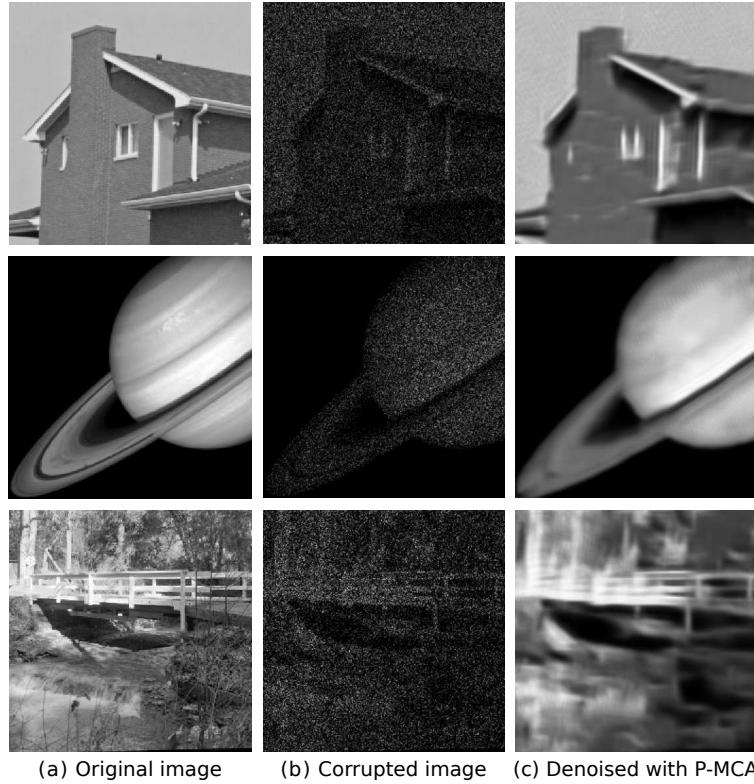


Figure 6: Denoising of the House, Saturn and Bridge images when corrupted by Poisson noise (at peak value 2). The figure illustrates the original and noisy images together with the reconstructed images using P-MCA trained with $H = 100$ and $D = 400$ for 100 variational EM iterations. The corresponding PSNR values of the reconstructed images are 24.66 (House), 28.39 (Saturn) and 20.15 (Bridge); while the PSNR values of the noisy images are 4.94 (House), 7.11 (Saturn) and 6.21 (Bridge). In addition, the denoised House image can be compared with Fig. 9 of Giryes and Elad (2014) which depicts the reconstructed images using BM3D, SPDA and NLSPCA methods.

and, given that $T(y) = y$ for Poisson and Exponential distributions (i.e., $M(\Theta) = W$), used the following formula⁵ to estimate the non-noisy image pixels:

$$(\vec{y}_{\text{estimate}}^{(n)})_d = \mathbb{E}_{q^{(n)}}[\bar{W}_d(\vec{s}, \Theta)], \quad (49)$$

where d corresponds to pixel d of the image patch n .

Poisson noise. After Gaussian denoising, Poisson noise removal benchmarks do presumably provide the most extensive opportunities for quantitative comparisons of different denoising algorithms. One possibility for denoising an image corrupted by Poisson noise is to simply ignore the specific properties of Poisson noise and employ a method that is tailored to Gaussian noise. For instance, BM3D (Dabov et al., 2007) can be applied directly. BM3D is a well-known denoising method which

5. Estimators of other distributions would use the corresponding $M(\Theta)$ matrix (see Mousavi et al., 2021, for denoising images corrupted by Beta noise).

often outperforms SC or k-SVD (Aharon et al., 2006) algorithms on image denoising benchmarks. The original method is tailored to 2D-images corrupted by Gaussian noise, however.

A frequently followed alternative approach is converting the Poisson noise to Gaussian, and then applying an algorithm which assumes Gaussian observables. Such methods apply a non-linear variance-stabilizing transformation (VST) such as Anscombe (Anscombe, 1948) or Fisz (Fisz, 1955) that produces a signal in which the noise can approximately be treated as Gaussian with unit variance. Approaches such as VST+BM3D use the Anscombe root transformation followed, in this case, by standard BM3D (e.g., Makitalo and Foi, 2010). Such approaches, however, are specific to Poisson noise as similar transformations are in general not available for all exponential family distributions. But also in the case of Poisson noise, transformation to Gaussian noise can be problematic. In fact, it is known that transformation-based methods can yield poor results for low-intensity signals (Salmon et al., 2014; Makitalo and Foi, 2010). It should be noted that the Poisson noise is not additive and the peak value of the original image determines the strength of the noise (the lower the peak is, the stronger the noise will be). For low peak values (e.g. lower than 3), VSTs are less effective. As a result, many approaches have been suggested to improve such transformations or exploit alternative approaches in the case of low-intensity signals (e.g., Salmon et al., 2014; Makitalo and Foi, 2010; Rond et al., 2016; Azzari and Foi, 2016; Giryes and Elad, 2014). One example is the P⁴IP approach (Rond et al., 2016) that applies a general plug-and-play prior approach on Poisson inverse problems. The method still relies on Gaussian-based denoising algorithms, however. Other studies (such as Makitalo and Foi, 2010; Azzari and Foi, 2016) attempt to improve the transformation techniques in order to increase the performance of the methods. A prominent work is an iterative VST framework, known as I+VST+BM3D, introduced by Azzari and Foi (2016) that, in a series of consecutive iterations, combines the noisy observation with a previously estimation of the image in order to improve the signal-to-noise ratio (SNR). The algorithm can then cope with extremely low SNR cases and produce state-of-the-art denoising results.

More relevant for comparison to our approach are methods that explicitly assume a Poisson noise. As discussed in the previous experiment, a data model that exploits the properties of the Poisson distribution (for the case of Poisson denoising) may be more suitable rather than a Gaussian-based model. In this direction, one important study is the work by Salmon et al. (NLSPCA; 2014) that uses Poisson PCA in which the link between latents and observables (as discussed before) is considered to be a weighted linear sum that sets the Poisson's natural parameter. In another study, Giryes and Elad (SPDA; 2014) introduce a Poisson sparse coding model capable of learning dictionaries from a Poisson distributed dataset. Their method likewise considers a linear superposition of the latents. In contrast, the P-MCA model that is used here uses maximization to directly set the means of the Poisson observables. Figure 6 illustrates the reconstructed 'House', 'Saturn' and 'Bridge' images using the P-MCA model. The estimated 'House' image can be further compared with Fig. 9 depicted by Giryes and Elad (2014) which presents the results of BM3D, SPDA and NLSPCA.

In addition to the approaches presented above, deep learning-based techniques have also been applied to Poisson denoising in recent years (see, e.g., Remez et al., 2017; Kumwilaisak et al., 2020; DeGuchy et al., 2019; Jin et al., 2018; Tolooshams et al., 2020). Importantly, Remez et al. (De-noiseNet; 2017) explored deep CNNs to denoise low-light images and reported state-of-the-art results for Poisson denoising at different peak values (also see Remez et al., 2018, for a further development). Another DNN-based algorithm for Poisson denoising is DCEA presented by Tolooshams et al. (2020) which, similar to DenoiseNet, considers a supervised setting for training. As mentioned earlier, the contribution considers distributions of the natural exponential family which includes the Poisson

Table 4: Comparison of the PSNR values (in terms of dB) for the considered Poisson denoising benchmarks. The algorithms are divided into two groups for peaks 0.2 and 0.5, and three groups for peaks 1 and 2 (values of DenoiseNet and DCEA-UC do not exist for peaks 0.2 and 0.5). NLSPCA, SPDA and P-MCA are LVMs that assume a Poisson noise while the others are task-specific and highly optimized algorithms that either assume a Gaussian denoiser (the BM3D-based methods and P⁴IP) or a DNN-based approach that is trained supervised (DenoiseNet and DCEA-UC). For each peak value and each image, the bold numbers denote the best PSNRs in comparison to the other models of the corresponding subgroup; and the underlined number denotes the best PSNR among all models. Listed values are averages over five noise realizations. See Secs. 7.2.3 and C.4 for details.

Method	Peak	House	Camera	Peppers	Saturn	Bridge	Flag
BM3D	0.2	18.37	17.35	17.10	22.02	17.09	14.28
VST+BM3D		17.79	16.90	16.96	21.38	17.12	13.53
I+VST+BM3D		19.68	18.40	<u>17.54</u>	23.13	<u>18.13</u>	17.49
P ⁴ IP		19.48	17.82	17.31	23.05	17.54	14.82
SPDA		17.83	16.93	16.75	21.52	16.80	16.58
NLSPCA		18.91	17.79	17.45	22.90	17.46	16.48
P-MCA ($H = 20$)		<u>19.69</u>	18.19	17.35	23.00	17.45	16.75
P-MCA ($H = 30$)		19.36	18.48	17.46	<u>23.30</u>	17.51	17.09
BM3D	0.5	20.27	18.83	18.49	23.86	18.24	15.87
VST+BM3D		19.61	18.46	18.41	23.75	18.26	15.58
I+VST+BM3D		21.54	<u>19.79</u>	19.05	25.78	<u>19.08</u>	18.60
P ⁴ IP		20.93	19.27	18.86	25.19	18.47	16.50
SPDA		20.51	18.90	18.66	25.50	18.46	19.67
NLSPCA		20.85	19.23	18.78	24.91	18.50	18.80
P-MCA ($H = 20$)		21.58	19.49	18.96	<u>25.79</u>	18.52	18.50
P-MCA ($H = 30$)		21.86	19.67	18.94	25.48	18.47	19.06
DenoiseNet	1	22.87	<u>21.59</u>	21.43	26.26	19.83	19.45
DCEA-UC		23.00	21.47	20.91	-	19.87	-
BM3D		22.35	20.37	19.89	25.89	19.22	18.31
VST+BM3D		21.64	20.19	19.71	25.82	19.43	18.46
I+VST+BM3D		23.04	21.07	20.44	27.27	19.86	19.74
P ⁴ IP		22.67	20.54	20.07	27.05	19.31	19.07
SPDA		22.73	20.23	19.99	27.02	19.20	22.54
NLSPCA		22.09	20.32	19.62	26.89	18.94	20.26
P-MCA ($H = 100$)		<u>23.42</u>	20.54	20.35	<u>27.46</u>	19.34	21.28
DenoiseNet	2	24.77	23.25	23.19	28.37	20.80	21.38
DCEA-UC		24.52	22.94	22.94	-	20.82	-
BM3D		24.18	22.13	21.97	27.42	20.31	20.81
VST+BM3D		23.79	21.97	22.02	27.95	20.49	20.79
I+VST+BM3D		24.62	22.25	21.93	28.85	20.69	21.18
P ⁴ IP		24.65	21.87	21.33	28.93	20.16	21.04
SPDA		<u>25.09</u>	21.54	21.23	<u>29.38</u>	20.15	24.92
NLSPCA		23.86	20.76	20.52	28.22	19.47	20.86
P-MCA ($H = 100$)		24.70	21.16	21.44	28.64	20.11	21.41

distribution. Tolooshams et al. investigated their approach in two different settings: Constrained and unconstrained DCEA, referred to as DCEA-C and DCEA-UN, respectively. Also compare related work by Kumwilaisak et al. (2020) who use multi-directional long-short term memory (LSTM) networks along with the CNNs. Nevertheless, there are substantial differences between, for instance, DenoiseNet or DCEA-UN, on the one hand, and P-MCA, on the other. P-MCA is trained directly on the noisy image and the approach does not require a-priori information about the peak value of the corrupted data; this contrasts with DenoiseNet or DCEA-UN which are both trained supervised.

Table 4 presents a quantitative performance comparison of different approaches using the standard measure of peak-signal-to-noise-ratio (PSNR). The methods we chose for comparison represent, to our best knowledge, state-of-the-art results for Poisson denoising at different peak values. We here chose four peak values and reported results for P-MCA considering different latent dimensionalities H and using models with $D = 20 \times 20$ observables. In general, we observed that lower values of H produce better results for low peak values while larger H values result in better performance for high peak values. An explanation could be that the information content of noisy data is lower (if the amount of data points remains the same) such that it becomes difficult to estimate the large number of parameters as required for a large H . Consequently, such a hyper-parameter can be further optimized for our learning algorithm (observe the different H values for P-MCA in Tab. 4).

As it can be observed in Tab. 4, P-MCA produces competitive results and, in some cases, outperforms the other models; but it can also be outperformed, e.g., by I+VST+BM3D, DenoiseNet or DCEA-UN in different settings. At a closer inspection, however, these approaches are using considerable fine-tuning specific to images. Furthermore, the major difference compared to DenoiseNet or DCEA-UN is their supervised settings where they use clean (non-noisy) data for training. P-MCA is, like SPDA and NLSPCA, not tailored to images. All these three approaches also have in common that they are directly based on the assumption of a Poisson distribution for the observables. Considering Tab. 4, P-MCA shows, in comparison to SPDA and NLSPCA, improved performance in terms of PSNR values in many of the investigated settings (especially at very low peak values). Such improvements for one example of the exponential family (P-MCA) may argue in favor of the general approach represented by the proposed LVMs and their optimization based on Theorems 1 and 2. It should be mentioned that we did not rigorously optimize the performance of the P-MCA model (and the other EF-MCA algorithms applied here) using, e.g., extensive parameter tuning. In general, task performance could be improved, e.g., by exploiting task-specific patch-averaging procedures or additional methods to avoid local optima.

Besides the comparisons presented above, we also compared the performance of the P-MCA model with other similar Gaussian-based LVMs trained with EVO. To this, we considered a version of MCA studied in Sheikh et al. (2019), binary sparse coding (BSC; Henniges et al., 2010), truncated variational autoencoder (TVAE; Drefs et al., 2022b) and evolutionary spike-and-slab sparse coding (ES3C; Drefs et al. (2022a); also compare Sheikh et al. (2014); Goodfellow et al. (2012)). These are all linear or non-linear LVMs that use binary latents (except ES3C that uses a spike-and-slab prior) and we have trained all of them using EVO (Drefs et al., 2022a) with identical hyper-parameters (the details are discussed in Appendix C.4). Figure 7 depicts the performance comparison of these approaches for Poisson denoising at peak value 1 (also see Fig. 12). As observed, P-MCA outperforms all the other models (that are tailored to the Gaussian noise) in both achieving higher ELBO values (although the models are not directly comparable due to the different number of parameters), and also higher PSNRs for the reconstruction. We also repeated the experiment at the peak value of 4. In that case, TVAE obtained the highest PSNRs followed by P-MCA being the second best model.

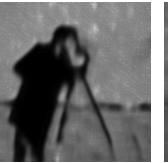
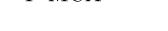
Model	Measure	House	Camera	Peppers				
BSC	PSNR	22.40	19.91	19.47		Noisy		BSC
	ELBO	-463.70	-418.86	-431.05				
ES3C	PSNR	22.37	19.78	19.60		ES3C		TVAE
	ELBO	-458.23	-414.08	-426.98				
TVAE	PSNR	22.84	20.13	19.57		MCA		P-MCA
	ELBO	-459.05	-416.48	-428.83				
MCA	PSNR	22.66	19.78	19.33		P-MCA		P-MCA
	ELBO	-457.14	-413.17	-426.07				

Figure 7: Comparison of the PSNR and ELBO values obtained by the considered LVMs for Poisson denoising at peak 1. The results correspond to one run of each algorithm. Best values are highlighted.

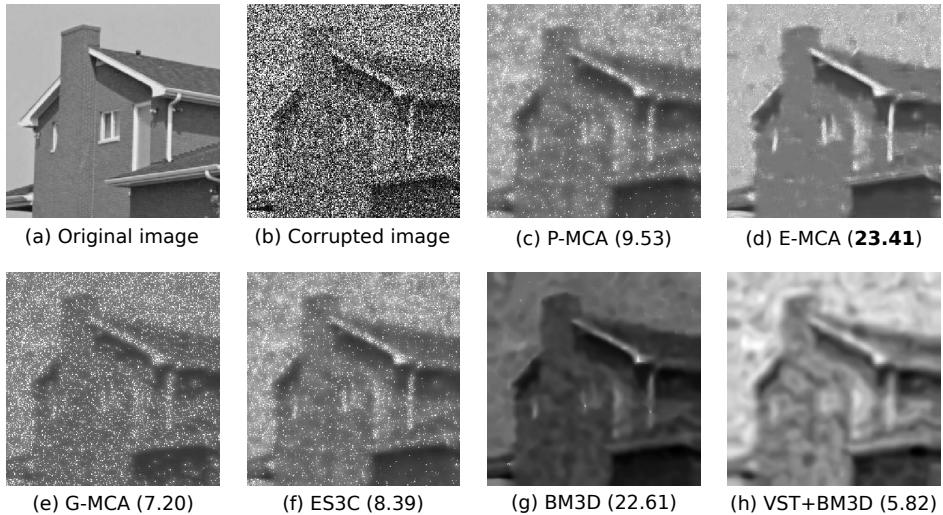


Figure 8: Denoising of the House image when corrupted by Exponential noise (we did not rescale the image before adding noise). Depicted are (a) original, (b) corrupted and (c)-(h) reconstructed images and the corresponding PSNRs in terms of dB obtained by different models (see text for details). Here, E-MCA achieves the highest PSNR in comparison to the others which is consistent with its Exponential noise assumption. Note that P-MCA and E-MCA use one-parameter distributions as noise models and only use the information of one dictionary (related to the mean) while the others (G-MCA, ES3C and BM3D) use additional information corresponding to the variance parameter.

Exponential noise. Exponential noise is treated much less frequently such that a comparison similar to Tab. 4 is not possible. Instead, we here used the example of Exponential noise to highlight the importance of using the correct noise distribution. That is, we exploited E-MCA as an instance of our LVMs with Exponential noise together with P-MCA and G-MCA (i.e., models assuming Poisson and Gaussian noise, respectively) for denoising the House image corrupted by Exponential noise. Additionally, we further applied the BM3D, VST+BM3D and ES3C algorithms. As illustrated in Fig. 8, E-MCA achieved the highest PSNR compared to all other approaches, which could be expected because E-MCA assumes the true image noise. Also, BM3D performed reasonably well;

however, in combination with the Anscombe transformation, the algorithm (VST+BM3D) achieved the lowest PSNR, which may be attributed to the Anscombe transformation being specific to Poisson noise. Visual inspection may even argue in favor of using the right noise model more than what PSNRs would suggest. In fact, Fig. 8 shows larger differences between the reconstructed images by E-MCA and BM3D than what may be expected by considering the PSNRs of the two approaches.

8. Discussion

Latent variable models (LVMs) have found very wide-spread use in Machine Learning, Statistics and Artificial Intelligence applications. A central element of basically all LVM approaches is the procedure which optimizes an LVM’s parameters given a set of data points. For non-Gaussian observables, optimization is typically much more challenging than for Gaussian observables, and especially so when we seek generic optimization methods for a broad class of observable distributions.

Here we studied an example of LVMs with a very generally defined range of observable distributions. By using binary latents in combination with maximization in order to link latents to the means of observables, we could address parameter optimization in general. While it would be challenging to use gradient based approaches for the maximum, the maximum combination of latents has convenient analytical properties. In our study, these properties allowed for the derivation of the main result: A general set of concise parameter update equations applicable to any (regular and minimal) distribution of the exponential family. For small scale problems, optimization can be performed based on full posteriors. For larger scale problems, we applied a novel variational acceleration (EVO; Drefs et al., 2022a), which is likewise generically and directly applicable (see Sec. 5.2.1).

The most closely related approach, EF-SC (Lee et al., 2009), used MAP-based optimization and distributions of the exponential family with a weighted linear sum to define the natural parameters of the used observation noise. Such a link from latents to observables has advantages in terms of MAP-based training (mono-modal posteriors are maintained). A disadvantage is that different observable distributions essentially define different non-linear superposition models, i.e., the observable means depend differently on the weighted summation of latents for different observation distributions. Also, how to realize a link from latents to two (or more) parameter distributions seems to be an open question for previous approaches. At least (and as previously stated), neither EF-SC nor PCA-like approaches which use the same link (Collins et al., 2002; Mohamed et al., 2009; Salmon et al., 2014) report numerical results for exponential family distributions with more than one parameter. Additionally, EF-SC would not be able to learn prior parameters because of the used MAP approaches (additional cross-validation would be required). In contrast, the family of LVMs considered here is generally defined with parameter update equations (Theorems 1 and 2) that are directly applicable to any choice of a (regular and minimal) exponential family distribution. The used exact inference (for small H) or variational approximation (for large H), furthermore, allows for learning of prior parameters. For the approach investigated here, we consequently argue that it represents an exceptionally flexible and generally applicable LVM approach for unsupervised learning. It encompasses, in principle, any choice of an exponential family distribution for its observables. As such, it allows for using many more distributions of the family than have here or elsewhere been considered including Categorical, Dirichlet, normal-gamma etc. Preliminary example algorithms based on one- and two-parameter distributions of the exponential family suggest that feasible and competitive learning algorithms are obtained when applying the results derived here.

The LVMs considered here share with the generalized linear model (GLM; Nelder and Wedderburn, 1972) that both approaches are relatively generally defined. For the here considered LVMs (EF-MCA), Theorems 1 and 2 in conjunction with EVO represent a generic optimization procedure; for GLMs an example of a generic optimization is based on iterative reweighted least square optimization which results in generally applicable (iterative) parameter updates (McCullagh and Nelder, 1989). The importance of the latter and similar approaches is reflected by a large variety of software packages for GLMs (Kosmidis, 2019). The class of EF-MCA models and the class of GLM models differ very significantly, however. Most salient is the focus of GLMs on regression compared to the focus on unsupervised learning for EF-MCA. However, also taking this difference aside, GLMs and EF-MCA strongly differ by how latents are linked to observables. GLMs are by definition (and by name) using weighted summation while EF-MCA uses weighted maximization. Furthermore, GLMs usually consider continuous latents while EF-MCA considers binary latents. Finally, also the treatment of multiple parameter distributions is different: GLMs can be defined with observable variances depending on the mean in a specific way, while the focus remains regression through modeling of the mean dependence. For EF-MCA, multiple matrices are used to link latents and observables (see Fig. 1) in the case when observable distributions with multiple parameters are used (compare Fig. 5). The use of multiple matrices is also a distinguishing feature compared to the other unsupervised learning approaches with more generally defined observable distributions (Collins et al., 2002; Mohamed et al., 2009; Lee et al., 2009).

Nevertheless, the central results here (i.e., the generic parameter update equations discussed in Theorems 1 and 2) are based on the assumption of binary latents and a maximum superposition, which may limit the application of our EF-MCA data models. While for many types of data, summation is a plausible assumption (e.g. acoustic waveforms for acoustic data), it has been argued that, e.g., for images or spectrogram representations of sounds, a maximum superposition from latents to observables is more closely aligned with the true generating process (see, e.g., Roweis, 2003; Lücke and Sahani, 2008; Bornschein et al., 2013; Frolov et al., 2015; Sheikh et al., 2019, for a discussion). Likewise, discrete latents may represent more realistic modeling choices in some cases (e.g. for diseases/symptoms relations, Singliar and Hauskrecht, 2006; Jernite et al., 2013; Mousavi et al., 2021) while for other data continuous latents are clearly more suitable.

In order to numerically investigate the derived optimization procedure, we used example applications which included structure finding in image patches, automatic estimation of noise distributions (also compare Vergari et al., 2019; Valera and Ghahramani, 2017), and denoising of images subject to non-Gaussian noises. In the case of Poisson noise, where a number of alternative approaches are available for comparison, we observed competitive performance. In particular, denoising results improved on models which used Poisson noise with linear superposition, namely the P-MCA algorithm improved on Non-Local Sparse Principal Component Analysis (NLSPCA; Salmon et al., 2014) and the Sparse Poisson Denoising Algorithm (SPDA; Giryes and Elad, 2014). Comparison to NLSPCA and SPDA may be particularly interesting in the context of this work. Both NLSPCA and SPDA use the standard linear link to observables via the natural parameter of the Poisson distribution, whereas P-MCA sets the mean of the Poisson using maximization. Besides, NLSPCA and SDPA exploit continuous latents while P-MCA uses binary latents. Tab. 4 shows, in general, a better denoising performance of P-MCA compared to both NLSPCA and SDPA at low-SNR values (although SPDA also yields very competitive results in some cases). Poisson noise is just one example of an exponential family distribution, and denoising is just one application. Still, we note that P-MCA is competitive for the task without using additional, task-specific mechanisms (compare NLSPCA and SDPA) and

without extensive fine-tuning of hyper-parameters. Further applications of our approach included observables following Bernoulli and Exponential distributions (both one-parameter distributions) and, importantly, observables following Gaussian, Gamma and Beta distributions (as two-parameter distributions). The Beta distribution has also been considered in an application of our results to medical data of hearing impairments (Mousavi et al., 2021).

In analogy to the literature on standard SC, further application examples would include inpainting, compression, feature learning, or compressed sensing using in principle any regular exponential family as observable distribution. Future work may also include the use of more richly structured prior distributions. As the derived update equations apply in general for any binary latents, the independent Bernoulli prior could also be replaced by more general distributions for binary variables (e.g. Dai et al., 2013) including those parameterized by deep models. Binary latents for deep generative graphical models are very common such that, e.g., deep SBNs (Saul et al., 1996; Gan et al., 2015) or deep restricted Boltzmann Machines (Larochelle and Bengio, 2008) could be used as well as deep models with deterministic non-linearities (e.g. Jang et al., 2017). Future research may also investigate how the current results can be extended to other forms of non-linear superpositions. The non-linearity (48) or an occlusion non-linearity as discussed by Dai and Lücke (2014) do maintain crucial properties required for Theorems 1 and 2, for instance. And non-linear superposition models could go further in the direction of deep models, e.g., by using repeated maximization with several layers of stochastic binary latents. Finally, also hybrid approaches combining results of LVMs with binary and continuous latents (cf. Sec. 4) could represent a promising future research direction.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within the excellence initiative (cluster of excellence H4a 2.0, EXC 2177/1 - project 390895286) and by the frame of the priority program SPP 2298 "Theoretical Foundations of Deep Learning" - project 464104047 (HM and JL). Furthermore, the contributions by FH were funded by the German Ministry of Research and Education (BMBF) in project 05M2020 (SPAplus, TP3), and the contributions by JD were funded within the DFG project 352015383 (HAPPAA, CRC 1330, B2). Large-scale numerical experiments were supported by the HPC Cluster of Oldenburg University (INST 184/225-1 FUGG) and by the HLRN network of HPC clusters (project nim00006).

Appendix A. Proof of the Main Theorems

We here provide the full proofs of Theorems 1 and 2 as well as further details. First consider the function $\mathcal{A}_{dh}(\vec{s}, \Theta)$ defined in (29). We show that the function has a useful property which is exploited in our proofs. For special cases of EF-MCA data models, this property has been observed before (Lücke and Sahani, 2008; Lücke and Eggert, 2010) but the following Lemma represents the required generalization for exponential family observables.

Lemma 1 *Consider $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ that are defined in (22). Then, for any well-behaved function g and any arbitrary $\vec{s} \in \{0, 1\}^H$, we have*

$$\mathcal{A}_{dh}(\vec{s}, \Theta) g(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)) = \mathcal{A}_{dh}(\vec{s}, \Theta) g(W_{dh}, V_{dh}). \quad (50)$$

Proof Given a vector \vec{s} , for each pair (d_o, h_o) either of the following applies:

$$h_o = h(d_o, \vec{s}, \Theta), \quad \text{or} \quad h_o \neq h(d_o, \vec{s}, \Theta),$$

where $h(d_o, \vec{s}, \Theta) = \operatorname{argmax}_h \{M_{d_o h}(\Theta) s_h\}$. First, let $h_o = h(d_o, \vec{s}, \Theta)$. It follows from (22) that

$$\begin{aligned} \mathcal{A}_{d_o h_o}(\vec{s}, \Theta) g(\bar{W}_{d_o}(\vec{s}, \Theta), \bar{V}_{d_o}(\vec{s}, \Theta)) &= \mathcal{A}_{d_o h_o}(\vec{s}, \Theta) g(W_{d_o h(d_o, \vec{s}, \Theta)}, V_{d_o h(d_o, \vec{s}, \Theta)}) \\ &= \mathcal{A}_{d_o h_o}(\vec{s}, \Theta) g(W_{d_o h_o}, V_{d_o h_o}). \end{aligned}$$

On the other hand, it follows from $h_o \neq h(d_o, \vec{s}, \Theta)$ and (29) that $\mathcal{A}_{d_o h_o}(\vec{s}, \Theta) = 0$ which trivially satisfies the claim in (50). \blacksquare

The lemma illustrates how multiplication by $\mathcal{A}_{dh}(\vec{s}, \Theta)$ simplifies the left-hand-side of the equation (50) to the right-hand-side expression in which the second factor is independent of \vec{s} . The function g can be any arbitrary function with arguments $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$. Moreover, it should be noted that we here assumed the case of $L = 2$ and presented the lemma using $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$. For the case of arbitrary L , the straightforward generalization of the lemma applies by considering the definition (45) and a function g with argument(s) $\bar{W}_d^{(1)}(\vec{s}, \Theta), \dots, \bar{W}_d^{(L)}(\vec{s}, \Theta)$. The general case will be used for the proof of Theorem 2.

A.1 Parameter Update Equations – Two-Parameters Case

We first prove Theorem 1 and show that the derivatives of $\mathcal{F}(q, \Theta)$ w.r.t. W_{dh} and V_{dh} yields the corresponding equations for W and V in (30) and (31).

Proof of Theorem 1 Consider a single dictionary element W_{dh} and, for the sake of brevity, let us abbreviate $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ by \bar{W}_d and \bar{V}_d , respectively. Then using the chain rule and Eqns. (11) and (27)-(28), we obtain:

$$\begin{aligned} \frac{\partial}{\partial W_{dh}} \log(p(y_d; \vec{\eta}_d(\vec{s}, \Theta))) &= \frac{\partial}{\partial W_{dh}} \log(p(y_d; \vec{\Phi}(\bar{W}_d, \bar{V}_d))) \\ &= \sum_{l=1}^2 \left(\frac{\partial}{\partial W_{dh}} \Phi_l(\bar{W}_d, \bar{V}_d) \right) \left(\frac{\partial}{\partial \eta_l} \log(p(y_d; \vec{\eta})) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d, \bar{V}_d)} \right) \\ &= \sum_{l=1}^2 \left(\frac{\partial}{\partial W_{dh}} \bar{W}_d \right) \left(\frac{\partial}{\partial w} \Phi_l(w, \bar{V}_d) \Big|_{w=\bar{W}_d} \right) \left(\frac{\partial}{\partial \eta_l} \log(p(y_d; \vec{\eta})) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d, \bar{V}_d)} \right) \\ &\quad + \sum_{l=1}^2 \underbrace{\left(\frac{\partial}{\partial W_{dh}} \bar{V}_d \right)}_0 \left(\frac{\partial}{\partial v} \Phi_l(\bar{W}_d, v) \Big|_{v=\bar{V}_d} \right) \left(\frac{\partial}{\partial \eta_l} \log(p(y_d; \vec{\eta})) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d, \bar{V}_d)} \right) \\ &= \sum_{l=1}^2 \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_l(w, \bar{V}_d) \Big|_{w=\bar{W}_d} \right) \left(T_l(y_d) - \frac{\partial A(\vec{\eta})}{\partial \eta_l} \right) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d, \bar{V}_d)}. \end{aligned} \tag{51}$$

Moreover, from (5) we know that for any regular distribution (i.e. with finite $A(\vec{\eta})$) of the exponential family, $A(\vec{\eta})$ satisfies:

$$\frac{\partial A(\vec{\eta})}{\partial \eta_1} = \mathbb{E}_{p(y; \vec{\eta})}[T_1(y)] \quad \text{and} \quad \frac{\partial A(\vec{\eta})}{\partial \eta_2} = \mathbb{E}_{p(y; \vec{\eta})}[T_2(y)]. \tag{52}$$

Thus, we can further simplify:

$$\begin{aligned}
 \frac{\partial}{\partial W_{dh}} \log(p(y_d; \vec{\eta}_d(\vec{s}, \Theta))) &= \sum_{l=1}^2 \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_l(w, \bar{V}_d) \Big|_{w=\bar{W}_d} \right) \\
 &\quad \times \left(T_l(y_d) - \mathbb{E}_{p(y; \vec{\eta})}[T_l(y)] \right) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d, \bar{V}_d)} \\
 &= \sum_{l=1}^2 \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_l(w, \bar{V}_d) \Big|_{w=\bar{W}_d} \right) \left(T_l(y_d) - \mathbb{E}_{p(y; \vec{\Phi}(\bar{W}_d, \bar{V}_d))}[T_l(y)] \right).
 \end{aligned} \tag{53}$$

Now, using Lemma 1 we obtain:

$$\begin{aligned}
 \frac{\partial}{\partial W_{dh}} \log(p(y_d; \vec{\eta}_d(\vec{s}, \Theta))) &= \sum_{l=1}^2 \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_l(w, V_{dh}) \Big|_{w=W_{dh}} \right) \\
 &\quad \times \left(T_l(y_d) - \mathbb{E}_{p(y; \vec{\Phi}(W_{dh}, V_{dh}))}[T_l(y)] \right).
 \end{aligned} \tag{54}$$

Note that the above equation depends on parameter \vec{s} of the hidden states only through the function $\mathcal{A}_{dh}(\vec{s}, \Theta)$. This is an important property of Lemma 1 that alleviates the complexity of the aforementioned equation and enables us to extract a set of concise update equations for dictionaries W and V . To see this, we derive (using Eqn. (26)):

$$\begin{aligned}
 \frac{\partial}{\partial W_{dh}} \mathcal{F}(q, \Theta) &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \left(\sum_{d'} \frac{\partial}{\partial W_{dh}} \log(p(y_{d'}^{(n)}; \vec{\eta}_{d'}(\vec{s}, \Theta))) \right) \\
 &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \left(\sum_{d'} \sum_{l=1}^2 \left(\mathcal{A}_{d'h}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_l(w, V_{d'h}) \Big|_{w=W_{d'h}} \right) \right. \\
 &\quad \times \left. \left(T_l(y_{d'}^{(n)}) - \mathbb{E}_{p(y; \vec{\Phi}(W_{d'h}, V_{d'h}))}[T_l(y)] \right) \delta_{dd'} \right) \\
 &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \sum_{l=1}^2 \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_l(w, V_{dh}) \Big|_{w=W_{dh}} \right) \\
 &\quad \times \left(T_l(y_d^{(n)}) - \mathbb{E}_{p(y; \vec{\Phi}(W_{dh}, V_{dh}))}[T_l(y)] \right) \\
 &= \sum_{l=1}^2 \left(\frac{\partial}{\partial w} \Phi_l(w, V_{dh}) \Big|_{w=W_{dh}} \right) \sum_n \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] \\
 &\quad \times \left(T_l(y_d^{(n)}) - \mathbb{E}_{p(y; \vec{\Phi}(W_{dh}, V_{dh}))}[T_l(y)] \right),
 \end{aligned} \tag{55}$$

where $\delta_{dd'}$ denotes the Kronecker delta, and using our mean value parametrization defined by $\vec{w} := \mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[\vec{T}(y)]$, we have:

$$\begin{aligned}
 \frac{\partial}{\partial W_{dh}} \mathcal{F}(q, \Theta) &= \left(\frac{\partial}{\partial w} \Phi_1(w, V_{dh}) \Big|_{w=W_{dh}} \right) \sum_n \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] \left(T_1(y_d^{(n)}) - W_{dh} \right) \\
 &\quad + \left(\frac{\partial}{\partial w} \Phi_2(w, V_{dh}) \Big|_{w=W_{dh}} \right) \sum_n \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] \left(T_2(y_d^{(n)}) - V_{dh} \right).
 \end{aligned} \tag{56}$$

Now, independently of the functions $\frac{\partial}{\partial w} \Phi_l(w, V_{dh})|_{w=W_{dh}}$ for $l = 1, 2$, the derivative of the ELBO w.r.t. W_{dh} is zero, i.e. $\frac{\partial \mathcal{F}}{\partial W_{dh}} = 0$, if it is the case that:

$$\sum_n \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] (T_1(y_d^{(n)}) - W_{dh}) = 0 \quad \text{and} \quad (57)$$

$$\sum_n \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] (T_2(y_d^{(n)}) - V_{dh}) = 0. \quad (58)$$

Rearranging terms yields (30) and (31) and completes the proof. The proof proceeds along the same lines for $\frac{\partial \mathcal{F}}{\partial V_{dh}}$ which results in the same set of equations. \blacksquare

A.2 Second Derivatives

Theorem 1 provides equations (30) and (31) for the matrices W and V . Fulfilling the equations implies that the derivatives of $\mathcal{F}(q, \Theta)$ w.r.t. W and V vanish. If we assume convergence of the corresponding fixed-point equations (38) and (39) that we use as M-steps, then W and V approach the values that solve (30) and (31) arbitrary closely. And in the numeric evaluations such convergence is observed. The M-step includes, for completeness, also the update for $\vec{\pi}$ in (40). If we further suppose for the M-step that W and V have fully converged (and that $\vec{\pi}$ has been updated) then all derivatives of $\mathcal{F}(q, \Theta)$ vanish, i.e., $\mathcal{F}(q, \Theta)$ has a fixed-point w.r.t. the model parameters $\Theta = (\vec{\pi}, W, V)$. But vanishing derivatives can also correspond to (local) minima or saddle points and not necessarily maxima. We can gather more information about the fixed-points by considering, e.g., the updates (38) for one element W_{dh} of the matrix W . If we now divert from the fixed-point by changing W_{dh} , we can ask if the objective $\mathcal{F}(q, \Theta)$ then increases or decreases. We do so by evaluating the second derivative at an (arbitrary) fixed-point of $\mathcal{F}(q, \Theta)$. We start with the first derivative:

$$\begin{aligned} \frac{\partial}{\partial W_{dh}} \mathcal{F}(q, \Theta) &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \frac{\partial}{\partial W_{dh}} \left(\sum_{d'} \log(p(y_{d'}^{(n)}; \vec{\eta}_{d'}(\vec{s}, \Theta))) \right) \\ &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \sum_{l=1}^2 \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_l(w, V_{dh})|_{w=W_{dh}} \right) \left(T_l(y_d^{(n)}) - \frac{\partial A(\vec{\eta})}{\partial \eta_l} \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} \\ &= \underbrace{\left(\frac{\partial}{\partial w} \Phi_1(w, V_{dh})|_{w=W_{dh}} \right)}_{:= G_1} \underbrace{\sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(T_1(y_d^{(n)}) - \frac{\partial A(\vec{\eta})}{\partial \eta_1} \right)}_{:= X_1} \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} \\ &\quad + \underbrace{\left(\frac{\partial}{\partial w} \Phi_2(w, V_{dh})|_{w=W_{dh}} \right)}_{:= G_2} \underbrace{\sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(T_2(y_d^{(n)}) - \frac{\partial A(\vec{\eta})}{\partial \eta_2} \right)}_{:= X_2} \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})}. \end{aligned} \quad (59)$$

Then, considering the definitions of the terms X_1 , X_2 , G_1 and G_2 above, the second derivative yields (at the fixed-point):

$$\begin{aligned} \frac{\partial^2}{(\partial W_{dh})^2} \mathcal{F}(q, \Theta) &= \frac{\partial G_1}{\partial W_{dh}} X_1 + G_1 \frac{\partial X_1}{\partial W_{dh}} + \frac{\partial G_2}{\partial W_{dh}} X_2 + G_2 \frac{\partial X_2}{\partial W_{dh}} \\ &= G_1 \frac{\partial X_1}{\partial W_{dh}} + G_2 \frac{\partial X_2}{\partial W_{dh}}, \end{aligned} \quad (60)$$

since $X_1 = X_2 = 0$ as W and V satisfy (30) and (31). Now, using the chain rule we get:

$$\begin{aligned}
 \frac{\partial X_1}{\partial W_{dh}} &= \left(\frac{\partial X_1}{\partial \eta_1} \frac{\partial \eta_1}{\partial W_{dh}} \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} + \left(\frac{\partial X_1}{\partial \eta_2} \frac{\partial \eta_2}{\partial W_{dh}} \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} \\
 &= - \left(\sum_{n, \vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \frac{\partial^2 A(\vec{\eta})}{\partial \eta_1 \partial \eta_1} \left(\frac{\partial}{\partial W_{dh}} \Phi_1(W_{dh}, V_{dh}) \right) \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} \\
 &\quad - \left(\sum_{n, \vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \frac{\partial^2 A(\vec{\eta})}{\partial \eta_2 \partial \eta_1} \left(\frac{\partial}{\partial W_{dh}} \Phi_2(W_{dh}, V_{dh}) \right) \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} \\
 &= - \sum_{n, \vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \left(\frac{\partial^2 A(\vec{\eta})}{\partial \eta_1 \partial \eta_1} G_1 + \frac{\partial^2 A(\vec{\eta})}{\partial \eta_2 \partial \eta_1} G_2 \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})}. \tag{61}
 \end{aligned}$$

Note that we here treat $\mathcal{A}_{dh}(\vec{s}, \Theta)$ as a constant since the second derivative of \bar{W}_d w.r.t. W_{dh} is equal to zero (recall Eqns. (27)-(28)). Similarly we can show:

$$\frac{\partial X_2}{\partial W_{dh}} = - \sum_{n, \vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \left(\frac{\partial^2 A(\vec{\eta})}{\partial \eta_1 \partial \eta_2} G_1 + \frac{\partial^2 A(\vec{\eta})}{\partial \eta_2 \partial \eta_2} G_2 \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})}. \tag{62}$$

Now, inserting (61) and (62) into (60), we get:

$$\begin{aligned}
 \frac{\partial^2}{(\partial W_{dh})^2} \mathcal{F}(q, \Theta) &= G_1 \frac{\partial X_1}{\partial W_{dh}} + G_2 \frac{\partial X_2}{\partial W_{dh}} \\
 &= G_1 \left(- \sum_{n, \vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \left(\frac{\partial^2 A(\vec{\eta})}{\partial \eta_1 \partial \eta_1} G_1 + \frac{\partial^2 A(\vec{\eta})}{\partial \eta_2 \partial \eta_1} G_2 \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} \right) \\
 &\quad + G_2 \left(- \sum_{n, \vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \left(\frac{\partial^2 A(\vec{\eta})}{\partial \eta_1 \partial \eta_2} G_1 + \frac{\partial^2 A(\vec{\eta})}{\partial \eta_2 \partial \eta_2} G_2 \right) \Big|_{\vec{\eta}=\vec{\Phi}(W_{dh}, V_{dh})} \right) \\
 &= - \sum_{n, \vec{s}} q^{(n)}(\vec{s}) \mathcal{A}_{dh}(\vec{s}, \Theta) \mathcal{G}^T(W_{dh}, V_{dh}) (\mathbf{H}_A) \mathcal{G}(W_{dh}, V_{dh}),
 \end{aligned}$$

where \mathbf{H}_A denotes the Hessian matrix of A (based on (5), it also corresponds to the covariance matrix between $T_1(y)$ and $T_2(y)$), and $\mathcal{G} := \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}$. The term $\mathcal{G}^T(W_{dh}, V_{dh})(\mathbf{H}_A)\mathcal{G}(W_{dh}, V_{dh})$ constructs a quadratic form, and the Hessian of the log-partition $A(\vec{\eta})$ is known to be positive semi-definite; and also positive definite for minimal representations (e.g., Wainwright and Jordan, 2008; Bickel and Doksum, 2015). As $q^{(n)}(\vec{s})$ and $\mathcal{A}_{dh}(\vec{s}, \Theta)$ are non-negative, we therefore obtain:

$$\frac{\partial^2}{(\partial W_{dh})^2} \mathcal{F}(q, \Theta) \leq 0. \tag{63}$$

The second derivative can, furthermore, be considered strictly smaller than zero under mild conditions. For instance, we can demand the condition that there exist indices n' and \vec{s}' such that $q^{(n')}(\vec{s}') \mathcal{A}_{dh}(\vec{s}', \Theta) \neq 0$, then it follows that $\mathbb{E}_{q^{(n')}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] \neq 0$. The condition can only be violated artificially with potentially few data points with specific properties, and the condition is satisfied for the types of data that are usually of interest (including all data considered in this work). If

the condition is combined with properties of $\vec{\Phi}$ ensuring non-zero $\mathcal{G}(W_{dh}, V_{dh})$ and with minimality (ensuring positive definite \mathbf{H}_A), then strictly smaller second derivatives are obtained. Consequently, the diversion of W_{dh} from the fixed-point result in a decrease of the objective, and (assuming convergence) updates (38) will increase $\mathcal{F}(q, \Theta)$ until a (local) maximum is reached.

Nonetheless, it is challenging to infer a more general behavior. First of all, the fixed-point updates (38) could in theory show intricate behavior such as convergence to an attractor for difference equations. Moreover, in practice all model parameter Θ are updated which would require the analysis of all second derivatives including all mixed second derivatives for all parameters (amounting to an exhaustive classification of all fixed-points). Further investigations would also require the combined optimization using E- as well as M-steps. Still, Eqn. (63) provides some information about the individual updates (38), and it applies for an arbitrary W_{dh} , of course, and for an arbitrary fixed-point. We also do note that the very same result is obtained for individual updates of V_{dh} using (39), i.e., the derivations will be analogous, with the same result of second derivatives being smaller than zero.

A.3 Parameter Update Equations – General Case

We derived Theorem 1 for the case of $L = 2$, i.e., for distributions of the exponential family with sufficient statistics of length two. This choice was for notational convenience only, and the proof of Theorem 1 suggests a straightforward generalization for arbitrary L . We already discussed the generalization of our formulations in Sec. 6 and stated Theorem 2 for the general case. Here we provide the proof of Theorem 2. First recall the details presented in Sec. 6. Then, similar to the $L = 2$ case, observe that the derivatives of $\mathcal{F}(q, \Theta)$ contain derivatives of $\bar{W}_d^{(l)}(\vec{s}, \Theta)$ for $l = 1, \dots, L$ w.r.t. the dictionary elements $W_{dh}^{(l)}$ for each d and h . For these derivatives the generalization of Eqns. (27)-(28) becomes:

$$\frac{\partial}{\partial W_{dh}^{(l)}} \bar{W}_d^{(l')}(\vec{s}, \Theta) = \begin{cases} \mathcal{A}_{dh}(\vec{s}, \Theta) & \text{if } l = l' \\ 0 & \text{otherwise} \end{cases} = \mathcal{A}_{dh}(\vec{s}, \Theta) \delta_{ll'}, \quad (64)$$

where $\mathcal{A}_{dh}(\vec{s}, \Theta)$ is given by (29). Lemma 1 can be also generalized by again considering the cases $h = h(d, \vec{s}, \Theta)$ and $h \neq h(d, \vec{s}, \Theta)$ separately, which then yields:

$$\mathcal{A}_{dh}(\vec{s}, \Theta) g(\bar{W}_d^{(1)}(\vec{s}, \Theta), \dots, \bar{W}_d^{(L)}(\vec{s}, \Theta)) = \mathcal{A}_{dh}(\vec{s}, \Theta) g(W_{dh}^{(1)}, \dots, W_{dh}^{(L)}). \quad (65)$$

Proof of Theorem 2 Consider a single parameter $W_{dh}^{(l)}$ for an arbitrary $l = 1, \dots, L$ and abbreviate $\bar{W}_d^{(l)}(\vec{s}, \Theta)$ by $\bar{W}_d^{(l)}$. Then, using the chain rule and Eqn. (43), we get:

$$\begin{aligned} \frac{\partial}{\partial W_{dh}^{(l)}} \log(p(y_d; \vec{\eta}_d(\vec{s}, \Theta))) &= \frac{\partial}{\partial W_{dh}^{(l)}} \log(p(y_d; \vec{\Phi}(\bar{W}_d^{(1)}, \dots, \bar{W}_d^{(L)}))) \\ &= \sum_{l'=1}^L \left(\frac{\partial}{\partial W_{dh}^{(l)}} \Phi_{l'}(\bar{W}_d^{(1)}, \dots, \bar{W}_d^{(L)}) \right) \left(\frac{\partial}{\partial \eta_{l'}} \log(p(y_d; \vec{\eta})) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d^{(1)}, \dots, \bar{W}_d^{(L)})} \right) \\ &= \sum_{l'=1}^L \left\{ \sum_{l''=1}^L \left(\frac{\partial}{\partial W_{dh}^{(l)}} \bar{W}_d^{(l'')} \right) \left(\frac{\partial}{\partial w} \Phi_{l'}(\bar{W}_d^{(1)}, \dots, w, \dots, \bar{W}_d^{(L)}) \Big|_{w=\bar{W}_d^{(l'')}} \right) \right\} \\ &\quad \times \left(\frac{\partial}{\partial \eta_{l'}} \log(p(y_d; \vec{\eta})) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d^{(1)}, \dots, \bar{W}_d^{(L)})} \right). \end{aligned} \quad (66)$$

Now, using (64) and also (5) that states for any regular distribution of the exponential family $\frac{\partial A(\vec{\eta})}{\partial \eta_{l'}} = \mathbb{E}_{p(y; \vec{\eta})}[T_{l'}(y)]$, we obtain:

$$\begin{aligned}
 \frac{\partial}{\partial W_{dh}^{(l)}} \log(p(y_d; \vec{\eta}_d(\vec{s}, \Theta))) &= \sum_{l'=1}^L \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_{l'}(\bar{W}_d^{(1)}, \dots, w, \dots, \bar{W}_d^{(L)}) \Big|_{w=\bar{W}_d^{(l)}} \right) \\
 &\quad \times \left(T_{l'}(y_d) - \frac{\partial A(\vec{\eta})}{\partial \eta_{l'}} \right) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d^{(1)}, \dots, \bar{W}_d^{(L)})} \\
 &= \sum_{l'=1}^L \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_{l'}(\bar{W}_d^{(1)}, \dots, w, \dots, \bar{W}_d^{(L)}) \Big|_{w=\bar{W}_d^{(l)}} \right) \\
 &\quad \times \left(T_{l'}(y_d) - \mathbb{E}_{p(y; \vec{\eta})}[T_{l'}(y)] \right) \Big|_{\vec{\eta}=\vec{\Phi}(\bar{W}_d^{(1)}, \dots, \bar{W}_d^{(L)})} \\
 &= \sum_{l'=1}^L \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_{l'}(\bar{W}_d^{(1)}, \dots, w, \dots, \bar{W}_d^{(L)}) \Big|_{w=\bar{W}_d^{(l)}} \right) \\
 &\quad \times \left(T_{l'}(y_d) - \mathbb{E}_{p(y; \vec{\Phi}(\bar{W}_d^{(1)}, \dots, \bar{W}_d^{(L)}))}[T_{l'}(y)] \right).
 \end{aligned}$$

Further, using (65) simplifies the above equation as follows:

$$\begin{aligned}
 \frac{\partial}{\partial W_{dh}^{(l)}} \log(p(y_d; \vec{\eta}_d(\vec{s}, \Theta))) &= \sum_{l'=1}^L \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_{l'}(W_{dh}^{(1)}, \dots, w, \dots, W_{dh}^{(L)}) \Big|_{w=W_{dh}^{(l)}} \right) \\
 &\quad \times \left(T_{l'}(y_d) - \mathbb{E}_{p(y; \vec{\Phi}(W_{dh}^{(1)}, \dots, W_{dh}^{(L)}))}[T_{l'}(y)] \right).
 \end{aligned}$$

Note that the equation now depends on the hidden state \vec{s} only through the function $\mathcal{A}_{dh}(\vec{s}, \Theta)$. Then, using Eqn. (26), it can be stated that:

$$\begin{aligned}
 \frac{\partial}{\partial W_{dh}^{(l)}} \mathcal{F}(q, \Theta) &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \left(\sum_{d'} \frac{\partial}{\partial W_{dh}^{(l)}} \log(p(y_{d'}^{(n)}; \vec{\eta}_{d'}(\vec{s}, \Theta))) \right) \\
 &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \sum_{l'=1}^L \left(\mathcal{A}_{dh}(\vec{s}, \Theta) \right) \left(\frac{\partial}{\partial w} \Phi_{l'}(W_{dh}^{(1)}, \dots, w, \dots, W_{dh}^{(L)}) \Big|_{w=W_{dh}^{(l)}} \right) \\
 &\quad \times \left(T_{l'}(y_d^{(n)}) - \mathbb{E}_{p(y; \vec{\Phi}(W_{dh}^{(1)}, \dots, W_{dh}^{(L)}))}[T_{l'}(y)] \right) \\
 &= \sum_{l'=1}^L \left(\frac{\partial}{\partial w} \Phi_{l'}(W_{dh}^{(1)}, \dots, w, \dots, W_{dh}^{(L)}) \Big|_{w=W_{dh}^{(l)}} \right) \\
 &\quad \times \sum_n \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] \left(T_{l'}(y_d^{(n)}) - W_{dh}^{(l')} \right),
 \end{aligned}$$

where for the last step we exploited the mean value parametrization defined by $\vec{w} := \mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[\vec{T}(y)]$. Therefore, independently of the functions $\frac{\partial}{\partial w} \Phi_{l'}(W_{dh}^{(1)}, \dots, w, \dots, W_{dh}^{(L)}) \Big|_{w=W_{dh}^{(l)}}$ for each l' , we

have $\frac{\partial \mathcal{F}}{\partial W_{dh}^{(l)}} = 0$ if for all $l = 1, \dots, L$ we have:

$$\sum_n \mathbb{E}_{q^{(n)}} [\mathcal{A}_{dh}(\vec{s}, \Theta)] \left(T_l(y_d^{(n)}) - W_{dh}^{(l)} \right) = 0, \quad (67)$$

which yields Eqn. (46) and completes the proof. \blacksquare

The general case does, of course, also include the $L = 1$ case and consequently Bernoulli, Exponential or the Poisson distribution. Moreover, similar to the $L = 2$ case, we can gather further information about points where $W_{dh}^{(l)}$ satisfies Eqn. (46). Considering the second derivatives again as in Sec. A.2, we can likewise show that fixed-point solutions for a single $W_{dh}^{(l)}$ and for fixed d and h correspond to (local) maxima if the Hessian matrix $\mathbf{H}_{\mathcal{F}}$ is positive definite.

Appendix B. Additional Details on Parameter Update Equations

B.1 Parametrization of the Gaussian-MCA

For Gaussian-MCA (G-MCA) in Example 2, the generative model was shown to be given by:

$$p(\vec{s} | \Theta) = \prod_{h=1}^H (\pi_h^{s_h} (1 - \pi_h)^{1-s_h}), \quad (68)$$

$$p(\vec{y} | \vec{s}, \Theta) = \prod_{d=1}^D \mathcal{N}(y_d; \bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta) - \bar{W}_d^2(\vec{s}, \Theta)), \quad (69)$$

where we can use the update equations of Theorem 1 for W and V . To obtain some intuition, we can also define the function:

$$\bar{\sigma}_d^2(\vec{s}, \Theta) = \bar{V}_d(\vec{s}, \Theta) - \bar{W}_d^2(\vec{s}, \Theta), \quad \forall d.$$

Because of the definition of $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$ in (22) we get:

$$\bar{\sigma}_d^2(\vec{s}, \Theta) = \sigma_{dh(d, \vec{s}, \Theta)}^2 \text{ where } h(d, \vec{s}, \Theta) = \operatorname{argmax}_h \{W_{dh} s_h\} \text{ and } \sigma_{dh}^2 = V_{dh} - W_{dh}^2,$$

such that σ^2 is a matrix with $D \times H$ entries. The considered generative model then becomes:

$$p(\vec{s} | \Theta) = \prod_{h=1}^H (\pi_h^{s_h} (1 - \pi_h)^{1-s_h}), \quad (70)$$

$$p(\vec{y} | \vec{s}, \Theta) = \prod_{d=1}^D \mathcal{N}(y_d; \bar{W}_d(\vec{s}, \Theta), \bar{\sigma}_d^2(\vec{s}, \Theta)). \quad (71)$$

The latents thus change the mean via the matrix W and the variance via the matrices V and W . We can now use the update rules for V_{dh} and W_{dh} and compute the elements σ_{dh}^2 as a result. Alternatively, we can also combine the update rules for V_{dh} and W_{dh} to directly obtain an update rule for σ_{dh}^2 :

$$\begin{aligned}
 (\sigma^2_{dh})^{\text{new}} &= V_{dh}^{\text{new}} - (W_{dh}^{\text{new}})^2 = V_{dh}^{\text{new}} - 2W_{dh}^{\text{new}}W_{dh}^{\text{new}} + (W_{dh}^{\text{new}})^2 \\
 &= \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] (y_d^{(n)})^2}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]} - 2 \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] y_d^{(n)}}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]} W_{dh}^{\text{new}} + (W_{dh}^{\text{new}})^2 \\
 &= \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] ((y_d^{(n)})^2 - 2y_d^{(n)}W_{dh}^{\text{new}} + (W_{dh}^{\text{new}})^2)}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]} \\
 &= \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] (y_d^{(n)} - W_{dh}^{\text{new}})^2}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]}. \tag{72}
 \end{aligned}$$

This form of update is more familiar as it evaluates the square deviation from the mean given by W_{dh}^{new} . Also note that we first have to compute W_{dh}^{new} before updating σ_{dh}^2 , which is in analogy, e.g., to standard Gaussian mixtures for variance (or covariance) updates. For the G-MCA application of Sec. 7.2.1, we showed W_{dh} and σ_{dh} (standard deviations) for whitened natural image patches.

B.2 Parametrization of the Gamma-MCA

Let Gamma be the noise distribution in (8)-(10); i.e., consider the following generative model:

$$p(\vec{s} | \Theta) = \prod_{h=1}^H (\pi_h^{s_h} (1 - \pi_h)^{1-s_h}), \tag{73}$$

$$p(\vec{y} | \vec{s}, \Theta) = \prod_{d=1}^D \text{Gamma}(y_d; \vec{\eta}_d(\vec{s}, \Theta)), \quad \text{for } y_d \in (0, \infty). \tag{74}$$

Given the shape and rate parameters $\alpha, \beta > 0$ of the Gamma distribution, the natural parameters and sufficient statistics are given by:

$$\vec{\eta} = (-\beta, \alpha - 1)^T, \quad \vec{T}(y) = (y, \log(y))^T.$$

Now, based on the mean value parametrization in (6) and also (7), we get:

$$w_1 = \mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[y] \text{ and } w_2 = \mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[\log(y)],$$

where $\mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[y]$ represents the mean of observables and is equal to $-\frac{\eta_2 + 1}{\eta_1}$. Also we know that

$$\mathbb{E}_{p(y; \vec{\Phi}(\vec{w}))}[\log(y)] = \psi(\eta_2 + 1) - \log(-\eta_1),$$

with the Digamma function $\psi(\cdot)$ defined by:

$$\psi(x) = \frac{d}{dx} \log(\Gamma(x)) = \log(x) - \frac{1}{2x} - \frac{1}{12x^2} + \dots \tag{75}$$

Thus, the relation between the mean value parameters and the natural parameters can be expressed by:

$$\vec{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} -\frac{\eta_2 + 1}{\eta_1} \\ \psi(\eta_2 + 1) - \log(-\eta_1) \end{pmatrix}.$$

In order to compute the function $\vec{\Phi}$, we further substitute $\eta_1 = -\frac{\eta_2+1}{w_1}$ (the first equation above) into $w_2 = \psi(\eta_2 + 1) - \log(-\eta_1)$ (the second equation above) that yields:

$$w_2 \approx \log(\eta_2 + 1) - \frac{1}{2(\eta_2 + 1)} - \log\left(\frac{\eta_2 + 1}{w_1}\right) = -\frac{1}{2(\eta_2 + 1)} + \log(w_1), \quad (76)$$

where we approximated the Digamma function with its first two terms. Then, after some simplifications, it can be easily seen that the above system of equations results in:

$$\eta_1 \approx \frac{-1}{2w_1(\log(w_1) - w_2)} \quad \text{and} \quad \eta_2 \approx \frac{1}{2(\log(w_1) - w_2)} - 1. \quad (77)$$

Therefore, using the natural parameters (77) and the definition of the link function in (11), we get:

$$\vec{\eta}_d(\vec{s}, \Theta) = \vec{\Phi}(\bar{W}_d(\vec{s}, \Theta), \bar{V}_d(\vec{s}, \Theta)) \approx \begin{pmatrix} \frac{-1}{2\bar{W}_d(\vec{s}, \Theta)(\log(\bar{W}_d(\vec{s}, \Theta)) - \bar{V}_d(\vec{s}, \Theta))} \\ \frac{1}{2(\log(\bar{W}_d(\vec{s}, \Theta)) - \bar{V}_d(\vec{s}, \Theta))} - 1 \end{pmatrix}.$$

The function $\vec{\Phi}$ defines the link between natural and mean value parameters which will be used in the EM algorithm. Observe that in each EM iteration, we require to compute $\vec{\eta}_d(\vec{s}, \Theta)$ (e.g., to compute the posteriors) given the mean value parameters $\bar{W}_d(\vec{s}, \Theta)$ and $\bar{V}_d(\vec{s}, \Theta)$. These mean values can be further used to estimate the variance of the Gamma distribution such as $\bar{\sigma}_d^2(\vec{s}, \Theta)$. Alternatively, we can use the variance formula of the Gamma distribution (given by $\sigma^2 = \frac{\eta_2+1}{\eta_1^2}$) and let:

$$\bar{\sigma}_d^2(\vec{s}, \Theta) \approx 2\bar{W}_d^2(\vec{s}, \Theta)(\log(\bar{W}_d(\vec{s}, \Theta)) - \bar{V}_d(\vec{s}, \Theta)), \quad \forall d, \quad (78)$$

where (similar to the Gaussian case)

$$\bar{\sigma}_d^2(\vec{s}, \Theta) = \sigma_{dh(d, \vec{s}, \Theta)}^2 \quad \text{and} \quad h(d, \vec{s}, \Theta) = \operatorname{argmax}_h \{W_{dh}s_h\}. \quad (79)$$

Likewise, the variance elements can be directly updated in each M-step by:

$$\begin{aligned} (\sigma_{dh}^2)^{\text{new}} &\approx 2(W_{dh}^{\text{new}})^2 \left(\log(W_{dh}^{\text{new}}) - V_{dh}^{\text{new}} \right) \\ &= 2(W_{dh}^{\text{new}})^2 \left(\frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] \log(W_{dh}^{\text{new}})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]} - \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] \log(y_d^{(n)})}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]} \right) \\ &= 2 \frac{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)] (W_{dh}^{\text{new}})^2 (\log(W_{dh}^{\text{new}}) - \log(y_d^{(n)}))}{\sum_{n=1}^N \mathbb{E}_{q^{(n)}}[\mathcal{A}_{dh}(\vec{s}, \Theta)]}. \end{aligned} \quad (80)$$

The above equation has been used for updating the variance of the Gamma-MCA model in our experiments in Sec. 7.2.2. Despite the approximation of the Digamma function, we observed that the aforementioned equation provides, in practice, a good estimation of the variance parameter. Other approximations of the Digamma function can be further used to improve the performance of the method. For instance, we also used the first three and also four terms of the Digamma function in (75) for our experiments and consistently observed a better approximation of the variance.

B.3 The Evidence Lower Bound (ELBO)

For completeness, we here show the ELBO / free energy (26) can be derived as a lower bound of the log-likelihood function. For this, we use the view as presented, e.g., by Neal and Hinton (1998); Bishop (2006) and introduce variational distributions $q^{(n)}(\vec{s})$ defined over the latent variables (as discussed before, these distributions can be full posteriors $p(\vec{s} | \vec{y}^{(n)}, \Theta)$, or denote approximations in the case of intractable posteriors). Then, for an EF-MCA data model (8)-(10) we have:

$$\begin{aligned}\mathcal{L}(\Theta) &= \sum_n \log \left(\sum_{\vec{s}} q^{(n)}(\vec{s}) \frac{p(\vec{y}^{(n)}, \vec{s} | \Theta)}{q^{(n)}(\vec{s})} \right) \geq \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \log \left(\frac{p(\vec{y}^{(n)}, \vec{s} | \Theta)}{q^{(n)}(\vec{s})} \right) \\ &= \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \log (p(\vec{y}^{(n)}, \vec{s} | \Theta)) - \sum_n \sum_{\vec{s}} q^{(n)}(\vec{s}) \log (q^{(n)}(\vec{s})) = \mathcal{F}(q, \Theta),\end{aligned}\quad (81)$$

where we used Jensen's inequality in the first line (see, e.g., Bishop, 2006, for the details). The last equation above then represents the free energy function $\mathcal{F}(q, \Theta)$ (or the ELBO) given by (26).

Appendix C. Additional Details on Numerical Experiments

C.1 The Bars Test

For the bars tests presented in Sec. 7.1, we set $\pi_h^{\text{gen}} = 0.2$ for $h = 1, \dots, H$ (i.e., two active bars on average per data point) and generated $N = 1000$ i.i.d. data points according to each of the considered EF-MCA models. The ground-truth parameter W was set to a value of 10 for the bars and 1 for the non-bar pixels in the case of E-MCA, and 0.99 for the bars and 0.01 for the non-bars in the case of B-MCA. Next, we trained E- and B-MCA models on their corresponding datasets and updated their parameters using (38) and (40) for 50 full EM iterations. We initialized W by computing the mean of the data and adding a small amount of Gaussian noise; also the priors π_h were initialized at 0.3 for each h . In each EM iteration, the fixed-point M-step equation was iterated only once, as we observed comparable results compared to running it multiple times for both E- and B-MCA models. Also for P-MCA, we used $H = 6$, $D = 9$ and set other parameters to be similar to the E-MCA model.

Relation to Previous MCA Approaches. In previous studies (e.g., Lücke and Sahani, 2008; Lücke et al., 2009; Lücke and Eggert, 2010; Bornschein et al., 2013), the maximum function has been commonly approximated using a smooth function for the sake of numerical stabilities. That is, for the case of Poisson or Gaussian distributions (where $M(\Theta) = W$), a smooth function \bar{W}_d^ρ is used to approximate the function \bar{W}_d as ρ approaches infinity:

$$\bar{W}_d^\rho(\vec{s}, \Theta) := \left(\sum_{h=1}^H (W_{dh} s_h)^\rho \right)^{\frac{1}{\rho}} \quad \text{where} \quad \lim_{\rho \rightarrow \infty} \bar{W}_d^\rho(\vec{s}, \Theta) = \bar{W}_d(\vec{s}, \Theta).$$

Parameter ρ does in fact control the level of non-linearity and it can be varied during the learning (as it has been done by Lücke et al., 2009), or it can be set to a large fixed value (similar to the procedure considered by Bornschein et al., 2013, who used, e.g., $\rho = 21$). Consequently, $\mathcal{A}_{dh}(\vec{s}, \Theta)$ can be rewritten by (see, e.g., Lücke and Sahani, 2008, for the details):

$$\mathcal{A}_{dh}(\vec{s}, \Theta) := \lim_{\rho \rightarrow \infty} \left(\frac{\partial}{\partial W_{dh}} \bar{W}_d^\rho(\vec{s}, \Theta) \right) = \lim_{\rho \rightarrow \infty} \frac{s_h (W_{dh})^\rho}{\sum_h s_h (W_{dh})^\rho}.$$

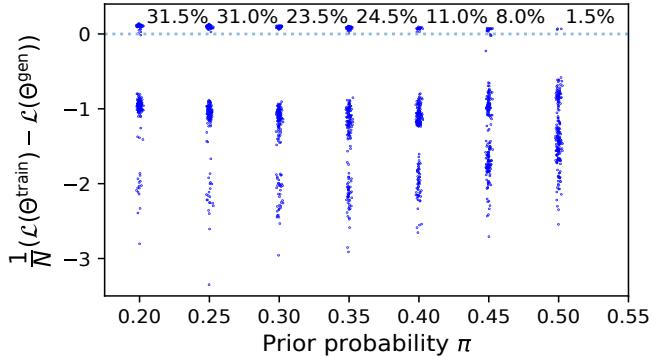


Figure 9: Differences of log-likelihoods between trained B-MCA models and ground-truth values for bars test datasets. Average active number of bars (πH) is set by different prior probabilities $0.2 \leq \pi \leq 0.5$ in step size 0.05. Individual runs are shown with small offsets along the π -axis for a better visualization. Percentages refer to the fraction of runs that converged to ground-truth values (in such cases, the learned log-likelihoods are usually higher than the ground-truth due to overfitting).

We here observed that such an approximation is beneficial to avoid numerical instabilities which can be used, e.g., along other approaches for avoiding local optima.

Avoiding Local Optima. In our bars test experiments, we executed the algorithms multiple times while holding the hyper-parameters fixed and using different realizations of the initial model parameters in each run. We observed that the best solutions recovered all bars with high accuracy. Nonetheless, as mentioned before, slight overfitting effects occurred in some cases that could be diminished by increasing the number of data points. We also frequently observed that the algorithms could not recover the ground-truth parameters. This can be seen as the effect of local optima which is different for each of the EF-MCA models. Local optima effects showed, for instance, to be more severe for Gamma-MCA compared to others. While annealing methods have frequently been applied as a measure against local optima effects (e.g., Ueda and Nakano, 1998; Lücke and Sahani, 2008), we here investigated an alternative approach in order to prevent the algorithm from converging to sub-optimal solutions. In detail, we first employed the Gaussian-MCA model as it is shown to be more robust against local optima and performed 50 EM iterations. We then used the learned Θ parameters (of the Gaussian-MCA) to initialize the parameters of the desired EF-MCA model. We observed that, e.g. for Gamma-MCA, such a method helps alleviate the effect of local optima.

Reliability For Binary Data. In addition to the local optima effect, the inherent complexity of the model can also affect its performance. That is, for instance, increasing the sparseness (i.e., $\sum_h \pi_h$ which in this case is equal to πH as we assumed $\pi_h = \pi$ for $h = 1, \dots, H$) will decrease the reliability of the model. In fact, one measure that has been of interest is how robust an algorithm is w.r.t. the average number of active bars, and how often it reaches the global vs. any local optimum. The probability of recovering all bars has been commonly termed *reliability* of the algorithm (Spratling, 2006). In the following, we investigate the reliability of the Bernoulli-MCA (B-MCA) model for different levels of sparseness. For our purposes, we notably did not optimize learning to

improve reliability (e.g., by introducing annealing procedures; Lücke and Eggert, 2010) but used the canonical form of B-MCA with exact E-steps (i.e., full posteriors) and updated W and $\vec{\pi}$ using Eqns. (38) and (40). We generated artificial datasets according to the B-MCA model similar to the procedure presented above with $W_{dh} \in \{0.01, 0.99\}$ and varied the value of π that determines the average number of active bars per data point. For each value of π , we generated 200 different bars datasets each with $N = 1000$ data points and then fitted a B-MCA model. We computed 50 full EM iterations with the same initialization as the experiments above and measured the reliability in terms of the percentage of all trained B-MCA models that achieved a higher log-likelihood value than the ground-truth. It should be noted that we commonly observed the overfitting effect as we used a finite sample size here and thus, the best runs were observed to have log-likelihood values that are slightly higher than the ground-truth log-likelihood. Results are presented in Fig. 9.

As it can be observed, the best runs reach higher log-likelihood values than the ground-truth. These runs do recover all bar patterns and the prior parameters. Even for values $\pi = 0.5$, i.e., for five out of ten bars per data point on average, one to two out of 100 runs do extract all bars. Thus, the best runs in terms of higher log-likelihood values can be automatically determined without knowledge of the ground-truth such that the proposed algorithm (B-MCA here) yields a reliable approach for extracting all bars for binary data (compare Spratling, 2006; Frolov et al., 2015).

C.2 Feature Extraction – Natural Image Patches

To initialize the model parameters $\Theta = \{\vec{\pi}, W, \sigma^2\}$ of the Gaussian-MCA model described in Sec. 7.2.1, we followed the procedure used in Mousavi et al. (2020): The π_h^{init} were randomly uniformly drawn from the interval $[0.1, 0.5]$ for $h = 1, \dots, H$. The generative fields (GFs) for component variances were set equal to the average variance per pixel, i.e., $(\sigma_{dh}^2)^{\text{init}} = \sigma_y^2 \forall d, h$ with $\sigma_y^2 = \frac{1}{DN} \sum_{d=1}^D \sum_{n=1}^N (y_d^{(n)} - \bar{y}_d)^2$ and $\bar{y}_d = \frac{1}{N} \sum_{n=1}^N y_d^{(n)}$. The initial values for the GFs for component means were obtained by clustering the dataset using a variational GMM algorithm (we used the var-GMM-S algorithm implemented in (Hirschberger, 2019) together with the default hyper-parameters; also see Hirschberger et al. (2022); Forster and Lücke (2018)), and then setting $\tilde{W}_h^{\text{init}}$ equal to the learned H cluster centers. In order to ensure positive values for $\bar{\sigma}$ (see, e.g., Fig. 5), we used the background model discussed in Mousavi et al. (2021) and enforced one hidden unit $s_1 = \text{const} = 1$ during learning. We performed 500 EM iterations, initially updating only W and $\vec{\pi}$, while σ^2 remained fixed at its initial value. After approximately half of the iterations, i.e., when we could no longer observe significant changes of the lower bound, we fixed W and $\vec{\pi}$ and only updated σ^2 for the remaining EM steps (we found this procedure to help avoiding oscillations of the lower bound which we observed when updating all parameters jointly from the beginning). Figure 10 depicts the $H = 1000$ learned GFs as well as the learned priors (the GFs for the background unit correspond to the first depicted dictionary element, the respective prior equals one).

Using compute resources of the HPC cluster CARL of the University of Oldenburg (which is equipped with Intel Xeon E5-2650 v4 12C, E5-2667 v4 8C and E7-8891 v4 10c CPUs), we observed an average runtime of approximately 10 minutes per EM step for our implementation when being executed distributedly on 556 CPU cores.

C.3 Noise Type Estimation

The problem of noise type estimation has been addressed in a number of contributions. Teymurazyan et al. (2013), for instance, studied the problem of distinguishing the type of noise in Positron

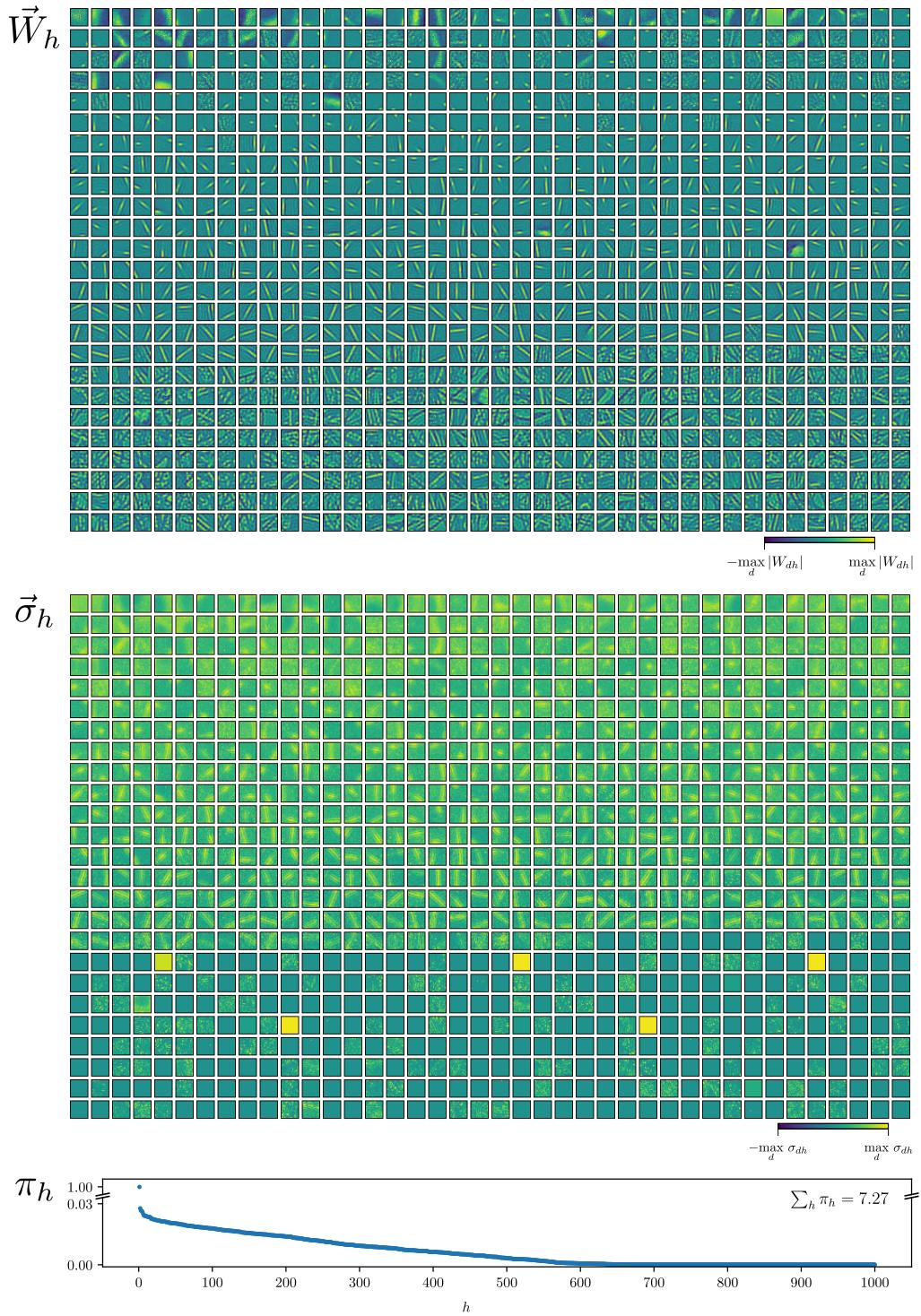


Figure 10: Complete dictionaries with $H = 1,000$ component means and component variances learned from natural image patches using Gaussian-MCA (for improved visibility, we illustrate standard deviations instead of variances; compare Sec. 7.2.1). The generative fields are ordered according to their activations, starting with the fields corresponding to the most active hidden units.

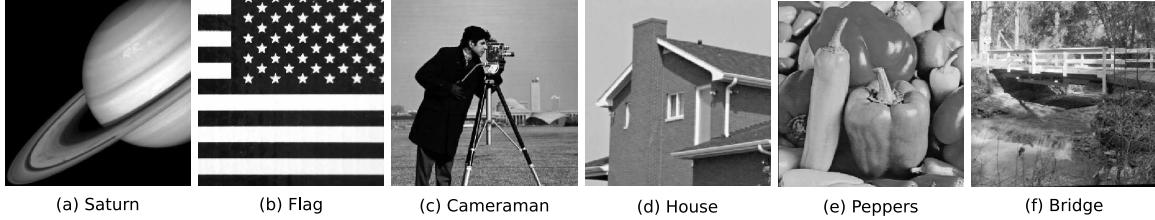


Figure 11: Original images used for denoising benchmarks.

Emission Tomography (PET) data (Teymurazyan et al., 2013). In PET, radioactive fluids are injected into humans or animals in order to obtain images for diagnostics or scientific studies. As a result, knowing the type of noise is crucial for different image processing routines (Teymurazyan et al., 2013; Mou et al., 2018). Other related work focused on machine learning automation (e.g., Valera and Ghahramani, 2017; Vergari et al., 2019). In Valera and Ghahramani (2017), a Bayesian approach is presented to infer the type of data in categories such as categorical, ordinal, count, real-valued, positive real-valued or interval (also see Molina et al., 2018). As discussed in Sec. 4, such approaches are different from our work here as they exploit (among other differences) a mixture of likelihood functions. We, on the other hand, trained each of the EF-MCA data models here and directly used the obtained ELBOs to infer the noise type of a given dataset.

For the experiments in Sec. 7.2.2, we used $D = 144$ and $H = 100$ (for images) or $H = 512$ (for acoustic data), and trained the EF-MCA data models for 100 variational EM iterations. The three examples in Tab. 3 that we selected from the CHiME dataset (Foster et al., 2015) were: CR-lounge-200110-1711.s0-chunk46.48kHz.wav, CR-lounge-200110-1601.s0-chunk25.48kHz.wav and CR-lounge-200110-1601.s0-chunk7.48kHz.wav. Amplitude spectrograms were computed as follows: Time domain signals were resampled to 22.050 Hz and cut into 2 seconds long mono segments. We then computed the STFT using a 2048-point FFT, 512 samples frame shift and Hann windowing. This resulted in spectrograms with 1025 frequency channels and 87 time steps. The amplitude spectrograms were then cut into patches of size $D = 12 \times 12$ resulting in $N = 77,064$ data points. After learning, we observed very small values of components σ_{dh} (between 0.1 and 0.2) for both Gaussian- and Gamma-MCA models which resulted in very high, positive free energies in Tab. 3.

C.4 Denoising

We used images with gray scales in the interval $[0, 255]$ (see the original images in Fig. 11) and, rescaled the pixel amplitudes by dividing by the maximum amplitude and then multiplying by the desired peak value. Similar to Drefs et al. (2022a), we further cut the images into smaller patches by moving a sliding window over the noisy versions. We then employed the EF-MCA algorithms with $D = 20 \times 20$ (similar to Giryes and Elad, 2014) and different H values (for Poisson noise), and $D = 12 \times 12$ and $H = 512$ (for Exponential noise). Each algorithm was trained for 100 variational EM iterations without any annealing or further pre- or post-processing. For both E- and P-MCA, we manually set the minimum value of W to 10^{-2} after each M-step in order to avoid zero values.

Tab. 4 lists the PSNR values for VST+BM3D, I+VST+BM3D and P⁴IP from Azzari and Foi (2016), BM3D, NLSPCA and SPDA from Giryes and Elad (2014), DenoiseNet from Remez et al. (2017) and DCEA-UC from Tolooshams et al. (2020). Also, Fig. 12 presents the reconstructed images using different generative models exploited for Poisson denoising.

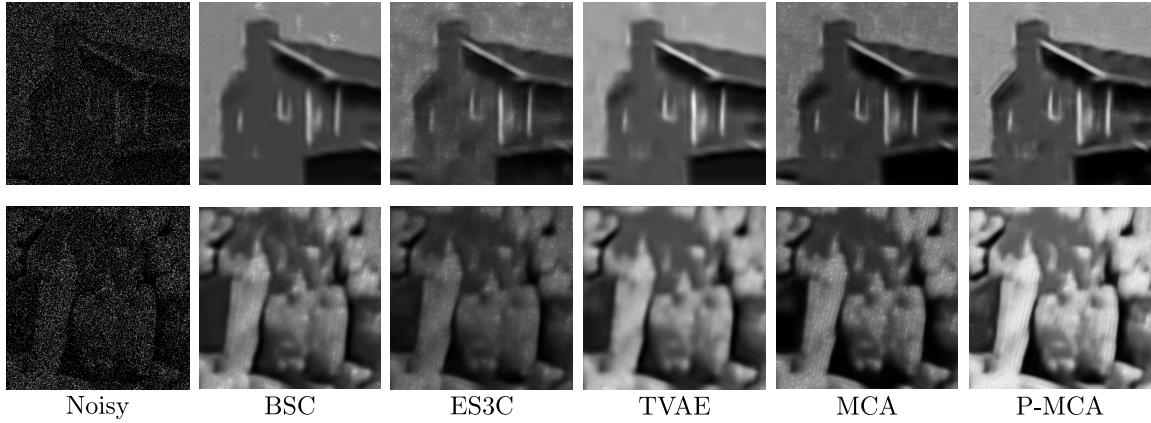


Figure 12: Reconstructed images obtained by different LVMs for Poisson denoising at peak value 1. The corresponding PSNR and ELBO values are presented in Fig. 7.

References

- M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.
- F. J. Anscombe. The transformation of Poisson, Binomial and Negative-Binomial data. *Biometrika*, 35(3/4):246–254, 1948.
- L. Azzari and A. Foi. Variance stabilization for noisy+estimate combination in iterative Poisson denoising. *IEEE Signal Processing Letters*, 23(8):1086–1090, 2016.
- D. Baby and H. Bourlard. Speech dereverberation using variational autoencoders. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5784–5788, 2021.
- D. J. Bartholomew. *Latent Variable Models and Factors Analysis*. Oxford University Press, Inc., 1987.
- M. Basbug and B. Engelhardt. Hierarchical compound Poisson factorization. In *International Conference on Machine Learning*, pages 1795–1803. PMLR, 2016.
- P. Berkes, B. White, and J. Fiser. No evidence for active sparsification in the visual cortex. In *Advances in Neural Information Processing Systems*, volume 22, pages 108–116, 2009.
- A. Berliner, G. Rotman, Y. Adi, R. Reichart, and T. Hazan. Learning discrete structured variational auto-encoder using natural evolution strategies. In *International Conference on Learning Representations*, 2022.
- P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Chapman and Hall/CRC, 2015.
- C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

- D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.
- S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7327–7347, 2022.
- J. Bornschein, M. Henniges, and J. Lücke. Are V1 simple cells optimized for visual occlusions? A comparative study. *PLOS Computational Biology*, 9(6):e1003062, 2013.
- H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising: Can plain neural networks compete with BM3D? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2392–2399, 2012.
- S. Chaudhury and H. Roy. Can fully convolutional networks perform well for general image restoration problems? In *Fifteenth IAPR International Conference on Machine Vision Applications*, pages 254–257. IEEE, 2017.
- Y. Chen and T. Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1256–1272, 2017.
- J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.
- M. Collins, S. Dasgupta, and R. E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, pages 617–624, 2002.
- P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.
- D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.
- K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.
- B. Dai, S. Ding, and G. Wahba. Multivariate Bernoulli distribution. *Bernoulli*, 19(4):1465–1483, 2013.
- Z. Dai and J. Lücke. Autonomous document cleaning – a generative approach to reconstruct strongly corrupted scanned texts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1950–1962, 2014.
- O. DeGuchy, F. Santiago, M. Banuelos, and R. F. Marcia. Deep neural networks for low-resolution photon-limited imaging. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3247–3251. IEEE, 2019.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–38, 1977.

- J. Drefs, E. Guiraud, and J. Lücke. Evolutionary variational optimization of generative models. *Journal of Machine Learning Research*, 23(21):1–51, 2022a.
- J. Drefs, E. Guiraud, F. Panagiotou, and J. Lücke. Direct evolutionary optimization of variational autoencoders with binary latents. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 357–372. Springer, 2022b.
- B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, pages 1189–1242, 1975.
- B. Efron. *Exponential Families in Theory and Practice*. Cambridge University Press, 2023.
- B. Everitt. *An Introduction to Latent Variable Models*. Chapman and Hall, 1984.
- G. Exarchakis and J. Lücke. Discrete sparse coding. *Neural Computation*, 29:2979–3013, 2017.
- G. Exarchakis, O. Oubari, and G. Lenz. A sampling-based approach for efficient clustering in large datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12403–12412, 2022.
- S. Ferrari and F. Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815, 2004.
- A. Fischer and C. Igel. Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.
- M. Fisz. The limiting distribution of a function of two independent random variables and its statistical application. In *Colloquium Mathematicum*, volume 3, pages 138–146, 1955.
- P. Földiák. Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics*, 64:165–170, 1990.
- D. Forster and J. Lücke. Can clustering scale sublinearly with its clusters? A variational EM acceleration of GMMs and k-means. In *International Conference on Artificial Intelligence and Statistics*, pages 124–132. PMLR, 2018.
- P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumley. Chime-home: A dataset for sound source recognition in a domestic environment. In *Proceedings of the 11th Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–5, 2015.
- A. A. Frolov, D. Húsek, and P. Y. Polyakov. Two expectation-maximization algorithms for boolean factor analysis. *Neurocomputing*, 130(0):83 – 97, 2014.
- A. A. Frolov, D. Húsek, and P. Y. Polyakov. Comparison of seven methods for boolean factor analysis and their evaluation by information gain. *IEEE Transactions on Neural Networks and Learning Systems*, 27(3):538–550, 2015.
- Z. Gan, R. Henao, D. Carlson, and L. Carin. Learning deep sigmoid belief networks with data augmentation. In *International Conference on Artificial Intelligence and Statistics*, pages 268–276. PMLR, 2015.

- R. Giryes and M. Elad. Sparsity-based Poisson denoising with dictionary learning. *IEEE Transactions on Image Processing*, 23(12):5057–5069, 2014.
- I. J. Goodfellow, A. Courville, and Y. Bengio. Scaling up spike-and-slab models for unsupervised feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1902–1914, 2012.
- C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, and O. Winther. scVAE: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.
- S. Gu, L. Zhang, W. Zuo, and X. Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2862–2869, 2014.
- M. Haft, R. Hofman, and V. Tresp. Generative binary codes. *Formal Pattern Analysis & Applications*, 6(4):269–284, 2004.
- H. Hälvä, S. Le Corff, L. Lehéricy, J. So, Y. Zhu, E. Gassiat, and A. Hyvärinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 34, pages 1624–1633, 2021.
- M. Henniges, G. Puertas, J. Bornschein, J. Eggert, and J. Lücke. Binary sparse coding. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 450–457. Springer, 2010.
- M. Henniges, R. E. Turner, M. Sahani, J. Eggert, and J. Lücke. Efficient occlusive components analysis. *Journal of Machine Learning Research*, 15:2689–2722, 2014.
- G. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.
- F. Hirschberger. vc-GMM. Bitbucket repository. <https://bitbucket.org/fhirschberger/clustering>, 2019. Accessed: 2019-09-17.
- F. Hirschberger, D. Forster, and J. Lücke. A variational EM acceleration for efficient clustering at very large scales. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9787–9801, 2022.
- M. D. Hoffman. Poisson-uniform nonnegative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5361–5364. IEEE, 2012.
- A. Hyttinen, V. B. Pacela, and A. Hyvärinen. Binary independent component analysis: A non-stationarity-based approach. In *Uncertainty in Artificial Intelligence*, pages 874–884. PMLR, 2022.
- A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *International Conference on Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

- A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.
- A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.
- A. Hyvärinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pages 859–868, 2019.
- R. Imamura, T. Itasaka, and M. Okuda. Zero-shot hyperspectral image denoising with separable image prior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1416–1420, 2019.
- E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations*, 2017.
- Y. Jernite, Y. Halpern, and D. Sontag. Discovering hidden variables in noisy-OR networks using quartet tests. In *Advances in Neural Information Processing Systems*, volume 26, pages 2355–2363, 2013.
- Q. Jin, O. Miyashita, F. Tama, J. Yang, and S. Jonic. Poisson image denoising by piecewise principal component analysis and its application in single-particle x-ray diffraction imaging. *IET Image Processing*, 12(12):2264–2274, 2018.
- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.
- Z. Kadkhodaie and E. Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In *Advances in Neural Information Processing Systems*, volume 34, pages 13242–13254, 2021.
- M. E. Khan, G. Bouchard, K. P. Murphy, and B. M. Marlin. Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems*, pages 1108–1116, 2010.
- I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.
- D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- I. Kosmidis. CRAN packages for generalized linear models and with related methods, 2019. URL <https://cran.r-project.org/web/packages/cranly/vignettes/glms.html>.
- W. Kumwilaisak, T. Piriyatharawet, P. Lasang, and N. Thatphithakkul. Image denoising with deep convolutional neural and multi-directional long short-term memory networks under Poisson noise environments. *IEEE Access*, 8:86998–87010, 2020.

- H. Larochelle and Y. Bengio. Classification using discriminative restricted Boltzmann machines. In *International Conference on Machine Learning*, pages 536–543, 2008.
- Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.
- H. Lee, R. Raina, A. Teichman, and A. Y. Ng. Exponential family sparse coding with application to self-taught learning. In *International Joint Conference on Artificial Intelligence*, volume 9, pages 1113–1119, 2009.
- J. Li and D. Tao. Simple exponential family PCA. In *International Conference on Artificial Intelligence and Statistics*, pages 453–460. PMLR, 2010.
- M. Lu, J. Z. Huang, and X. Qian. Sparse exponential family principal component analysis. *Pattern Recognition*, 60:681–691, 2016.
- J. Lücke. Truncated variational expectation maximization. *arXiv preprint, arXiv:1610.03113*, 1610, 2019.
- J. Lücke and J. Eggert. Expectation truncation and the benefits of preselection in training generative models. *Journal of Machine Learning Research*, 11:2855–900, 2010.
- J. Lücke and M. Sahani. Maximal causes for non-linear component extraction. *Journal of Machine Learning Research*, 9:1227–67, 2008.
- J. Lücke, R. Turner, M. Sahani, and M. Henniges. Occlusive components analysis. In *Advances in Neural Information Processing Systems*, volume 22, pages 1069–77, 2009.
- D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.
- M. Makitalo and A. Foi. Optimal inversion of the anscombe transformation in low-count Poisson image denoising. *IEEE Transactions on Image Processing*, 20(1):99–109, 2010.
- J. D. McAuliffe and D. M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, pages 121–128, 2008.
- P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Routledge, 1989.
- S. Mohamed, Z. Ghahramani, and K. A. Heller. Bayesian exponential family PCA. In *Advances in Neural Information Processing Systems*, pages 1089–1096, 2009.
- S. Mohamed, K. A. Heller, and Z. Ghahramani. Bayesian and L1 approaches for sparse unsupervised learning. In *International Conference on Machine Learning*, 2012.
- A. Molina, A. Vergari, N. Di Mauro, S. Natarajan, F. Esposito, and K. Kersting. Mixed sum-product networks: A deep architecture for hybrid domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

- T. Mou, J. Huang, and F. O’Sullivan. The Gamma characteristic of reconstructed PET images: Implications for ROI analysis. *IEEE Transactions on Medical Imaging*, 37(5):1092–1102, 2018.
- H. Mousavi, J. Drefs, and J. Lücke. A double-dictionary approach learns component means and variances for V1 encoding. In *International Conference on Machine Learning, Optimization, and Data Science*, pages 240–244. Springer, 2020.
- H. Mousavi, M. Buhl, E. Guiraud, J. Drefs, and J. Lücke. Inference and learning in a latent variable model for Beta distributed interval data. *Entropy*, 23(5):552, 2021.
- 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.
- J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.
- B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–9, 1996.
- M. Opper and D. Saad. Tutorial on variational approximation methods. In *Advanced mean field methods: theory and practice.*, pages 129–159. MIT Press, 2000.
- J. Puertas, J. Bornschein, and J. Lücke. The maximal causes of natural scenes are edge filters. In *Advances in Neural Information Processing Systems*, volume 23, pages 1939–1947, 2010.
- T. Remez, O. Litany, R. Giryes, and A. M. Bronstein. Deep convolutional denoising of low-light images. *arXiv preprint arXiv:1701.01687*, 2017.
- T. Remez, O. Litany, R. Giryes, and A. M. Bronstein. Class-aware fully convolutional Gaussian and Poisson denoising. *IEEE Transactions on Image Processing*, 27(11):5707–5722, 2018.
- D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, 2014.
- J. T. Rolfe. Discrete variational autoencoders. In *International Conference on Learning Representations*, 2017.
- A. Rond, R. Giryes, and M. Elad. Poisson inverse problems by the plug-and-play scheme. *Journal of Visual Communication and Image Representation*, 41:96–108, 2016.
- S. Roweis. EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pages 626–632, 1998.
- S. T. Roweis. Factorial models and refiltering for speech separation and denoising. In *Eighth European Conference on Speech Communication and Technology*, volume 7, pages 1009–1012, 2003.
- S. Sajama and A. Orlitsky. Semi-parametric exponential family PCA. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

- J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett. Poisson noise reduction with non-local PCA. *Journal of Mathematical Imaging and Vision*, 48(2):279–294, 2014.
- L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4(1):61–76, 1996.
- A.-S. Sheikh and J. Lücke. Select-and-sample for spike-and-slab sparse coding. In *Advances in Neural Information Processing Systems*, volume 29, pages 3927–3935, 2016.
- A.-S. Sheikh, J. A. Shelton, and J. Lücke. A truncated EM approach for spike-and-slab sparse coding. *Journal of Machine Learning Research*, 15:2653–2687, 2014.
- A.-S. Sheikh, N. S. Harper, J. Drefs, Y. Singer, Z. Dai, R. E. Turner, and J. Lücke. STRFs in primary auditory cortex emerge from masking-based statistics of natural sounds. *PLOS Computational Biology*, 15(1):e1006595, 2019.
- J. A. Shelton, P. Sterne, J. Bornschein, A.-S. Sheikh, and J. Lücke. Why MCA? Nonlinear sparse coding with spike-and-slab prior for neurally plausible image encoding. In *Advances in Neural Information Processing Systems*, volume 25, pages 2285–2293, 2012.
- J. A. Shelton, J. Gasthaus, Z. Dai, J. Lücke, and A. Gretton. GP-select: Accelerating EM using adaptive subspace preselection. *Neural Computation*, 29(8):2177–2202, 2017.
- A. Shocher, N. Cohen, and M. Irani. “Zero-shot” super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3118–3126, 2018.
- T. Šingliar and M. Hauskrecht. Noisy-OR component analysis and its application to link analysis. *Journal of Machine Learning Research*, 7:2189–2213, 2006.
- M. W. Spratling. Pre-synaptic lateral inhibition provides a better architecture for self-organising neural networks. *Network: Computation in Neural Systems*, 10:285–301, 1999.
- M. W. Spratling. Learning image components for object recognition. *Journal of Machine Learning Research*, 7(5):793–815, 2006.
- H. Sprikeler, T. Zito, and L. Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *Journal of Machine Learning Research*, 15(1):921–947, 2014.
- Y. Tai, J. Yang, X. Liu, and C. Xu. MemNet: A persistent memory network for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4539–4547, 2017.
- A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 47(10):2807–2820, 1999.
- A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.
- A. Teymurazyan, T. Riauka, H.-S. Jans, and D. Robinson. Properties of noise in positron emission tomography images reconstructed with filtered-backprojection and row-action maximum likelihood algorithm. *Journal of Digital Imaging*, 26(3):447–456, 2013.

- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.
- M. K. Titsias and M. Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems*, volume 24, pages 2339–2347, 2011.
- B. Tolooshams, A. Song, S. Temereanca, and D. Ba. Convolutional dictionary learning based auto-encoders for natural exponential-family distributions. In *International Conference on Machine Learning*, pages 9493–9503. PMLR, 2020.
- N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271–82, 1998.
- I. Valera and Z. Ghahramani. Automatic discovery of the statistical types of variables in a dataset. In *International Conference on Machine Learning*, pages 3521–3529, 2017.
- W. J. van der Linden and R. K. Hambleton. *Handbook of Modern Item Response Theory*. Springer Science & Business Media, 2013.
- J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366, 1998.
- A. Vergari, A. Molina, R. Peharz, Z. Ghahramani, K. Kersting, and I. Valera. Automatic Bayesian density analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5207–5215, 2019.
- M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends® in Machine Learning. Now Publishers, Inc., 2008.
- M. Yin, Y. Yue, and M. Zhou. ARSM: Augment-REINFORCE-swap-merge estimator for gradient backpropagation through Categorical variables. In *International Conference on Machine Learning*, 2019.
- K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.
- K. Zhang, W. Zuo, and L. Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.
- D. Zhou and X.-X. Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE. In *Advances in Neural Information Processing Systems*, volume 33, pages 7234–7247, 2020.
- M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative Binomial process and Poisson factor analysis. In *International Conference on Artificial Intelligence and Statistics*, pages 1462–1471, 2012.