Abstract

We propose rectified factor networks (RFNs) as generative unsupervised models, which learn robust, very sparse, and non-linear codes with many code units. RFN learning is a variational expectation maximization (EM) algorithm with unknown prior which includes (i) rectified posterior means, (ii) normalized signals of hidden units, and (iii) dropout. Like factor analysis, RFNs explain the data variance by their parameters. For pretraining of deep networks on MNIST, rectangle data, convex shapes, NORB, and CIFAR, RFNs were superior to restricted Boltzmann machines (RBMs) and denoising autoencoders. On CIFAR-10 and CIFAR-100, RFN pretraining always improved the results of deep networks for different architectures like AlexNet, deep supervised net (DSN), and a simple “Network In Network” architecture. With RFNs success is guaranteed.

1. Introduction

The advent of deep learning (Hinton & Salakhutdinov, 2006; Bengio et al., 2007) and its success in both academic challenges and industrial applications is to a large part based on better input representations compared to previous approaches (Sutskever et al., 2014; Dahl et al., 2012; Krizhevsky et al., 2012; Graves et al., 2013; Ciresan et al., 2012). These representations are sparse, highly overcomplete, and hierarchical. Input codes at higher levels capture more complex and higher non-linear explanatory factors of the observed input (Bengio et al., 2013). Rectified linear units (ReLU) (Nair & Hinton, 2010; Glorot et al., 2011; Zeiler et al., 2013) and dropout (Hinton et al., 2012; Warde-Farley et al., 2013; Baldi & Sadowski, 2014) are techniques to obtain very sparse representations. Beyond complex and abstract representation on different levels, in particular these sparse representations increased the performance of deep learning architectures. Sparse representations were originally motivated by findings in sensory neural systems and are well known objectives in machine learning (Olshausen & Field, 1996; Hinton & Ghahramani, 1997; Girolami, 2001). The key advantage of sparse representations is that dependencies between coding units are much easier to model and much easier to interpret. Most importantly, distinct concepts are much less likely to interfere in sparse representations. Using sparse representations, structures in data often break down to co-occurrences of features, i.e. features that are simultaneously present in samples. In bioinformatics sparse codes excelled in biclustering of gene expression data (Hochreiter et al., 2010) and in finding DNA sharing patterns between humans and Neanderthals (Hochreiter, 2013). Successful deep learning methods like autoencoders and restricted Boltzmann machines (RBMs) construct highly overcomplete sparse representations with hundreds or thousands of coding units.

Our goal is to advance sparse, overcomplete, and hierarchical representations by generative models like factor analysis for: (1) modeling the input noise, (2) decorrelating the hidden units, (3) estimating the information content in hidden units. Consequently, noisy input components are
down-weighted during learning and coding units regard the noise in the detected patterns. In contrast, highly predictive inputs are up-weighted and thereby allow identifying interesting parts in the input. Decorrelation leads to a better coverage of the feature space and forces more variety into hidden units. Thus, rare and small input structures are more likely to be detected. Furthermore, the sparseness of hidden units is increased by decorrelation, since sparse variables have low correlations as known from independent component analysis. High information content of a hidden unit is reflected by a large activation which in turn hints at higher specificity, higher precision, or a more consistent input structure.

However generative models with many sparse hidden units face several severe problems. First, generative models prefer to explain parts of data by noise instead of a small signal. Since additive noise assumes equal noise for each sample, small or rare signals are explained away by noise. The second problem is that sparse priors like Laplacians for the coding units are computationally very expensive when calculating the code for a new input (Olshausen & Field, 1996; Hoyer, 2004; Rozell et al., 2008; Gregor & LeCun, 2010). The rectified factor network constructs sparse codes with many coding units, which possess the advantages of generative models but still select models using a variational Bayesian learning procedure, does not yield sparse posteriors (Harva & Kaban, 2010). Variational updates are expensive as they have to compute the inverse of the hidden units’ empirical covariance matrix. Furthermore, generative models with sparse priors do not ensure sparse posterior means. For example, rectified factor analysis, which rectifies Gaussian priors and selects models using a variational Bayesian learning procedure, does not yield sparse posteriors (Harva & Kaban, 2005, 2007). We propose rectified factor networks (RFNs) which possess the advantages of generative models but still construct sparse codes with many coding units.

2. Rectified Factor Network

A rectified factor network (RFN) consists of a single or stacked factor analysis model(s) with many sparse, non-linear hidden units (factors). The factor analysis model is

$$ v = Wh + \epsilon. $$

The prior $h \sim N(0, I)$ on the hidden units $h \in \mathbb{R}^l$, i.e. the factors, and the noise $\epsilon \sim N(0, \Psi)$ on visible units $v \in \mathbb{R}^m$, i.e. the input or observations, are independent. The model parameters are the weight (factor loading) matrix $W \in \mathbb{R}^{m \times l}$ and the noise covariance matrix $\Psi \in \mathbb{R}^{m \times m}$.

We assume diagonal $\Psi$ to explain correlations between input components by the hidden units (signal) and not by correlated noise. The factor analysis model is depicted in Fig. 1. Given the mean-centered data $\{v\} = \{v_1, \ldots, v_n\}$ and the covariance matrix $C = \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T$, we base RFN learning on the expectation maximization (EM) update for maximizing the likelihood of the factor analysis model. In the EM algorithm, the hidden unit posterior $p(h_i | v_i)$ is Gaussian with mean $\mu_{h_i | v_i}$ and covariance matrix $\Sigma_{h_i | v_i}$. The posterior is computed in the E-step for the old parameters ($W, \Psi$) which are then updated in the M-step. See supplementary Section S10 for maximum likelihood factor analysis.

We modify this EM algorithm with respect to four issues: (i) **rectifying** the posterior values (enforcing non-negative means) leads to sparse and non-linear codes; (ii) **normalizing** the posterior values (signal part of the posterior) across data samples increases the information in the coding units and keeps the reconstruction error low; (iii) **dropout** of posterior means regularizes and increases sparseness; (iv) **gradient descent** update in the M-step in the Newton direction allows stochastic updates by mini-batches and dropout while ensuring efficient implementations on GPUs. Rectifying can be implemented by the posterior constraint method (Ganchev et al., 2010; Graca et al., 2007). The posterior constraint method minimizes the Kullback-Leibler distance between the variational distribution and the posterior with particular constraints on the posterior. In our case the constraints are non-negative posterior means. However, this approach leads to a constraint quadratic (in the number of factors) optimization problem for computing the posterior means (see supplementary Section S6.1). Hence, the posterior constraint method is computationally too complex for codes with many units. Furthermore, normalization, which is essential for our approach, cannot be realized by posterior constraints. Instead of the Kullback-Leibler distance of the posterior constraint method, we minimize
3.1. Fixed Point

We derive a fixed point of the algorithm which is independent of the posterior modification. The fixed point solution explains the variation in the data by noise and signal like factor analysis. How much data variance is explained by signal and how much by noise depends on the information in the hidden units about the input. The following theorem states that the RFN model explains the data covariance matrix $C$ by a noise part $\Psi$ and a signal part $WSW^T$, where $S$ is an estimate of the second moments of the hidden units.

Theorem 1 (Fixed Point). For fixed points of RFN learning with full $\Psi$ update holds

$$C = \Psi + WSW^T.$$  

Proof. The fixed point equation for the $W$ update is

$$\Delta W = U S^{-1} - W = 0 \Rightarrow U - W S = 0 \Rightarrow U W^T - W S W^T = 0,$$  

where the last equation is obtained by multiplying the previous from the right hand side by $W^T$. The fixed point equation for the full (not only diagonal) $\Psi$ update is

$$\Psi = C - U W^T - W U^T + WSW^T$$  

$$= C - WU^T,$$

where we inserted Eq. (3). Since $WSW^T$ in Eq. (3) is symmetric, we can insert $WU^T = UW^T = C - \Psi$ in Eq. (5) and obtain the equation of the theorem. \qed

Like with standard factor analysis, the data variance is explained by the model via the parameters $\Psi$ (noise) and $W$ (signal). In contrast to standard factor analysis, the second moment $S$ of the unknown prior is estimated. With dropout and mini-batches, $\mu_{h_i|v_i}$ become random variables and $U$ and $S$ are replaced by their expectations.

How much data variation is explained by signal $WSW^T$ and how much is explained by noise $\Psi$ depends on (a) the model complexity like the number of factors, (b) the measurement noise on the visible units, and (c) the amount of information about the visible units $v$ that is coded in the hidden units $\mu_{h_i|v_i}$. If the $\mu_{h_i|v_i}$ contain sufficient amount of information about $v_i$, then the fixed point equation can be fulfilled when restricting $\Psi$ to a diagonal matrix. We ensure sufficient amount of information in the hidden units by minimizing the Fréchet distance between the variational distribution and the posterior. That a diagonal $\Psi$ is sufficient, can be shown by a large $U$ stemming from correlated $v_i$ and $\mu_{h_i|v_i}$. A large $U$ leads to a large $W$ which in turn keeps both $\Sigma_{h_i|v_i}$ and $\Psi$ small. $\Psi$ is basically the least squares error (see more details in the supplementary Section S8).

3. Analysis of RFN Learning

In this section RFN learning is analyzed and justified. The RFN objective is unknown because the modification of the posterior disconnects any prior from the posterior. The modified posterior also has a prior but it is not known, thus the likelihood cannot be computed. First, we show via a fixed point analysis that trained RFN models explain the data variance via noise and signal like factor analysis. Next, RFN learning is formulated as a variational EM algorithm that maximizes the data likelihood for an unknown prior. Finally, we derive the Newton update rule that is used by RFN learning.
3.2. Variational EM

The following theorem states that RFN learning can be described as a variational EM algorithm (Ghahramani & Hinton [1998]; Neal & Hinton [1998]; Jordan et al. [1999]; Beal 2003; Girolami 2001; Palma et al. 2006).

**Theorem 2** (variational EM). Learning rectified factor networks (RFNs) is equivalent to a variational expectation maximization (EM) algorithm for factor analysis with an unknown prior on the factors.

**Proof.** We consider maximum likelihood factor analysis with a variational EM algorithm. The Gaussian prior is parameterized by the variational parameter \( \xi \):

\[
p(h; \xi) = (2\pi)^{-\frac{n}{2}} \exp \left(-\frac{1}{2} (h - \xi)^T (h - \xi) \right). \tag{5}
\]

Using a variational distribution \( Q(h | v) \), the data likelihood can be bounded:

\[
\log p(v) - D_{KL}(Q(h | v) \parallel p(h | v)) = - \int Q(h | v) \log \frac{Q(h | v)}{p(h | v)} dh = \int Q(h | v) \log p(v | h) dh - D_{KL}(Q(h | v) \parallel p(h)) = F(W, \Psi, \xi | \tilde{W}, \tilde{\Psi}, \tilde{\xi}),
\]

where \( D_{KL} \) denotes the Kullback-Leibler divergence. \((W, \Psi, \xi)\) are the parameters that describe \( Q(h | v) \), while \((\tilde{W}, \tilde{\Psi}, \tilde{\xi})\) are the parameters of the prior \( p(h | v) \), and \( \xi \) are the parameters of the prior \( p(h) \). The variational EM maximizes \( F \) in its M-step with respect to the parameters \((W, \Psi)\) and in its E-step with respect to \( Q \) given by \((\tilde{W}, \tilde{\Psi}, \tilde{\xi})\). The posterior \( p(h | v) \) is a Gaussian with following mean and covariance matrix:

\[
\mu_{h|v} = (I + W^T \Psi^{-1} W)^{-1} (W^T \Psi^{-1} v + \xi), \quad \Sigma_{h|v} = (I + W^T \Psi^{-1} W)^{-1}. \tag{7}
\]

The variational EM minimizes in its E-step \( D_{KL}(Q(h | v) \parallel p(h | v)) \). We approximate the unknown posterior \( p(h_i | v_i) \) by the variational distribution \( Q(h_i | v_i) \). We set \( Q(h_i | v_i) = p(h_i | v_i; \xi_i) = N(h_i; \mu_{h|v}, \Sigma_{h|v}) \) which is the posterior of the model with Gaussian prior \( p(h; \xi) \) for some \( \xi_i \). \( \mu_Q \) is the result of RFN learning after steps (i)—(iii): rectifying, normalizing, and dropout. We set \( \mu_{h|v} = \mu_Q \) and solve Eq. (7) for \( \xi \):

\[
\xi = (W^T \Psi^{-1} W + I)^{-1} \mu_Q - W^T \Psi^{-1} v. \tag{8}
\]

See more details in the supplementary Section [S7]. In the variational EM, the distribution \( Q \) is a variational approximation to the posterior (Ghahramani & Hinton [1998]).

Also RFN learning approximates the posterior of a prior which can be expressed by a parametrized family of shifted Gaussians.

\[
\int p(h | \xi) p(\xi) d\xi \approx \sum_i p(h | \xi_i). \tag{9}
\]

\( \mathcal{S} \) estimates the second moment of the prior via \( Q \), which is an exact approximation of the posterior after learning converged. The RFN objective is to maximize \( \int Q(h | v) \log p(v | h) dh \) with respect to \((W, \Psi)\) in the M-step. Thus, the M-step minimizes the reconstruction error weighted by \( \Psi \) for given hidden activations \( \mu \). See more details on the RFN objective in supplementary Section [S9].

3.3. Gradient Descent Update Rule

For RFN learning, we derive the gradient descent update rule in the Newton direction. The M-step maximizes the objective

\[
\mathcal{L} = \int_{\mathbb{R}^n} Q_i(h_i | v_i) \log \left[ p(v_i | h_i; W, \Psi) \right] dh_i.
\]

\[
\log \mathcal{L} = - \frac{n}{2} \log (2\pi) - \frac{n}{2} \log |\Psi| - \frac{1}{2} \sum_{i=1}^n v_i^T \Psi^{-1} v_i + \sum_{i=1}^n E_{h_i | v_i} (v_i^T \Psi^{-1} W h_i)
\]

\[
- \frac{1}{2} E_{h_i | v_i} \left( \sum_{i=1}^n h_i^T W^T \Psi^{-1} W h_i \right)
\]

The Hessian \( H_W \) of \( \frac{\partial}{\partial \log \mathcal{L}} \) with respect to \( W \) as a vector is:

\[
H_W = \frac{\partial \text{vec} \left( \frac{\partial \mathcal{L}}{\partial \log \mathcal{L}} \right)}{\partial \text{vec}(W)^T} = \frac{\partial \text{vec} (\Psi^{-1} U - \Psi^{-1} W S)}{\partial \text{vec}(W)^T} = - S \otimes \Psi^{-1}, \tag{10}
\]

where \( \otimes \) is the Kronecker product of matrices. For the product of the negative inverse Hessian with the gradient we have:

\[
- H_W^{-1} \text{vec} (\Psi^{-1} U - \Psi^{-1} W S) = (S^{-1} \otimes \Psi) \text{vec} (\Psi^{-1} U - \Psi^{-1} W S) = \text{vec} (\Psi (\Psi^{-1} U - \Psi^{-1} W S) S^{-1}) = \text{vec} (U S^{-1} - W).
\]

Thus, if we apply a Newton update then the update direction for \( W \) in the M-step is

\[
\Delta W = U S^{-1} - W. \tag{12}
\]
We recover the exact EM update if the learning rate is 1. The Newton update converges fast, but comes at the cost of inverting $S$. For large $S$ matrices, $\Delta W = U - W S$ is an alternative update rule, though it converges slower.

The Hessian $H_\Psi$ of $\frac{1}{2} \log \mathcal{L}$ with respect to $\Psi$ as a vector is a diagonal matrix with $[H_\Psi]_{ij} = 0$ for $i \neq j$ and for $i = j$:

$$
[H_\Psi]_{ii} = \frac{1}{\Psi_{ii}^2} - \frac{2 c_i}{\Psi_{ii}} = \frac{1}{\Psi_{ii}^2} \left(1 - \frac{2 c_i}{\Psi_{ii}}\right),
$$

\[c_i = [C - U W^T - W U^T + W S W^T]_{ii}\]

Since we maximize the bound, we have to ensure that the Hessian is negative definite, that is, $[H_\Psi]_{ii} < 0$ or $\Psi_{ii} < 2 c_i$. Therefore we replace $c_i$ by $\Psi_{ii}$ and get $[H_\Psi]_{ii} = -1/\Psi_{ii}^2$. For another approach see supplementary Section S5.2

The Newton update direction is:

$$
\Delta \Psi_{ii} = -\frac{[\nabla \Psi \log \mathcal{L}]_{ii}}{[H_\Psi]_{ii}} = c_i - \Psi_{ii}.
$$

Again we recover the EM update if the learning rate is 1.

### 4. Experiments

#### 4.1. Comparison of RFNs to Other Unsupervised Methods

In this section, we assess the performance of rectified factor networks (RFNs) as unsupervised methods for data representation. The number of components are fixed to 50 for each method. We compare

1. RFN: rectified factor networks,
2. RFNn: RFNs without normalization,
3. DAE: denoising autoencoders with rectified linear units,
4. RBM: restricted Boltzmann machines with Gaussian visible units and hidden binary units,
5. FAsp: factor analysis with Jeffrey’s prior ($p(z) \propto 1/z$) on the hidden units which is sparser than a Laplace prior,
6. FAlap: factor analysis with Laplace prior on the hidden units,
7. ICA: independent component analysis by FastICA (Hyvärinen & Oja, 1999),
8. SFA: sparse factor analysis with a Laplace prior on the parameters,
9. FA: standard factor analysis, and
10. PCA: principal component analysis.

We generated nine different benchmark datasets (D1 to D9), where each dataset consists of 100 instances for averaging the results. Each instance consists of 100 samples and 100 features resulting in a $100 \times 100$ data matrix. Into these data matrices, structures are implanted as biclusters (Hochreiter et al., 2010). A bicluster is a pattern across a particular number of features which is found in a particular number of samples. The size of the bicluster is given by the number of features that form the pattern and by the number of samples in which the pattern is found. The datasets had different noise levels and different bicluster sizes. We considered large and small bicluster sizes, where large biclusters have 20–30 samples and 20–30 features, while small biclusters have 3–8 samples and 3–8 features. The signal strength (scaling factor) of a pattern in a sample was randomly chosen according to the Gaussian $\mathcal{N}(1, 1)$. Finally, to each data matrix background noise was added, where the noise is distributed according to a zero-mean Gaussian with standard deviation 1, 5, or 10. The datasets are described in Tab. II.

| Table 1. Overview over the datasets. Shown is the background noise (“noise”), the number of large biclusters ($n_1$), and the number of small biclusters ($n_2$). |
|-------------------|---|---|---|---|---|---|---|---|---|
|                  | D1 | D2 | D3 | D4 | D5 | D6 | D7 | D8 | D9 |
| noise            | 1  | 5  | 10 | 1  | 5  | 10 | 1  | 5  | 10 |
| $n_1$            | 10 | 10 | 10 | 15 | 15 | 15 | 5  | 5  | 5  |
| $n_2$            | 10 | 10 | 10 | 5  | 5  | 5  | 15 | 15 | 15 |

We evaluated the methods according to the sparseness of the components. Sparseness is the percentage of the components for which their absolute value is smaller than 0.01. We aim at very sparse representations. The second criterion is the reconstruction error when reconstructing the input from the values of the hidden units. For computing the reconstruction error, we used the Frobenius norm of the matrix difference between reconstructed matrix and data matrix. Thus, we used the sum of the squared reconstruction errors across samples. We aim at keeping the information of the input in the hidden units. If applicable, the third criterion is the error in modeling the covariance by the generative model. Again we used the Frobenius norm of the matrix difference between model covariance matrix and data covariance matrix. We aim at representing the covariance structure of the data by the model parameters. See supplementary Sections S3 and S4 for more details on the data and supplementary Section S2 for information on hyperparameter selection for the different methods. The results are given in Tab. II by the mean over 100 instances for each dataset. In the supplementary Section S3 we also confirm the result for models with 100 and 150 coding units. Furthermore, we confirmed the results for different noise levels (See supplementary Section S4). RFNs yield the most sparse code with a low reconstruction error and a good approximation of the covariance matrix.
Table 2. Comparison of RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. Criteria are: sparseness of the factors (SP) reported in percent (%) of the samples with a factor smaller than 0.01, reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean of 100 instances. The maximal value in the table was set to 999. RFNs yield the most sparse code with a low reconstruction error and a good approximation of the covariance matrix.

|       | SP  | ER  | CO  | SP  | ER  | CO  | SP  | ER  | CO  | SP  | ER  | CO  |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| RFN   | 72  | 74  | 11  | 75  | 240 | 72  | 75  | 462 | 260 | 72  | 79  | 12  |
| RFNn  | 68  | 122 | 32  | 74  | 285 | 97  | 74  | 537 | 331 | 65  | 144 | 48  |
| DAE   | 61  | 82  | —   | 66  | 243 | —   | 66  | 461 | —   | 60  | 88  | —   |
| RBM   | 22  | 106 | —   | 11  | 301 | —   | 10  | 566 | —   | 22  | 113 | —   |
| FAsp  | 37  | 469 | 999 | 40  | 999 | 999 | 40  | 999 | 999 | 40  | 999 | 999 |
| FAlap | 4   | 50  | 392 | 4   | 228 | 135 | 5   | 443 | 406 | 4   | 51  | 477 |
| ICA   | 16  | 35  | —   | 16  | 168 | —   | 16  | 327 | —   | 16  | 35  | —   |
| SFA   | 1   | 42  | 26  | 1   | 210 | 61  | 1   | 409 | 220 | 1   | 41  | 17  |
| FA    | 1   | 42  | 13  | 1   | 210 | 58  | 1   | 409 | 214 | 1   | 41  | 17  |
| PCA   | 0   | 35  | —   | 0   | 168 | —   | 0   | 327 | —   | 0   | 35  | —   |

We have shown that RFNs construct very sparse codes of the input with low reconstruction errors that explain the covariance structure of the input. Besides RFNs, only restricted Boltzmann machines (RBMs) and (denoising) autoencoders (DAEs) are able to generate sparse codes with many hidden units. Next we show the benefits of sparse codes computed by RFNs when compared to RBMs and DAEs at pretraining deep networks.

4.2. Unsupervised Pretraining for Permutation Invariant Networks

In this section, we assess the performance of rectified factor networks (RFNs) as a pretraining procedure for the first layers of deep networks. We stacked RFNs in the same way as described by [Vincent et al., 2010], namely first training a single layer RFN and then passing the resulting representation as input for training to the next RFN. We conducted experiments with deep network architectures where the first layer was pretrained by RFNs (RFN-1). We also tested few architectures, where we stacked 3 RFNs giving a 3-layered RFN. The classification performance of deep networks with RFN pretrained layers was compared to (i) support vector machines, (ii) deep networks pretrained by stacking denoising autoencoders (SDAE), (iii) stacking regular autoencoders (SAE), (iv) restricted Boltzmann machines (RBM), and (v) stacking restricted Boltzmann machines (DBN).

The benchmark datasets are from previous publications [LeCun et al., 2004; Vincent et al., 2010; Larochelle et al., 2007; Krizhevsky, 2009] and contain: (i) MNIST (original MNIST), (ii) basic (a smaller subset of MNIST for training), (iii) bg-rand (MNIST digits with random noise background), (iv) bg-img (MNIST digits with random image background), (v) rect (discrimination between tall and wide rectangles), (vi) rect-img (discrimination between tall and wide rectangular images overlayed on different background images), (vii) convex (discrimination between convex and concave shapes), (viii) CIFAR-10 (60k 32x32 color
Rectified Factor Networks

Figure 2. Various randomly selected filters trained on benchmark data sets using an RFN with 1024 hidden units. The panels (a)–(c) show filters for MNIST variations, while (d) and (e)–(f) show filters for convex and rectangle, respectively. Filters learned from CIFAR-10 and NORB are shown in panel (g)–(h). RFNs learned various kinds of complex filters, such as stroke, local and global blob detectors. Panel (c) shows that RFN filters unaffected by the background noise.

images in 10 classes, with 6k images per class), and (ix) NORB (29,160 stereo image pairs of 50 toys belonging to 5 generic categories). The dataset characteristics in terms of size of training, validation and test set is given in the second column of Tab. 3. In all experiments we apply no further data preprocessing except median centering and learn the representation in an unsupervised fashion. We followed Vincent et al. (2010) and selected the models based on the validation set performance. The RFNs have as possible hyperparameters: (i) the number of units in the layers from $\{1024, 2048, 4096\}$ and (ii) the dropout rate from $\{0.0, 0.25, 0.5, 0.75\}$. The learning rates were fixed to $\eta_W = 0.1$ and $\eta_\Psi = 0.01$ which we found to be well suited for RFN learning on separate artificial toy datasets, where the task was to find biclusters (Hochreiter et al., 2010). For supervised fine-tuning with stochastic gradient descent, we selected the learning rate from $\{0.1, 0.01, 0.001\}$, the masking noise from $\{0.0, 0.25\}$, and the number of layers from $\{1, 3\}$. Again following Vincent et al. (2010), fine-tuning was stopped early, where the stopping time was selected based on the validation set performance.

The results of the comparison of deep networks with RFN pretrained low layers and other models are given in Tab. 3 while Fig. 2 shows some examples of the learned filters. The test error rate is reported for seven classification problems. A 95% confidence interval is computed according to Vincent et al. (2010). The result of the best performing method is given in bold, as well as the result of those methods for which confidence intervals overlap. In only one case RFN low layer pretraining was significantly worse than the best method but still the second best method. In six out of the nine experiments RFN low layer pretraining performed best, where in four cases it was significantly better than all other methods.

4.3. Unsupervised Pretraining for Convolution Networks

As convolutional networks are the state-of-the-art in vision, we assess the performance of RFNs for pretraining the first layer of deep convolutional networks on two benchmark datasets: CIFAR-10 and CIFAR-100. We do not use RFNs for pretraining the whole network but only the first layers. All higher layers are randomly initialized as described in the publications where they were introduced. We conducted experiments with three deep convolutional network architectures where the first convolution layer is pretrained by RFNs: (i) the AlexNet (Krizhevsky et al., 2012), (ii) Deeply Supervised Networks (DSN) (Lee et al., 2014), and (iii) our 5-Convolution-Network-In-Network (5C-NIN) similar to Lin et al. (2013)’s Network In Networks. Next, we describe the datasets and the architecture of the 5C-NIN net.

The CIFAR datasets contain 60k RGB-color images of size 32x32, divided in into 5k training and 10k testing images, split between 10 and 100 categories, respectively. Both datasets were preprocessed in the same way as described by Goodfellow et al. (2013), namely by global contrast normalization and ZCA whitening. Additionally, the datasets were augmented by padding the images with four zero pixels at all sides. For data augmentation, at the beginning of every epoch, images in the training set were distorted...
Table 3. Comparison of deep networks with first layers pretrained by RFNs and other models. Test error rate on all considered classification problems is reported together with a 95% confidence interval. The result of the best performing method is given in bold, as well as those for which confidence intervals overlap. The first column gives the dataset, the second the size of training, validation and test set, the last column indicates the number of hidden layers of the RFN pretrained deep network which is chosen according to the validation set performance. In only one case RFN lower layer pretraining was significantly worse than the best method but still the second best method. In six out of the nine experiments RFN lower layer pretraining performed best, where in four cases it was significantly better than all other methods.

| Dataset    | SVM            | RBM            | DBN            | SAE            | SDAE           | RFN            |
|------------|----------------|----------------|----------------|----------------|----------------|----------------|
| MNIST      | 1.40 ± 0.23    | 1.21 ± 0.21    | 1.24 ± 0.22    | 1.40 ± 0.23    | 1.28 ± 0.22    | 1.27 ± 0.22 (1) |
| basic      | 3.03 ± 0.15    | 3.94 ± 0.17    | 3.11 ± 0.15    | 3.46 ± 0.16    | 2.84 ± 0.15    | 2.66 ± 0.14 (1) |
| bg-rand    | 14.58 ± 0.31   | 9.80 ± 0.26    | 6.73 ± 0.22    | 11.28 ± 0.28   | 10.30 ± 0.27   | 7.94 ± 0.24 (3) |
| bg-img     | 22.61 ± 0.37   | 16.15 ± 0.32   | 16.31 ± 0.32   | 23.00 ± 0.37   | 16.68 ± 0.33   | 15.66 ± 0.32 (1) |
| rect       | 2.15 ± 0.13    | 4.71 ± 0.19    | 2.60 ± 0.14    | 2.41 ± 0.13    | 1.99 ± 0.12    | 0.63 ± 0.06 (1) |
| rect-img   | 24.04 ± 0.37   | 23.69 ± 0.37   | 22.50 ± 0.37   | 24.05 ± 0.37   | 21.59 ± 0.36   | 20.77 ± 0.36 (1) |
| convex      | 19.13 ± 0.34   | 19.92 ± 0.35   | 18.63 ± 0.34   | 18.41 ± 0.34   | 19.06 ± 0.34   | 16.41 ± 0.32 (1) |
| NORB       | 11.6 ± 0.40    | 8.31 ± 0.35    | -              | 10.10 ± 0.38   | 9.50 ± 0.37    | 7.00 ± 0.32 (1) |
| CIFAR      | 62.7 ± 0.95    | 40.39 ± 0.96   | 43.38 ± 0.97   | 43.25 ± 0.97   | -              | 41.29 ± 0.95 (1) |

Table 4. The upper part of table shows three state-of-the-art convolutional deep networks pretrained in the first convolution layer with RFNs. The results are compared to the same architectures with random initialization for the first layer. The first column gives the network architecture, namely, AlexNet (Krizhevsky et al., 2012), Deeply Supervised Networks (DSN) (Lee et al., 2014), and our 5-Convolution-Network-In-Network (5C-NIN). The remaining columns report the test error rate with (“RFN”) and without (“org”) RFN pretraining for the benchmark datasets CIFAR-10 and CIFAR-100. For CIFAR-100 the DSN model is not available, therefore we could not test RFN pretraining. The lower part of the table reports the performance of other state-of-the-art networks: Network In Network (NIN), Maxout Networks (MN), and Graham’s DeepCNN. Our 5C-NIN outperformed, with the exception of DeepCNN, all other networks and gave the second best results reported so far on these datasets. In all cases pretraining with RFNs decreased the test error rate.

| Dataset | CIFAR-10 | CIFAR-100 |
|---------|----------|-----------|
| org RFN | org RFN augmented |
| AlexNet | 18.21 | 18.04 | 46.18 | 45.80 |   |
| DSN     | 7.97 | 7.74 | 34.57 |   | ✓  |
| 5C-NIN  | 7.81 | 7.63 | 29.96 | 29.75 | ✓  |
| NIN     | 8.81 | - | 35.68 | - | ✓  |
| MN      | 9.38 | - | 38.57 | - | ✓  |
| DeepCNN | 6.28 | - | 24.30 | - | ✓  |

5. Conclusion

We have introduced rectified factor networks (RFNs) for constructing very sparse and non-linear input representations with many coding units in a generative framework.
Like factor analysis, RFN learning explains the data variance. Furthermore, RFN learning can be considered as a variational EM algorithm. In contrast to generative methods with sparse priors, RFNs have a lower reconstruction error, explain the data variance better, and yield sparser codes. On MNIST, rectangle data, convex shapes, NORB, and CIFAR, RFNs led to better results than RBMs and autoencoders. On CIFAR-10 and CIFAR-100, RFNs improved the results of all tested deep architectures like AlexNet, deep supervised net, and a simple “Network In Network” architecture. RFNs are geared to large datasets, sparse coding, and many representational units, therefore they have high potential as unsupervised deep learning techniques.

Acknowledgment
The Tesla K40 used for this research was donated by the NVIDIA Corporation.

References
Baldi, P. and Sadowski, P. The dropout learning algorithm. Artificial Intelligence, 210:78–122, 2014. doi: 10.1016/j.artint.2014.02.004.

Beal, M. J. Variational algorithms for approximate Bayesian inference. PhD thesis, University College London, The Gatsby Computational Neuroscience Unit, 17 Queen Square London WC1N 3AR, 2003.

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. Greedy layer-wise training of deep networks. In Schölkopf, B., Platt, J. C., and Hoffman, T. (eds.), Advances in Neural Information Processing Systems 19 (NIPS 2006), pp. 153–160. MIT Press, 2007.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8):1798–1828, 2013.

Ciresan, D. C., Meier, U., and Schmidhuber, J. Multi-column deep neural networks for image classification. In IEEE Conference on Computer Vision and Pattern Recognition CVPR 2012, 2012. Long preprint arXiv:1202.2745v1 [cs.CV].

Dahl, G. E., Yu, D., Deng, L., and Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1):33–42, 2012.

Dowson, D. C. and Landau, B. V. The Fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis, 12:450–455, 1982.

Fréchet, M. Sur la distance de deux lois de probabilité. C. R. Acad. Sci. Paris, 244:689–692, 1957.

Ganchev, K., Graca, J., Gillenwater, J., and Taskar, B. Posterior regularization for structured latent variable models. Journal of Machine Learning Research, 11:2001–2049, 2010.

Ghahramani, Z. and Hinton, G. E. Variational learning for switching state-space models. Neural Computation, 12:963–996, 1998.

Girolami, M. A variational method for learning sparse and overcomplete representations. Neural Comput., 13(11):2517–2532, 2001.

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In Gordon, G., Dunson, D., and Dudk, M. (eds.), JMLR W&CP: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011), volume 15, pp. 315–323, 2011.

Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. Maxout networks. ArXiv e-prints, 2013.

Graca, J. V., Ganchev, K., and Taskar, B. Expectation maximization and posterior constraints. In Platt, J.C., Koller, D., Singer, Y., and Roweis, S.T. (eds.), Advances in Neural Information Processing Systems, volume 20, pp. 569–576, 2007.

Graham, Benjamin. Fractional max-pooling. CoRR, abs/1412.6071, 2014. URL http://arxiv.org/abs/1412.6071

Graves, A., Mohamed, A.-R., and Hinton, G. E. Speech recognition with deep recurrent neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649. IEEE, 2013.

Gregor, K. and LeCun, Y. Learning fast approximations of sparse coding. In Fünkranz, J. and Joachims, T. (eds.), Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 399–406. Omnipress 2010, ISBN 978-1-60558-907-7, 2010.

Harva, M. and Kaban, A. A variational bayesian method for rectified factor analysis. In Proc. Int. Joint Conf. on Neural Networks (IJCNN’05), pp. 185–190, 2005.

Harva, M. and Kaban, A. Variational learning for rectified factor analysis. Signal Processing, 87(3):509–527, 2007. doi: 10.1016/j.sigpro.2006.06.006.

Hinton, G. E. and Ghahramani, Z. Generative models for discovering sparse distributed representations. Philos. T. R. Soc. B, 352:1177–1190, 1997.

Hinton, G. E. and Salakhutdinov, R. Reducing the dimensionality of data with neural networks. Science, 313(5786):504–507, 2006.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Improving neural networks by preventing co-adaptation of feature detectors. CoRR, abs/1207.0580, 2012.

Hochreiter, S. HapFABIA: Identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Res., 41(22):e202, 2013. doi: 10.1093/nar/gkt1013. URL http://nar.oxfordjournals.org/content/41/22/e202.abstract

Rectified Factor Networks
Hochreiter, S., Bodenstein, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., VanSanden, S., Lin, D., Talloren, W., Bijnens, L., et al. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010. doi: 10.1093/bioinformatics/btq227.

Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469, 2004.

Hyvärinen, A. and Oja, E. A fast fixed-point algorithm for independent component analysis. *Neural Comput.*, 9(7):1483–1492, 1999.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Krizhevsky, A. Learning multiple layers of features from tiny images. Master’s thesis, Department of Computer Science, University of Toronto, 2009.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

Kullback, S. and Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. An empirical evaluation of deep architectures on problems with many factors of variation. In Ghahramani, Z. (ed.), *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML’07)*, pp. 473–480. ACM, New York, NY, USA, 2007. doi: 10.1145/1273496.1273556.

LeCun, Y., Huang, F.-J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Press, 2004.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. Deeply-Supervised Nets. *ArXiv e-prints*, 2014.

Lin, Min, Chen, Qiang, and Yan, Shuicheng. Network in network. *CoRR*, abs/1312.4400, 2013. URL [http://arxiv.org/abs/1312.4400](http://arxiv.org/abs/1312.4400).

Nair, V. and Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In Fünkranz, J. and Joachims, T. (eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814. Omnipress 2010, ISBN 978-1-60558-907-7, 2010.

Neal, R. and Hinton, G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I. (ed.), *Learning in Graphical Models*, pp. 355–368. MIT Press, Cambridge, MA, 1998.

Olshausen, B. A. and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Palmer, J., Wipf, D., Kreutz-Delgado, K., and Rao, B. Variational EM algorithms for non-Gaussian latent variable models. In *Advances in Neural Information Processing Systems*, volume 18, pp. 1059–1066, 2006.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., and Olshausen, B. A. Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20(10):2526–2563, 2008. doi: 10.1162/neco.2008.03-07-486.

Sutskever, I., Vinyals, O., and Le, Q. V. V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3104–3112. Curran Associates, Inc., 2014. URL [http://papers.nips.cc/paper/5346](http://papers.nips.cc/paper/5346).

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010.

Warde-Farley, D., Goodfellow, I. J., Courville, A., and Bengio, Y. An empirical analysis of dropout in piecewise linear networks. *ArXiv e-prints*, 2013.

Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q. V., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., and Hinton, G. E. On rectified linear units for speech processing. In *38th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3517–3521. IEEE, 2013.
Supplementary Information

S1 Introduction 12
S2 Hyperparameters Selected for Method Assessment 12
S3 Dataset I 13
S4 Dataset II 17
S5 Generalized Expectation Maximization: Gradient Based M-Step 21
  S5.1 Gradient Ascent .......................................................... 21
  S5.2 Newton Update ......................................................... 22
S6 Posterior Constraint vs. Fréchet Distance 24
  S6.1 Posterior Constraint Method ................................. 24
  S6.2 Fréchet Distance .................................................. 26
S7 Posterior for Gaussian Shift Prior 27
S8 RFN Fixed Point 28
S9 RFN Objective 29
S10 Maximum Likelihood Factor Analysis 36
S1. Introduction

This supplement contains additional information complementing the main manuscript and is structured as follows: First, additional information on the selected hyperparameters of the benchmark methods is given in Section S2. The sections S3 and S4 describe the data generation of the benchmark datasets and report the results for three different experimental settings, namely for extracting 50 (undercomplete), 100 (complete) or 150 (overcomplete) factors / hidden units. Next, RFN’s modified EM algorithm via gradient descent step and via Newton step is reported in the Subsection S5.1 and Subsection S5.2 respectively. In Subsection S6.1 we show, that minimizing the Kullback-Leibler distance in the posterior constraint method would lead to a quadratic optimization procedure for each E-step and therefore minimize the Fréchet distance, which is describes in the Subsection S6.2. Next, Section S7 describes RFN learning as a variational EM algorithm. In Section S8 we derive a fixed point of the algorithm which is independent of the posterior modification. Next, we describe the RFN objective in Section S9. Finally, Section S10 describes the maximum likelihood factor analysis model and the model selection by the EM-algorithm.

S2. Hyperparameters Selected for Method Assessment

In this section, we report the selected hyperparameters, that were used to assess the performance of the compared methods on the benchmark datasets in sections S3 and S4. For all methods, we used the default setting as given by the original publications or the software, so far they were applicable. The number of iterations were chosen to ensure that the algorithms converged. Convergence was estimated on additional instances of the datasets. The Laplace weight decay rate for SFA was chosen to be small in order to ensure low reconstruction error. Tab. S5 lists the selected hyperparameters. The performance of RFNs as unsupervised methods for data representation was compared to:

(1) RFN: rectified factor networks,
(2) RFNn: RFNs without normalization,
(3) DAE: denoising autoencoders with rectified linear units,
(4) RBM: restricted Boltzmann machines with Gaussian visible units and hidden binary units,
(5) FAsp: factor analysis with Jeffrey’s prior \( p(z) \propto \frac{1}{z} \) on the hidden units which is sparser than a Laplace prior,
(6) FAlap: factor analysis with Laplace prior on the hidden units,
(7) ICA: independent component analysis by FastICA [Hyvärinen & Oja 1999],
(8) SFA: sparse factor analysis with a Laplace prior on the parameters,
(9) FA: standard factor analysis,
(10) PCA: principal factor analysis.

Table S5. Hyperparameters of all methods that were used to assess the performance of rectified factor networks (RFNs) as unsupervised methods for data representation. We choose the default parameters of the methods as given by the original publications or the software. The number of iterations were chosen to ensure that the algorithms converged. Convergence was estimated on additional instances of the datasets. The Laplace weight decay rate for SFA was chosen to be small in order to ensure low reconstruction error.

| Method | Used hyperparameters |
|--------|----------------------|
| RFN    | \{learning rate=0.1, iterations=1000\} |
| RFNn   | \{learning rate=0.1, iterations=1000\} |
| DAE    | \{corruption level=0.2, learning rate=1e-04, iterations=1000\} |
| RBM    | \{learning rate=0.01, iterations=1000\} |
| FAsp   | \{iterations=500\} |
| FAlap  | \{iterations=500\} |
| SFA    | \{Laplace weight decay factor=5e-05, iterations=500\} |
S3. Dataset I

This section describes how the benchmark datasets were generated and reports the results of the evaluation for three different experimental settings. Namely, for extracting 50 (undercomplete), 100 (complete) or 150 (overcomplete) factors / hidden units.

We generated nine different benchmark datasets (D1 to D9), where each dataset consists of 100 instances for averaging the results. Each instance consists of 100 samples and 100 features resulting in a $100 \times 100$ data matrix. Into these data matrices, structures are implanted as biclusters [Hochreiter et al., 2010]. A bicluster is a pattern consisting of a particular number of features which is found in a particular number of samples. The size of a bicluster is given by the number of features that form the pattern and by the number of samples in which the pattern is found. The datasets had different noise levels and different bicluster sizes. We considered large and small bicluster sizes, where large biclusters have 20–30 samples and 20–30 features, while small biclusters have 3–8 samples and 3–8 features. The signal strength (scaling factor) of a pattern in a sample was randomly chosen according to the Gaussian $\mathcal{N}(1, 1)$. Finally, to each data matrix background noise was added, where the noise is distributed according to a zero-mean Gaussian with standard deviation 1, 5, or 10. The datasets are described in Tab. I. Features and samples that do not belong to a bicluster have a random bicluster membership drawn from $\mathcal{N}(0, 0.5)$.

The tables S6, S7, and S8 show the results on the benchmark datasets I for 50 (undercomplete), 100 (complete) and 150 (overcomplete) extracted factors / hidden units, respectively. The factors / hidden units were extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machines (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. For the overcomplete case with 150 components, the number of components are restricted to 100 for ICA and PCA. The assessment criteria are: sparseness of the factors (SP) reported in percent (%), reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively.

The results are given in tables S6, S7, and S8 by the mean over 100 instances for each dataset. RFNs yield the most sparse codes with low reconstruction errors and a good approximations of the covariance matrices. We have shown on this benchmark dataset I that RFNs construct very sparse codes of the input with low reconstruction errors that explain the covariance structure of the input. With more factors RFN, like DAEs and RBMs, yield lower reconstruction errors and sparser codes.
Table S6. Comparison for 50 factors / hidden units extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. Criteria are: sparseness of the factors (SP) reported in percent (%) of the samples the factor/component is smaller than 0.01, reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean of 100 instances for each dataset together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively. For 50 factors, RFNs yield the most sparse codes with low reconstruction errors and a good approximations of the covariance matrices.

|       | D1     | D2     | D3     | D4     | D5     |
|-------|--------|--------|--------|--------|--------|
|       | SP     | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     |
| RFN   | 72±1   | 74±2   | 11±1   | 75±0   | 240±3  | 72±2   | 75±0   | 462±5  | 260±6  |
| RFNn  | 68±1   | 122±5  | 32±4   | 74±0   | 285±4  | 97±3   | 74±0   | 537±7  | 331±8  |
| DAE   | 61±0   | 82±2   | —      | 66±0   | 243±2  | —      | 66±0   | 461±4  | —      |
| RBM   | 22±1   | 106±3  | —      | 11±1   | 301±3  | —      | 10±0   | 566±6  | —      |
| FAsp  | 37±1   | 469±38 | 999±99 | 40±1   | 999±99 | 999±99 | 40±2   | 999±99 | 999±99 |
| FAapl | 4±0    | 50±1   | 392±66 | 4±0    | 228±5  | 135±13 | 5±0    | 443±9  | 406±18 |
| ICA   | 16±1   | 35±0   | —      | 16±0   | 168±2  | —      | 16±0   | 327±4  | —      |
| SFA   | 1±0    | 42±1   | 26±3   | 1±0    | 210±5  | 61±2   | 1±0    | 409±8  | 220±6  |
| FA    | 1±0    | 42±1   | 13±2   | 1±0    | 210±4  | 58±2   | 1±0    | 409±8  | 214±6  |
| PCA   | 0±0    | 35±0   | —      | 0±0    | 168±2  | —      | 0±0    | 327±4  | —      |
|       | D6     | D7     | D8     | D9     | average|
|       | SP     | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     |
| RFN   | 75±0   | 464±6  | 264±6  | 73±0   | 68±2   | 9±1    | 75±0   | 237±3  | 69±1   |
| RFNn  | 74±0   | 541±6  | 336±8  | 71±1   | 106±4  | 23±3   | 74±0   | 279±3  | 91±2   |
| DAE   | 66±0   | 465±4  | —      | 62±0   | 75±2   | —      | 66±0   | 238±2  | —      |
| RBM   | 10±1   | 570±6  | —      | 20±1   | 97±3   | —      | 11±0   | 294±3  | —      |
| FAsp  | 41±1   | 999±99 | 999±99 | 38±1   | 335±32 | 999±99 | 41±1   | 999±99 | 999±99 |
| FAapl | 5±0    | 447±9  | 413±19 | 4±0    | 49±1   | 292±57 | 4±0    | 227±5  | 123±11 |
| ICA   | 16±1   | 329±4  | —      | 16±1   | 35±0   | —      | 16±1   | 167±2  | —      |
| SFA   | 1±0    | 412±8  | 223±7  | 1±0    | 42±1   | 19±3   | 1±0    | 209±4  | 59±2   |
| FA    | 1±0    | 412±8  | 217±7  | 1±0    | 42±1   | 10±1   | 1±0    | 409±4  | 57±2   |
| PCA   | 0±0    | 329±4  | —      | 0±0    | 35±0   | —      | 0±0    | 167±2  | —      |
Table S7. Comparison for 100 factors / hidden units extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. Criteria are: sparseness of the factors (SP) reported in percent (%) of the samples the factor/component is smaller than 0.01, reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean of 100 instances for each dataset together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively. For 100 factors, RFNs yield the most sparse codes with low reconstruction errors and a good approximations of the covariance matrices.

|       | D1       | D2       | D3       | D4       | D5       |
|-------|----------|----------|----------|----------|----------|
|       | SP       | ER       | CO       | SP       | ER       | CO       | SP       | ER       | CO       | SP       | ER       | CO       |
| RFN   | 76±1     | 34±3     | 4±1      | 82±1     | 67±8     | 18±3     | 82±1     | 124±16   | 63±12    | 75±1     | 38±3     | 5±1      |
| RFNn  | 71±1     | 110±27   | 25±4     | 79±0     | 180±5    | 42±2     | 80±0     | 331±8    | 139±7    | 65±2     | 143±9    | 47±8     |
| DAE   | 63±0     | 66±2     |         | 70±0     | 142±2    |         | 70±0     | 264±3    |         | 62±0     | 73±2     |         |
| RBM   | 12±1     | 100±3    |         | 5±0      | 282±4    |         | 4±0      | 522±6    |         | 12±1     | 106±3    |         |
| FAsp  | 71±0     | 474±38   | 999±99   | 62±0     | 999±53   | 999±99   | 56±1     | 999±99   | 999±99   | 70±0     | 616±44   | 999±99   |
| FAlap | 6±0      | 21±2     | 425±28   | 6±0      | 40±2     | 827±35   | 6±0      | 75±6     | 999±99   | 6±0      | 23±2     | 523±32   |
| ICA   | 43±9     | 0±0      |         | 32±7     | 0±0      |         | 29±8     | 0±0      |         | 42±8     | 0±0      |         |
| SFA   | 1±0      | 10±0     | 71±7     | 1±0      | 15±0     | 84±4     | 1±0      | 28±1     | 254±8    | 1±0      | 12±0     | 87±8     |
| FA    | 1±0      | 10±0     | 48±5     | 1±0      | 15±0     | 59±3     | 1±0      | 28±1     | 189±7    | 1±0      | 12±1     | 61±6     |
| PCA   | 4±0      | 0±0      |         | 2±0      | 0±0      |         | 1±0      | 0±0      |         | 3±0      | 0±0      |         |

|       | D6       | D7       | D8       | D9       | average  |
|-------|----------|----------|----------|----------|----------|
|       | SP       | ER       | CO       | SP       | ER       | CO       | SP       | ER       | CO       | SP       | ER       | CO       |
| RFN   | 82±1     | 127±17   | 65±14    | 77±1     | 30±3     | 3±1      | 82±1     | 64±8     | 17±4     | 82±1     | 123±15   | 62±13    |
| RFNn  | 80±0     | 334±8    | 141±7    | 74±1     | 86±4     | 14±2     | 79±0     | 174±4    | 39±2     | 80±0     | 329±7    | 137±6    |
| DAE   | 70±0     | 266±2    |         | 64±0     | 57±2     |         | 70±0     | 138±1    |         | 70±0     | 262±2    |         |
| RBM   | 4±0      | 527±6    |         | 11±1     | 92±2     |         | 4±0      | 274±4    |         | 4±0      | 518±6    |         |
| FAsp  | 56±0     | 999±99   | 999±99   | 71±0     | 338±33   | 999±99   | 62±1     | 999±42   | 999±99   | 56±1     | 999±99   | 999±99   |
| FAlap | 6±0      | 75±6     | 999±89   | 6±0      | 18±2     | 337±24   | 6±0      | 40±3     | 793±37   | 6±0      | 74±6     | 999±89   |
| ICA   | 30±7     | 0±0      |         | 44±8     | 0±0      |         | 33±7     | 0±0      |         | 30±7     | 0±0      |         |
| SFA   | 1±0      | 28±1     | 260±9    | 1±0      | 8±0      | 52±7     | 1±0      | 15±0     | 76±3     | 1±0      | 28±1     | 248±7    |
| FA    | 1±0      | 28±1     | 193±8    | 1±0      | 8±0      | 33±5     | 1±0      | 15±0     | 54±2     | 1±0      | 28±1     | 185±6    |
| PCA   | 1±0      | 0±0      |         | 4±0      | 0±0      |         | 2±0      | 0±0      |         | 1±0      | 0±0      |         |
Table S8. Comparison for 150 factors / hidden units extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. For the overcomplete case with 150 components, the number of components are restricted to 100 for ICA and PCA. Criteria are: sparseness of the factors (SP) reported in percent (%) of the samples the factor/component is smaller than 0.01, reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean of 100 instances for each dataset together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively. For 150 factors, RFNs yield the most sparse codes with low reconstruction errors and a good approximations of the covariance matrices.

|       | D1     | D2     | D3     | D4     | D5     |
|-------|--------|--------|--------|--------|--------|
| SP    | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     |
| RFN   | 81±4   | 12±2   | 1±1    | 86±0   | 16±1   | 4±1    | 86±0   | 29±4   | 15±5   | 80±1   | 15±5   | 2±2    | 86±1   | 17±5   | 5±3    |
| RFNn  | 72±1   | 100±8  | 19±4   | 80±0   | 137±4  | 24±1   | 81±0   | 254±6  | 83±4   | 66±0   | 113±3  | 52±5   | 80±0   | 141±4  | 26±2   |
| DAE   | 64±0   | 62±2   | —      | 71±0   | 125±2  | —      | 72±0   | 232±2  | —      | 63±0   | 69±2   | —      | 71±0   | 129±2  | —      |
| RBM   | 8±0    | 101±3  | —      | 4±0    | 284±8  | —      | 3±0    | 521±6  | —      | 8±0    | 106±3  | —      | 4±0    | 289±4  | —      |
| FAsp  | 81±1   | 474±38 | 999±99 | 79±0   | 999±53 | 999±99 | 77±1   | 999±49 | 999±99 | 80±1   | 616±44 | 999±99 | 79±1   | 999±60 | 999±99 |
| FAap  | 4±0    | 21±2   | 607±34 | 4±0    | 40±2   | 879±40 | 3±0    | 75±6   | 999±96 | 4±0    | 23±2   | 749±42 | 4±0    | 42±3   | 926±45 |
| ICA   | 43±9   | 0±0    | —      | 32±7   | 0±0    | —      | 29±8   | 0±0    | —      | 42±8   | 0±0    | —      | 33±5   | 0±0    | —      |
| SFA   | 1±0    | 10±0   | 103±9  | 1±0    | 15±0   | 204±7  | 1±0    | 28±1   | 656±12 | 1±0    | 12±0   | 126±10 | 1±0    | 16±0   | 220±8  |
| FA    | 1±0    | 10±0   | 87±8   | 1±0    | 15±0   | 187±5  | 1±0    | 28±1   | 611±11 | 1±0    | 12±1   | 108±9  | 1±0    | 16±0   | 200±6  |
| PCA   | 4±0    | 0±0    | —      | 2±0    | 0±0    | —      | 1±0    | 0±0    | —      | 3±0    | 0±0    | —      | 2±0    | 0±0    | —      |
|       | D6     | D7     | D8     | D9     | average |
| SP    | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     | SP     | ER     | CO     |
| RFN   | 86±1   | 29±7   | 15±6   | 82±1   | 10±3   | 1±1    | 86±1   | 17±10  | 5±9    | 86±1   | 31±19  | 16±13  | 84±1   | 20±6   | 7±4    |
| RFNn  | 81±0   | 255±6  | 84±3   | 76±1   | 74±5   | 9±2    | 81±0   | 133±3  | 23±1   | 81±0   | 250±7  | 81±4   | 77±0   | 162±5  | 45±3   |
| DAE   | 72±0   | 234±2  | —      | 65±0   | 53±2   | —      | 72±0   | 122±1  | —      | 72±0   | 230±2  | —      | 69±0   | 140±2  | —      |
| RBM   | 3±0    | 525±6  | —      | 8±0    | 93±3   | —      | 3±0    | 273±4  | —      | 3±0    | 517±6  | —      | 5±0    | 301±4  | —      |
| FAsp  | 77±1   | 999±99 | 999±99 | 81±1   | 338±31 | 673±99 | 79±0   | 999±42 | 999±99 | 77±1   | 999±99 | 999±99 | 79±1   | 999±74 | 999±99 |
| FAap  | 3±0    | 75±6   | 999±94 | 4±0    | 18±2   | 479±31 | 4±0    | 40±3   | 831±43 | 3±0    | 74±6   | 999±95 | 4±0    | 45±3   | 999±58 |
| ICA   | 30±7   | 0±0    | —      | 44±8   | 0±0    | —      | 33±7   | 0±0    | —      | 30±7   | 0±0    | —      | 35±7   | 0±0    | —      |
| SFA   | 1±0    | 28±1   | 668±12 | 1±0    | 8±0    | 78±8   | 1±0    | 15±0   | 188±5  | 1±0    | 28±1   | 644±9  | 1±0    | 18±1   | 321±9  |
| FA    | 1±0    | 28±1   | 622±11 | 1±0    | 8±0    | 64±7   | 1±0    | 15±0   | 173±4  | 1±0    | 28±1   | 599±9  | 1±0    | 18±1   | 294±8  |
| PCA   | 1±0    | 0±0    | —      | 4±0    | 0±0    | —      | 2±0    | 0±0    | —      | 1±0    | 0±0    | —      | 2±0    | 0±0    | —      |
S4. Dataset II

This dataset was generated as described in Section S3, but now features and samples that do not belong to a bicluster have a random bicluster membership drawn from $\mathcal{N}(0, 0.01)$.

The tables S9, S10, and S11 show the results on the benchmark datasets II for 50 (undercomplete), 100 (complete) and 150 (overcomplete) extracted factors / hidden units, respectively. The factors / hidden units were extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. The assessment criteria are: sparseness of the factors (SP) reported in percent (%) of the samples the factor/component is smaller than 0.01, reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean of 100 instances for each dataset together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively.

The results for benchmark dataset II are given in tables S9, S10, and S11 by the mean over 100 instances for each dataset. Also with lower noise level, RFNs yield the most sparse codes with low reconstruction errors and a good approximation of the covariance matrices. We have shown on this benchmark dataset I that RFNs construct very sparse codes of the input with low reconstruction errors that explain the covariance structure of the input. With more factors RFN, like DAEs and RBMs, yield lower reconstruction errors and sparser codes. These results confirmed the results from dataset I, where the noise level was higher.
Table S9. Comparison for 50 factors / hidden units extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. Criteria are: sparseness of the factors (SP) reported in percent (%), reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively. For 50 factors, RFNs yield the most sparse codes with low reconstruction errors and a good approximations of the covariance matrices.
Table S10. Comparison for 100 factors / hidden units extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. Criteria are: sparseness of the factors (SP) reported in percent (%) of the samples the factor/component is smaller than 0.01, reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. Results reported here, are the mean of 100 instances for each dataset together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively. For 100 factors, RFNs yield the most sparse codes with low reconstruction errors and a good approximation of the covariance matrices.

|        | D1          | D2          | D3          | D4          | D5          |
|--------|-------------|-------------|-------------|-------------|-------------|
|        | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO |
| RFN    | 79±1 | 23±3 | 2±0 | 82±1 | 63±9 | 16±3 | 82±1 | 120±17 | 61±15 | 78±1 | 27±3 | 2±1 | 82±1 | 62±7 | 16±3 |
| RFNn   | 77±0 | 61±4 | 6±1 | 80±0 | 169±4 | 36±2 | 80±0 | 326±8 | 135±6 | 76±1 | 73±4 | 9±2 | 79±0 | 171±5 | 37±2 |
| DAE    | 67±0 | 48±2 — | — | 70±0 | 134±1 — | — | 70±0 | 260±2 — | — | 67±0 | 54±2 — | — | 70±0 | 137±1 — | — |
| RBM    | 14±1 | 81±3 — | — | 4±0 | 266±3 — | — | 4±0 | 514±6 — | — | 15±1 | 88±2 — | — | 4±0 | 270±3 — | — |
| FAsp   | 72±0 | 233±32 | 499±99 | 62±0 | 999±4 | 999±99 | 56±0 | 999±99 | 999±99 | 71±0 | 320±34 | 878±99 | 62±0 | 999±49 | 999±99 |
| FAlap  | 6±0 | 27±3 | 202±17 | 6±0 | 38±3 | 756±33 | 6±0 | 74±5 | 999±83 | 6±0 | 31±3 | 274±23 | 6±0 | 39±3 | 778±34 |
| ICA    | 42±11 | 0±0 — | — | 31±8 | 0±0 — | — | 29±7 | 0±0 — | — | 40±7 | 0±0 — | — | 31±7 | 0±0 — | — |
| SFA    | 1±0 | 6±0 | 30±5 | 1±0 | 14±0 | 68±3 | 1±0 | 28±1 | 243±8 | 1±0 | 8±0 | 38±5 | 1±0 | 15±0 | 72±3 |
| FA     | 1±0 | 6±0 | 18±3 | 1±0 | 14±0 | 50±2 | 1±0 | 28±1 | 182±7 | 1±0 | 8±0 | 24±4 | 1±0 | 15±0 | 52±2 |
| PCA    | 4±0 | 0±0 — | — | 2±0 | 0±0 — | — | 1±0 | 0±0 — | — | 4±0 | 0±0 — | — | 2±0 | 0±0 — | — |

|        | D6          | D7          | D8          | D9          | average |
|--------|-------------|-------------|-------------|-------------|---------|
|        | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO | SP | ER | CO |
| RFN    | 82±1 | 120±16 | 60±13 | 80±1 | 18±2 | 1±0 | 82±1 | 61±7 | 15±3 | 82±1 | 122±13 | 60±11 | 81±1 | 68±9 | 26±6 |
| RFNn   | 80±0 | 329±7 | 137±6 | 78±0 | 49±3 | 4±1 | 80±0 | 165±4 | 34±1 | 80±0 | 325±7 | 134±6 | 79±0 | 185±5 | 59±3 |
| DAE    | 70±0 | 261±2 — | — | 68±0 | 39±2 — | — | 70±0 | 132±1 — | — | 70±0 | 259±2 — | — | 69±0 | 147±2 — | — |
| RBM    | 4±0 | 517±6 — | — | 12±1 | 71±2 — | — | 4±0 | 261±3 — | — | 4±0 | 512±5 — | — | 7±1 | 287±4 — | — |
| FAsp   | 56±1 | 999±99 | 999±99 | 73±0 | 149±28 | 237±62 | 62±0 | 999±34 | 999±99 | 56±0 | 999±99 | 999±99 | 63±0 | 999±65 | 999±99 |
| FAlap  | 6±0 | 74±6 | 999±91 | 6±0 | 22±3 | 134±14 | 6±0 | 37±2 | 733±28 | 6±0 | 73±6 | 999±84 | 6±0 | 46±4 | 985±45 |
| ICA    | 29±6 | 0±0 — | — | 40±9 | 0±0 — | — | 29±8 | 0±0 — | — | 28±7 | 0±0 — | — | 33±8 | 0±0 — | — |
| SFA    | 1±0 | 28±1 | 247±8 | 1±0 | 5±0 | 21±5 | 1±0 | 14±0 | 64±2 | 1±0 | 27±1 | 240±7 | 1±0 | 16±1 | 114±5 |
| FA     | 1±0 | 28±1 | 184±8 | 1±0 | 5±0 | 11±3 | 1±0 | 14±0 | 47±2 | 1±0 | 27±1 | 179±7 | 1±0 | 16±1 | 83±4 |
| PCA    | 1±0 | 0±0 — | — | 4±0 | 0±0 — | — | 2±0 | 0±0 — | — | 1±0 | 0±0 — | — | 2±0 | 0±0 — | — |
### Table S11. Comparison for 150 factors / hidden units extracted by RFN, RFN without normalization (RFNn), denoising autoencoder (DAE), restricted Boltzmann machine (RBM), factor analysis with a very sparse prior (FAsp), factor analysis with a Laplace prior (FAlap), independent component analysis (ICA), sparse factor analysis (SFA), factor analysis (FA), and principal component analysis (PCA) on nine datasets. For the overcomplete case with 150 components, the number of components are restricted to 100 for ICA and PCA. Criteria are: sparseness of the factors (SP) reported in percent (%) of the samples the factor/component is smaller than 0.01, reconstruction error (ER), and the difference between the empirical and the model covariance matrix (CO). The lower right column block gives the average SP (%), ER and CO. The lower right column block gives the average SP, ER and CO. Results reported here, are the mean of 100 instances for each dataset together with the standard deviation of 100 instances. The maximal value in the table and the maximal standard deviation was set to 999 and to 99, respectively. For 150 factors, RFNs yield the most sparse codes with low reconstruction errors and a good approximations of the covariance matrices.

|     | D1          | D2          | D3          | D4          | D5          |
|-----|-------------|-------------|-------------|-------------|-------------|
|     | SP          | ER          | CO          | SP          | ER          | CO          | SP          | ER          | CO          |
| RFN | 83 ±1       | 7 ±2        | 0 ±1        | 86 ±1       | 33 ±20      | 18 ±23      | 83 ±1       | 9 ±2        | 1 ±0        |
| RFNn| 79 ±0       | 48 ±3       | 4 ±1        | 81 ±0       | 129 ±3      | 21 ±1       | 81 ±0       | 250 ±7      | 80 ±4       |
| DAE | 68 ±0       | 44 ±2       | —           | 72 ±0       | 118 ±1      | —           | 72 ±0       | 229 ±2      | —           |
| RBM | 10 ±1       | 81 ±3       | —           | 3 ±0        | 265 ±3      | —           | 3 ±0        | 514 ±6      | —           |
| FAsp| 83 ±1       | 233 ±32     | 340 ±71     | 79 ±0       | 999 ±40     | 999 ±99     | 77 ±0       | 999 ±99     | 999 ±99     |
| FAlap| 4 ±0       | 27 ±3       | 295 ±25     | 4 ±0        | 38 ±3       | 791 ±41     | 3 ±0        | 74 ±5       | 999 ±91     |
| ICA | 42 ±11      | 0 ±0        | —           | 31 ±8       | 0 ±0        | —           | 29 ±7       | 0 ±0        | —           |
| SFA | 1 ±0        | 6 ±0        | 49 ±7       | 1 ±0        | 14 ±0       | 173 ±4      | 1 ±0        | 28 ±1       | 632 ±10     |
| FA  | 1 ±0        | 6 ±0        | 40 ±5       | 1 ±0        | 14 ±0       | 160 ±4      | 1 ±0        | 28 ±1       | 590 ±10     |
| PCA | 4 ±0        | 0 ±0        | —           | 2 ±0        | 0 ±0        | —           | 1 ±0        | 0 ±0        | —           |
|     | D6          | D7          | D8          | D9          | average     |
|     | SP          | ER          | CO          | SP          | ER          | CO          | SP          | ER          | CO          |
| RFN | 86 ±1       | 30 ±13      | 15 ±16      | 84 ±2       | 5 ±3        | 0 ±1        | 86 ±1       | 14 ±1       | 3 ±1        |
| RFNn| 81 ±0       | 251 ±6      | 81 ±3       | 80 ±0       | 37 ±3       | 2 ±0        | 81 ±0       | 126 ±3      | 20 ±1       |
| DAE | 72 ±0       | 230 ±2      | —           | 70 ±0       | 36 ±2       | —           | 72 ±0       | 116 ±1      | —           |
| RBM | 3 ±0        | 516 ±6      | —           | 8 ±1        | 71 ±2       | —           | 3 ±0        | 260 ±4      | —           |
| FAsp| 77 ±0       | 999 ±99     | 999 ±99     | 84 ±0       | 149 ±28     | 168 ±55     | 80 ±0       | 999 ±99     | 999 ±99     |
| FAlap| 3 ±0       | 74 ±6       | 999 ±97     | 4 ±0        | 22 ±3       | 198 ±17     | 4 ±0        | 37 ±2       | 768 ±40     |
| ICA | 29 ±6       | 0 ±0        | —           | 40 ±9       | 0 ±0        | —           | 29 ±8       | 0 ±0        | —           |
| SFA | 1 ±0        | 28 ±1       | 640 ±11     | 1 ±0        | 5 ±0        | 34 ±6       | 1 ±0        | 14 ±0       | 164 ±3      |
| FA  | 1 ±0        | 28 ±1       | 596 ±10     | 1 ±0        | 5 ±0        | 27 ±5       | 1 ±0        | 14 ±0       | 153 ±3      |
| PCA | 1 ±0        | 0 ±0        | —           | 4 ±0        | 0 ±0        | —           | 2 ±0        | 0 ±0        | —           |

**Rectified Factor Networks (RFN):**
- **DAE:** Denoising Autoencoder
- **RBM:** Restricted Boltzmann Machine
- **ICA:** Independent Component Analysis
- **SFA:** Sparse Factor Analysis
- **FA:** Factor Analysis
- **PCA:** Principal Component Analysis
S5. Generalized Expectation Maximization: Gradient Based M-Step

We modify the EM algorithm via gradient descent step in the M-step instead of computing the maximum. This approach is called Generalized Expectation Maximization (GEM) and is typically used if maximization cannot be realized but an increase of the lower bound on the likelihood can be achieved. We require gradient ascent to allow for stochastic gradients because of dropout and mini batches as well as for efficient implementations on GPUs.

S5.1. Gradient Ascent

The gradients in the M-step are:

\[
\frac{2}{n} \nabla_W \log L = \frac{1}{n} \sum_{i=1}^{n} \psi_i^{-1} v_i E_{h_i|v_i} (h_i) - \frac{1}{n} \sum_{i=1}^{n} \psi_i^{-1} W E_{h_i|v_i} (h_i h_i^T)
\]

and

\[
\nabla_\psi \log L = -\frac{n}{2} \psi^{-1} + \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}_{h_i|v_i} \left[ (\psi_i^{-1} (v_i - Wh_i) (v_i - Wh_i)^T \psi_i^{-1} ) \right].
\]

This gives the gradients:

\[
\frac{2}{n} \nabla_W \log L = \psi_i^{-1} \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) - \psi_i^{-1} W \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T)
\]

and

\[
\frac{2}{n} \nabla_\psi \log L = \text{diag} \left( \left[ -I + \psi_i^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T - \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) W^T - \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i) v_i + W \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) W^T \right) \right] \psi_i^{-1} \right).
\]

which is the diagonal of

\[
\left( -I + \psi_i^{-1} \left( \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T - \frac{2}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) W^T + W \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) W^T \right) \right) \psi_i^{-1}.
\]

For using subterms of previous gradient \( \nabla_W \log L \), this can be written as

\[
\left( -I + \psi_i^{-1} \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T + \left[ -\frac{2}{n} \psi_i^{-1} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) + \psi_i^{-1} W \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) W^T \right] \right) \psi_i^{-1}.
\]

The first term enters with a factor of 2 in contrast to the gradient \( \nabla_W \log L \). Therefore only the sums

\[
\sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i)
\]

and

\[
\sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T)
\]

must be computed for both gradients.
\[ \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T \]  

(22)

is the estimated covariance matrix (matrix of second moments for zero mean).

The generalized EM algorithm update rules are:

\[ C = \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T \]  

(23)

\[ \text{E-step:} \]

\[ \mu_{h_i|v_i} = W^T (W W^T + \Psi)^{-1} v_i = (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} v_i, \]

\[ \Sigma_{h_i|v_i} = I - W^T (W W^T + \Psi)^{-1} W = (I + W^T \Psi^{-1} W)^{-1}, \]

\[ E_{h_i|v_i}(h_i) = \mu_{h_i|v_i} \]

\[ E_{h_i|v_i}(h_i, h_i^T) = \mu_{h_i|v_i} \mu_{h_i|v_i}^T + \Sigma_{h_i|v_i} \]

\[ U = \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i}(h_i) \]

\[ S = \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i}(h_i, h_i^T) \]

\[ \text{M-step:} \]

\[ \Delta W = \Psi^{-1} U - \Psi^{-1} W S \]

\[ \Delta \Psi = \text{diag} \left( (-I + (\Psi^{-1} C + (-2 \Psi^{-1} U + \Psi^{-1} W S) W^T) \Psi^{-1}) \right). \]

**S5.2. Newton Update**

Instead of gradient ascent, we now consider a Newton update step. The gradients are

\[ \frac{2}{n} \nabla_W \log \mathcal{L} = \Psi^{-1} \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i}^T(h_i) - \Psi^{-1} W \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i}(h_i, h_i^T) \]  

(25)

and

\[ \frac{2}{n} \nabla_{\Psi} \log \mathcal{L} = \]

(26)
The EM update of $\Psi$ is

$$\Psi = \text{diag} \left( \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} \left( (v_i - W h_i) (v_i - W h_i)^T \right) \right)$$

$$= \text{diag} \left( \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T - \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) W^T - \frac{1}{n} \sum_{i=1}^{n} W E_{h_i|v_i} (h_i) v_i^T + W \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i \ h_i^T) W^T \right)$$

$$= \text{diag} \left( C - U W^T - U W^T + W S W^T \right),$$

where $c_i$ is computed using the optimal $W$ from the $W$-update. The Newton update for finding the roots of $\frac{\partial f}{\partial v}$ is

$$v_{n+1} = v_n - \eta H^{-1} \nabla_v f(v_n),$$

where $\eta$ is a small step size and $H$ is the Hessian of $f$ with respect to $v$ evaluated at $v_n$. We denote the update direction by

$$\Delta v = - H^{-1} \nabla_v f(v_n).$$

**Update of $W$.** The Hessian $H_W$ of $\left( \frac{1}{n} \log \mathcal{L} \right)$ with respect to $W$ as a vector is:

$$H_W = \frac{\partial \text{vec} \left( \frac{1}{n} \nabla_W \log \mathcal{L} \right)}{\partial \text{vec}(W)^T} = \frac{\partial \text{vec} \left( \Psi^{-1} U - \Psi^{-1} W S \right)}{\partial \text{vec}(W)^T} = - S \otimes \Psi^{-1},$$

where $\otimes$ is the Kronecker product of matrices and vec makes a vector out of a matrix. The inverse of $H_W$ is

$$H_W^{-1} = - S^{-1} \otimes \Psi.$$

For the product of the negative inverse Hessian with the gradient we have:

$$- H_W^{-1} \text{vec} \left( \Psi^{-1} U - \Psi^{-1} W S \right) = \text{vec} \left( \Psi \left( \Psi^{-1} U - \Psi^{-1} W S \right) S^{-1} \right) = \text{vec} \left( U S^{-1} - W \right).$$

Thus, if we apply a Newton update then the update direction for $W$ in the M-step is

$$\Delta W = U S^{-1} - W.$$  

This is the exact EM update if stepsize is 1. Since the objective is a quadratic function in $W$, one Newton update would lead to the exact solution. However for rectified linear posteriors and for dropout we want to have small changes of the matrix $W$. For the Newton update we require $S^{-1}$. $S$ was defined as

$$S = \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i \ h_i^T) = \frac{1}{n} \sum_{i=1}^{n} \mu_{h_i|v_i} \mu_{h_i|v_i}^T + \Sigma_{h|v}.$$  

For rectified linear and for dropout the posterior mean $\mu_{h_i|v_i}$ is different for each example. Therefore, the sum $\frac{1}{n} \sum_{i=1}^{n} \mu_{h_i|v_i} \mu_{h_i|v_i}^T$ must be computed which requires $l^2$ update steps for each sample $v_i$. This gives $O(nl^2)$ updates. For computing $S^{-1}$ we require $l^3$ steps. Therefore computing $S^{-1}$ requires $O(l^2(l + n))$ computations.

**Update of $\Psi$.** We restrict everything to the diagonal entries of $\Psi$:

$$\Delta \Psi = \text{diag} \left( - \Psi^{-1} + \Psi^{-1} \left( C + ( - 2 U + W S ) W^T \right) \Psi^{-1} \right).$$
If we define

\[ c_i = [C + (-2 U + WS) W^T]_{ii} \]  

(37)

the update is

\[ \Delta \Psi_{ii} = [\nabla \Psi \log L]_{ii} = -\frac{1}{\Psi_{ii}} + \frac{1}{\Psi_{ii}^2} c_i . \]  

(38)

We exploited that \( \Psi \) is diagonal. The Hessian \( H_{\Psi} \) of \( (\frac{2}{n} \log L) \) with respect to \( \Psi \) as a vector is:

\[ [H_{\Psi}]_{ij} = \begin{cases} 0 & \text{for } i \neq j \\ \frac{1}{\Psi_{ii}^2} - \frac{2}{\Psi_{ii}^3} c_i & \text{for } i = j. \end{cases} \]  

(39)

Therefore the Hessian \( H_{\Psi} \) is a diagonal matrix. Since we maximize the likelihood, we have to ensure that the Hessian is negative definite:

\[ \frac{1}{\Psi_{ii}^2} - \frac{2}{\Psi_{ii}^3} c_i = \frac{1}{\Psi_{ii}^2} \left( 1 - \frac{2 c_i}{\Psi_{ii}} \right) < 0 \]  

(40)

\[ \Psi_{ii} < 2 c_i . \]  

(41)

The Newton update direction for component \( \Psi_{ii} \) is the negative inverse Hessian multiplied by the gradient. The Hessian is diagonal, therefore for the update direction each component of the gradient is divided by the corresponding negative component of the Hessian:

\[ \Delta \Psi_{ii} = -\frac{[\nabla \Psi \log L]_{ii}}{[H_{\Psi}]_{ii}} = \frac{1}{\Psi_{ii}} - \frac{1}{\Psi_{ii}^2} c_i = \frac{\Psi_{ii}^2 - \Psi_{ii} c_i}{\Psi_{ii} - 2 c_i} = \frac{c_i - \Psi_{ii}}{2 c_i - \Psi_{ii}} . \]  

(42)

To ensure \( \Psi_{ii} < 2c_i \) we either set in the denominator of the \( \Psi_{ii} \) update \( \Psi_{ii} = c_i \) and obtain

\[ \Delta \Psi_{ii} = \frac{c_i - \Psi_{ii}}{c_i} \]  

(43)

or, alternatively, we set \( c_i = \Psi_{ii} \) and obtain

\[ \Delta \Psi_{ii} = c_i - \Psi_{ii} . \]  

(44)

The latter is the EM solution minus the actual \( \Psi_{ii} \).

S6. Posterior Constraint vs. Fréchet Distance

We want to minimize the distance between the posterior \( p \) and the variational \( Q \), where \( Q \) is constraint to non-negative mean values (rectifying). To minimize the distance enforces high information content in the posterior means about the input.

S6.1. Posterior Constraint Method

Rectifying can be implemented by the posterior constraint method \( \text{(Ganchev et al., 2010; Graca et al., 2007)} \). However, this approach leads to a constraint quadratic (in the number of factors) optimization problem. If \( Q \) approximates the posterior \( p \), then the posterior constraint method uses the Kulback-Leibler divergence (KL-divergence) \( \text{(Kullback & Leibler, 1951)} \):

\[ d(Q, p) = D_{KL}(Q \parallel p) = \int Q(h) \frac{Q(h)}{p(h \mid \nu)} dh . \]  

(45)

\( Q \) is from a Gaussian family

\[ Q(h) \sim (2\pi)^{-\frac{n}{2}} |\Sigma_q|^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (h - \mu_q)^T \Sigma_q^{-1} (h - \mu_q) \right) . \]  

(46)
and the posterior is Gaussian, too:

\[ p(h \mid v) \sim (2\pi)^{-\frac{d}{2}} |\Sigma_p|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (h - \mu_p)^T \Sigma_p^{-1} (h - \mu_p) \right). \]  

(47)

The Kullback-Leibler divergence between \( Q \) and \( p \) is

\[ D_{KL}(Q \parallel p) = \frac{1}{2} \left\{ \text{Tr} \left( \Sigma_p^{-1} \Sigma_q \right) + (\mu_p - \mu_q)^T \Sigma_p^{-1} (\mu_p - \mu_q) - l - \ln |\Sigma_q|/|\Sigma_p| \right\}. \]  

(48)

We assume that \( Q \) is from the mean rectified family of normal distributions in analogy to the rectified linear units used for neural networks. That means we consider the family of normal distributions which have mean larger equal zero. This leads to the constraints:

\[ \mu_q \geq 0, \]  

(49)

where the inequality is meant per component.

We now minimize Eq. (48) for \( Q \) in this family. Eq. (48) has a quadratic form in \( \mu_q \), where \( \Sigma_q \) does not enter, and terms in \( \Sigma_q \), where \( \mu_q \) does not enter. Therefore we can separately minimize for \( \Sigma_q \) and for \( \mu_q \).

For the minimization with respect to \( \Sigma_q \), we require

\[ \frac{\partial}{\partial \Sigma_q} \text{Tr} \left( \Sigma_p^{-1} \Sigma_q \right) = \Sigma_p^{-T} \]  

(50)

and

\[ \frac{\partial}{\partial \Sigma_q} \ln |\Sigma_q| = \Sigma_q^{-T}. \]  

(51)

For optimality the derivative of the objective \( D_{KL}(Q \parallel p) \) with respect to \( \Sigma_q \) must be zero:

\[ \frac{\partial}{\partial \Sigma_q} D_{KL}(Q \parallel p) = \Sigma_p^{-T} - \Sigma_q^{-T} = 0. \]  

(52)

This gives

\[ \Sigma_q = \Sigma_p. \]  

(53)

Using the second order derivative of the matrices as vectors, we confirm that it is a minimum.

If we use only terms in \( \mu_q \), the minimization with respect to \( \mu_q \) gives the constraint minimization problem:

\[ \min_{\mu_q} \frac{1}{2} \mu_q^T \Sigma_p^{-1} \mu_q - \mu_q^T \Sigma_p^{-1} \mu_q \]  

s.t. \( \mu_q \geq 0. \)  

(54)

If \( \lambda \) is the Lagrange multiplier for the constraints, then the dual is

\[ \min_{\lambda} \frac{1}{2} \lambda^T \Sigma_p \lambda + \mu^T \lambda \]  

s.t. \( \lambda \geq 0. \)  

(55)

The Karush-Kuhn-Tucker conditions require for the optimal solution for each component \( j \):

\[ \lambda_j \mu_{qj} = 0. \]  

(56)

Further the derivative of the Lagrangian with respect to \( \mu_q \) gives

\[ \Sigma_p^{-1} \mu_q - \Sigma_p^{-1} \mu_p - \lambda = 0 \]  

(57)

which can be written as

\[ \mu_q - \mu_p - \Sigma_p \lambda = 0. \]  

(58)

Concluding, the posterior constraint method (Ganchev et al., 2010; Graca et al., 2007) leads to a constraint quadratic (in the number of factors) optimization problem.
S6.2. Fréchet Distance

Since the Kullback-Leibler distance (Kullback & Leibler, 1951) would lead to a quadratic optimization procedure for each E-step, we move to another distance measure for normal distributions. In 1957 Fréchet introduced a distance between multivariate normal distributions $G$ and $H$ (Fréchet, 1957). The Fréchet distance is defined as the minimum over an expectation over all distribution $Z$ which have $G$ and $H$ as marginals. $X$ and $Y$ are random variables that are distributed according to $G$ and $H$ and take values $v$ and $h$, respectively. $Z$ is a distribution on $(v, h)$ which has as marginals

$$Z_1(v) = \int Z(v, h) \, dh = G(v)$$

and

$$Z_2(h) = \int Z(v, h) \, dv = H(h).$$

The expectation under $Z$ of the quadratic distance between $X$ and $Y$ is defined as

$$E_Z |X - Y|^2 = \int ||v - h||^2 \, dZ(v, h).$$

The Fréchet distance between distributions $G$ and $H$ is defined as

$$d^2(G, H) = \min_{Z: Z_1 = G, Z_2 = H} E_Z |X - Y|^2.$$

The minimization is taken over all distributions $Z$ which have $G$ and $H$ as marginals. The minimization is governed by optimizing the covariance structure between the marginals. If the $X$ and $Y$ have zero means and covariance matrices $\Sigma_G$ and $\Sigma_H$, respectively, then

$$E_Z |X - Y|^2 = \text{Tr}\left[\Sigma_G + \Sigma_H - \Sigma_G \Sigma_H - \Sigma_G^T \Sigma_H^T\right],$$

where $\Sigma_{GH}$ is the covariance matrix between $X$ and $Y$ determined by $Z$. Dowson and Landau (Dowson & Landau, 1982) showed that the Fréchet distance for multivariate normal distributions $G$ and $H$ is

$$d^2(G, H) = ||\mu_G - \mu_H||^2 + \text{Tr}\left[\Sigma_G + \Sigma_H - 2 \left(\Sigma_G \Sigma_H\right)^{1/2}\right],$$

where $\mu_G, \mu_H$ and $\Sigma_G, \Sigma_H$ are the respective means and covariance matrices of $G$ and $H$, respectively.

For the minimization of the squared Fréchet distance $d^2(Q, p)$ with respect to $\Sigma_q$, we require

$$\frac{\partial}{\partial \Sigma_q} \text{Tr}\left(\Sigma_q + \Sigma_p - 2 \left(\Sigma_q \Sigma_p\right)^{1/2}\right) = I - 2 \frac{1}{2} \left(\Sigma_q \Sigma_p\right)^{-1/2} \Sigma_p = 0.$$

This gives

$$\Sigma_q = \Sigma_p$$

as for the KL divergence. Again using the second order derivative of the matrices as vectors, we confirm that it is a minimum. This solution follows immediately from the fact that

$$\text{Tr}\left(\Sigma_q + \Sigma_p - 2 \left(\Sigma_q \Sigma_p\right)^{1/2}\right)$$

defines a metric on the covariance matrices (Dowson and Landau, 1982). If we use only terms in $\mu_q$, the minimization with respect to $\mu_q$ gives the constraint minimization problem:

$$\min_{\mu_q} \frac{1}{2} \mu_q^T \mu_q - \mu_q^T \mu$$

s.t. $\mu_q \geq 0$. 

The Lagrangian with multiplier $\lambda \geq 0$ is
\begin{equation}
L = \frac{1}{2} \mu_q^T \mu_q - \mu_p^T \mu_q - \lambda^T \mu_q .
\end{equation}

The derivative with respect to $\mu_q$ is
\begin{equation}
\frac{\partial L}{\partial \mu_q} = \mu_q - \mu_p - \lambda = 0 .
\end{equation}

The Karush-Kuhn-Tucker (KKT) conditions require for the optimal solution that for each constraint $j$:
\begin{equation}
\lambda_j \mu_{qj} = 0 .
\end{equation}

If $0 < \mu_{pj}$ then Eq. (70) requires $0 < \mu_{qj}$ because the Lagrangian $\lambda_j$ is larger than or equal to zero: $0 \leq \lambda_j$. From the KKT conditions Eq. (71) follows that $\lambda_j = 0$ and, therefore, $0 < \mu_{qj} = \mu_{pj}$. If $\mu_{pj} < 0$ then Eq. (70) requires $0 < \lambda_j$ because the constraints of the primal problem require $0 \leq \mu_{qj}$. From the KKT conditions Eq. (71) follows that $\mu_{pj} = \mu_{qj} = \lambda_j = 0$.

Therefore the solution of problem Eq. (68) is
\begin{equation}
\mu_{qj} = \begin{cases} 
\mu_{pj} & \text{for } \mu_{pj} > 0 \text{ and } \lambda_j = 0 \\
0 & \text{for } \mu_{pj} \leq 0 \text{ and } \lambda_j = -\mu_{pj} 
\end{cases} .
\end{equation}

This is the rectifying update rule of the RFN algorithm. Therefore, the RFN algorithm ensures that the mean values of the posterior still contains information on the input. Normalization of the hidden mean values does not change the information content since normalization is a fixed rescaling procedure for a learned RFN model.

**S7. Posterior for Gaussian Shift Prior**

We want to formulate RFN learning as a variational EM algorithm. Toward this end we have to introduce a Gaussian shift prior. Here we derive the posterior of the factor analysis model given a Gaussian shift prior. We consider the posterior on the hidden factor if the prior is a Gaussian shifted by the variational parameter $\xi$. The Gaussian prior is parametrized by the variational parameter $\xi$:
\begin{equation}
p(h; \xi) = (2\pi)^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (h - \xi)^T (h - \xi) \right) .
\end{equation}

For the prior on the latent variables we have
\begin{align}
E(h) &= \xi \\
\text{var}(h) &= I \\
E(h^T) &= \text{var}(h) + E(h)E(h)^T = I + \xi \xi^T
\end{align}

We obtain for the first two moments of the data:
\begin{align}
E(v) &= E(Wh + \epsilon) = WE(h) + E(\epsilon) = W\xi , \\
E(v^T) &= E((Wh + \epsilon)(Wh + \epsilon)^T) = WE(hh^T)W^T \\
&\quad + WE(h)E(\epsilon^T) + E(\epsilon)E(h^T)W^T + E(\epsilon \epsilon^T) \\
&\quad = W(I + \xi \xi^T)W^T + \Psi \\
\text{var}(v) &= E(vv^T) - E(v)E(v)^T = WW^T + \Psi .
\end{align}

Therefore, the marginal distribution for $v$ is
\begin{equation}
v \sim \mathcal{N}(W\xi , WW^T + \Psi) .
\end{equation}
Rectified Factor Networks

For the covariances we have:

\[ E(hv^T) = E(h h^T) W^T = (I + \xi \xi^T) W^T \]
\[ \text{Cov}(h, v) = E(hv^T) - E(h)E(v^T) = W^T \]
\[ E(vh^T) = W (I + \xi \xi^T) \]
\[ \text{Cov}(v, h) = E(vh^T) - E(v)E(h^T) = W . \]

We need some algebraic identities. Woodbury matrix identity gives

\[ I - W^T (W W^T + \Psi)^{-1} W = (I + W^T \Psi^{-1} W)^{-1} . \]

Using again the Woodbury matrix identity, we have

\[ (W W^T + \Psi)^{-1} W = \Psi^{-1} W - \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} W \Psi^{-1} W - \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} (I + W^T \Psi^{-1} W) - \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} W = \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} (I + W^T \Psi^{-1} W - W^T \Psi^{-1} W) = \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} \]

and obtain

\[ W^T (W W^T + \Psi)^{-1} \xi = (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} \xi . \]

The conditional distribution \((a \mid u)\) of two random variables \(a\) and \(u\) that both follow a Gaussian distribution is:

\[ a \sim \mathcal{N}(\mu_a, \Sigma_{aa}) , \]
\[ u \sim \mathcal{N}(\mu_u, \Sigma_{uu}) , \]
\[ \Sigma_{aa} = \text{Cov}(u, a) , \]
\[ \Sigma_{au} = \text{Cov}(a, u) , \]
\[ a \mid u \sim \mathcal{N}(\mu_a + \Sigma_{au} \Sigma_{uu}^{-1} (u - \mu_u) , \Sigma_{aa} - \Sigma_{au} \Sigma_{uu}^{-1} \Sigma_{au}) . \]

Thus, the mean of \(p(h \mid v, \xi)\) is

\[ \mu_{h|v} = \xi + W^T (W W^T + \Psi)^{-1} (v - W \xi) = (I - W^T (W W^T + \Psi)^{-1} W) \xi + W^T (W W^T + \Psi)^{-1} v = (I + W^T \Psi^{-1} W)^{-1} \xi + (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} v = (I + W^T \Psi^{-1} W)^{-1} (W^T \Psi^{-1} v + \xi) . \]

Furthermore it follows that the covariance matrix of \(p(h \mid v, \xi)\) is

\[ \Sigma_{h|v} = I - W^T (W W^T + \Psi)^{-1} W = (I + W^T \Psi^{-1} W)^{-1} . \]

**S8. RFN Fixed Point**

The fixed point equation for the \(W\) update is

\[ \Delta W = U S^{-1} - W = 0 \Rightarrow U - W S = 0 \]
\[ \Rightarrow U W^T - W S W^T = 0 . \]
where for the last equation we multiplied the previous from the right hand side by $W^T$. The fixed point equation for the full (not only diagonal) $\Psi$ update is

$$\Psi = C - U W^T - W U^T + W S W^T = C - W U^T,$$

(95)

where we inserted Eq. (94). Since both $C$ and $\Psi$ are symmetric we have $W U^T = U W^T = C - \Psi$. Inserting $U W^T$ in last Eq. (94) gives

$$C = \Psi + W S W^T.$$

(96)

Therefore the model corresponding to the fixed point explains the empirical matrix of second moments $C$ by a noise part $\Psi$ and a signal part $W S W^T$. Like factor analysis the data variance is explained by the model via the parameters $\Psi$ (noise) and $W$ (signal).

In contrast to standard factor analysis, $S$ estimates the second moments of the prior. The model distribution’s mean is

$$E(v) = E(W h + \epsilon) = W E(h).$$

(97)

The model distribution’s second moment is

$$E(v v^T) = E((W h + \epsilon)(W h + \epsilon)^T) = W E(h^T h) W^T + W E(h) E(\epsilon^T) + E(\epsilon) E(h^T) W^T + E(\epsilon^T)$$

$$= W S W^T + \Psi.$$  

The variance of model distribution is

$$\text{var}(v) = E(v v^T) - E(v)E(v)^T = W (S - E(h) E(h^T)) W^T + \Psi.$$  

(99)

The fixed point equation Eq. (94) gives $W = U S^{-1}$. If $v_i$ and $\mu_{h_i|v_i}$ are correlated, then $U$ is large. The eigenvalues of $\Sigma_{h_i|v_i}$ are smaller than one. The normalization step keeps the variance of $\mu_{h_i|v_i}$ to one. Therefore the eigenvalues of $S$ are bounded. The bound is 2 for decorrelated inputs and hidden units while it goes to 1 if $\Sigma_{h_i|v_i}$ goes to zero. Thus, large $U$ leads to large $W$. Therefore $\Sigma_{h_i|v_i} = (I + W^T \Psi^{-1} W)^{-1}$ is small, and Eq. (94) becomes

$$W = \left(\frac{1}{n} \sum_{i=1}^n v_i \mu_{h_i|v_i} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mu_{h_i|v_i} \mu_{h_i|v_i}^T \right)^{-1}.$$  

(100)

In this case, $W$ is a least squares estimate for reconstruction errors $\epsilon_i = v_i - W \mu_{h_i|v_i}$. For large $W$ we obtain $W \Sigma_{h_i|v_i} W^T = W W^T (W W^T + \Psi)^{-1} \Psi \approx \Psi$. Since the covariance of the input noise is $\Psi = \frac{1}{n} \sum_{i=1}^n \epsilon_i \epsilon_i^T$, we have $\Psi \approx \frac{1}{2n} \sum_{i=1}^n \epsilon_i \epsilon_i^T$. Therefore $\Psi$ becomes small as the least square estimate keeps $\epsilon_i$ small. In summary, correlation between $v_i$ and $\mu_{h_i|v_i}$ leads to large $W$, small $\Sigma_{h_i|v_i}$, and small $\Psi$. Therefore $\Psi$ can be constrained to a diagonal matrix without changing the solution.

**S9. RFN Objective**

For the $i$-th sample the data vector or the vector of visible units is $v_i$ and the vector of hidden units is $h_i$. For the moment we assume to have only sample $v_i$. The model likelihood is $p(v_i)$, the probability that the model produces visible units $v_i$. The prior on hidden units is $p(h_i)$, the probability to generate hidden unit activations without further information. The posterior of the hidden units is $p(h_i | v_i)$, the probability of hidden unit activations if data $v_i$ has been observed. We defined a variational distribution $Q(h_i)$ which is computed from $v_i$ using current parameter estimates but is independent
of the parameters. The variational distribution $Q(h_i)$ is introduced to bound the log-likelihood $\log p(v_i)$:

$$\log p(v_i) - D_{KL}(Q(h_i) \parallel p(h_i \mid v_i))$$

$$= \int Q(h_i) \log p(v_i) \, dh_i - \int Q(h_i) \log \frac{Q(h_i)}{p(h_i \mid v_i)} \, dh_i$$

$$= - \int Q(h_i) \log \frac{Q(h_i)}{p(h_i \mid v_i)} \, dh_i$$

$$= - \int Q(h_i) \log \frac{Q(h_i)}{p(h_i \mid v_i)} \, dh_i + \int Q(h_i) \log p(v_i \mid h_i) \, dh_i$$

$$= \int Q(h_i) \log p(v_i \mid h_i) \, dh_i - D_{KL}(Q(h_i) \parallel p(h_i))$$

$$= \mathcal{F}.$$  

$\mathcal{F}$ is the **objective** in our framework which has to be maximized. $\mathcal{F}$ is called the negative free energy. $\mathcal{F}$ is equal to the log-likelihood $\log p(v_i)$ if

$$D_{KL}(Q(h_i) \parallel p(h_i \mid v_i)) = 0,$$

that is,

$$Q(h_i) = p(h_i \mid v_i).$$

Therefore $\mathcal{F}$ is equal to the log-likelihood $\log p(v_i)$ if the variational distribution is equal to the posterior.

We focus on the decomposition of $\mathcal{F}$ in the two terms in the last but one line of Eq. (101). The first term is the expected reconstruction error $\mathcal{E}$:

$$\mathcal{E} = \int_{\mathbb{R}^l} Q(h_i) \log (p(v_i \mid h_i)) \, dh_i.$$  

(104)

The expectation is over hidden variables $h_i$ distributed according to $Q(h_i)$. The second term in the objective is the Kullback-Leibler divergence $D_{KL}(Q(h_i) \parallel p(h_i))$ between the variational distribution $Q(h_i)$ and the prior $p(h_i)$. $D_{KL}$ penalizes distributions $Q(h_i)$ which deviate from the desired prior distribution $p(h_i)$. Therefore $Q(h_i)$ should contain information on $v_i$ to allow for a low reconstruction error but at the same time it should be close to a desired distribution. Therefore $D_{KL}$ and $p(h_i)$ determine how information on $v_i$ is stored in $h_i$, while $\mathcal{E}$ enforces to store information on $v_i$ in $h_i$. The negative free energy $\mathcal{F}$ is

$$\mathcal{F} = \int_{\mathbb{R}^l} Q(h_i) \log (p(v_i \mid h_i)) \, dh_i - D_{KL}(Q(h_i) \parallel p(h_i)).$$

(105)

or with $\mathcal{E}$

$$\mathcal{F} = \mathcal{E} - D_{KL}(Q(h_i) \parallel p(h_i)).$$

(106)

If $\mathcal{E}$ is viewed as energy and $p(h_i)$ as distribution of the initial or default state, then $Q(h_i)$ indicates how far the energy $\mathcal{E}$ pushes the hidden states $h_i$ away from their initial states. The maximal value of $\mathcal{F}$ is an equilibrium between large $\mathcal{E}$ and small $D_{KL}$.

We can generalize to $n$ samples. The expected reconstruction error $\mathcal{E}$ for $n$ data samples is

$$\mathcal{E} = \sum_{i=1}^{n} \int_{\mathbb{R}^l} Q(h_i) \log (p(v_i \mid h_i)) \, dh_i$$

(107)

and objective becomes

$$\mathcal{F} = \mathcal{E} - \sum_{i=1}^{n} D_{KL}(Q(h_i) \parallel p(h_i)).$$

(108)
Next we define the model in order to compute the likelihood and the reconstruction error. We utilize a factor analysis model with parameters $W$ and $\Psi$ given by
\begin{equation}
    v_i = W h_i + \epsilon_i ,
\end{equation}
where the Gaussian noise $\epsilon_i$ is
\begin{equation}
    \epsilon_i \sim N(0, \Psi) .
\end{equation}
Thus, the conditional probability of data $v_i$ given hidden variables $h_i$ is:
\begin{equation}
    p(v_i | h_i, W, \Psi) = (2\pi)^{-m/2} |\Psi|^{-1/2} \exp\left(-\frac{1}{2} (v_i - Wh_i)^T \Psi^{-1} (v_i - Wh_i)\right) .
\end{equation}
The expected reconstruction error $\mathcal{E}$ using parameters $(W, \Psi)$ is
\begin{equation}
    \mathcal{E} = \sum_{i=1}^n \int_{\mathbb{R}^d} Q(h_i) \log (p(v_i | h_i, W, \Psi)) \, dh_i .
\end{equation}
For Gaussian noise, the component-wise reconstruction errors are weighted by $\Psi$. The variational distribution $Q(h_i)$ is computed from $v_i$ using estimates $\hat{W}$ and $\hat{\Psi}$ of the parameters $W$ and $\Psi$. We assume that $Q(h_i)$ is independent of the parameters $W$ and $\Psi$.

For computing the expected weighted reconstruction error $\mathcal{E}$ we need only the mean and the variance of $Q(h_i)$, because $p(v_i | h_i, W, \Psi)$ is Gaussian and its log is a quadratic function in $h_i$. The mean of $Q(h_i)$ is
\begin{equation}
    \mu_{h_i|v_i} = E_{\hat{h}_i|v_i}(h_i)
\end{equation}
and the variance
\begin{equation}
    \Sigma_{h_i|v_i} = \text{Var}(h_i) .
\end{equation}
The second moment of $Q(h_i)$ can be obtained by
\begin{equation}
    E_{\hat{h}_i|v_i}(h_i h_i^T) = E_{\hat{h}_i|v_i}(h_i) E_{\hat{h}_i|v_i}(h_i) + \text{Var}(h_i) = \mu_{h_i|v_i} \mu_{h_i|v_i}^T + \Sigma_{h_i|v_i} .
\end{equation}
The objective $\mathcal{F}$ is a function of the parameters $(W, \Psi)$ and the first and second (central) moments $\{\mu_{h_i|v_i}, \Sigma_{h_i|v_i}\}$ of $Q(h_i)$:
\begin{equation}
    \mathcal{F} = \mathcal{F}(W, \Psi, \{\mu_{h_i|v_i}, \Sigma_{h_i|v_i}\}) .
\end{equation}
$\mathcal{F}$ is maximized with respect to both $(W, \Psi)$ and $\{\mu_{h_i|v_i}, \Sigma_{h_i|v_i}\}$. The parameters $(W, \Psi)$ enter only $\mathcal{E}$ but not $D_{\text{Kullback-Leibler}}(Q \parallel p)$. The parameters $(W, \Psi)$ are adjusted to make the data $v_i$ more likely to be produced by the model given the current $Q(h_i)$. Thus, $\mathcal{E}$ is maximized with respect to $(W, \Psi)$ and, thereby, made less negative. However, there is a trade-off when updating the moments $\{\mu_{h_i|v_i}, \Sigma_{h_i|v_i}\}$ of the variational distribution $Q(h_i)$: these moments should explain the data via large (less negative) $\mathcal{E}$ and at the same time they should push $Q(h_i)$ close to the prior $p(h_i)$ to decrease $D_{\text{Kullback-Leibler}}(Q(h_i) \parallel p(h_i))$.

The next step in model design is to choose the variational and the prior distribution. We assume that both the variational distribution and the prior are Gaussian.

- the variational distribution $Q(h_i)$ is Gaussian distributed:
\begin{equation}
    Q(h_i) = N(\mu_{h_i|v_i}, \Sigma_{h_i|v_i})
\end{equation}

- the prior $p(h_i)$ is a zero mean Gaussian with the identity as covariance matrix (standardized Gaussian):
\begin{equation}
    p(h_i) = N(0, I) ,
\end{equation}
where $\mathcal{N}$ denotes a Gaussian distribution with mean vector and covariance matrix as parameters. Therefore, the variational distribution $Q(h_i)$ is

$$ Q(h_i) = (2\pi)^{-n/2} |\Sigma_{h_i|v_i}|^{-1/2} \exp \left( -\frac{1}{2} (h_i - \mu_{h_i|v_i})^T \Sigma_{h_i|v_i}^{-1} (h_i - \mu_{h_i|v_i}) \right). \quad (119) $$

The objective $\mathcal{F}$ becomes

$$ \mathcal{F} = \mathcal{E} - \sum_{i=1}^n D_{KL}(\mathcal{N}(\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \parallel \mathcal{N}(0, I)). \quad (120) $$

Since we have chosen $Q(h_i)$, we can compute $\mathcal{E}$:

$$ \mathcal{E} = -\frac{m \cdot n}{2} \log(2\pi) - \frac{n}{2} \log|\Psi| $$

$$ + \frac{1}{2} \sum_{i=1}^n E_{h_i|v_i} \left((v_i - WH_i)^T \Psi^{-1} (v_i - WH_i)\right) $$

$$ - \frac{1}{2} \sum_{i=1}^n v_i^T \Psi^{-1} v_i + \frac{1}{2} \sum_{i=1}^n E_{h_i|v_i} \left(v_i^T \Psi^{-1} WH_i\right) $$

$$ + \frac{1}{2} \sum_{i=1}^n E_{h_i|v_i} \left(h_i^T W^T \Psi^{-1} v_i\right) $$

$$ - \frac{1}{2} \sum_{i=1}^n E_{h_i|v_i} \left(h_i^T W^T \Psi^{-1} W h_i\right) $$

$$ = -\frac{m \cdot n}{2} \log(2\pi) - \frac{n}{2} \log|\Psi| $$

$$ - \frac{1}{2} \text{Tr} \left( \Psi^{-1} \sum_{i=1}^n v_i v_i^T \right) + \frac{1}{2} \text{Tr} \left( \Psi^{-1} W \sum_{i=1}^n E_{h_i|v_i} (h_i) v_i^T \right) $$

$$ + \frac{1}{2} \text{Tr} \left( \Psi^{-1} \sum_{i=1}^n v_i^T E_{h_i|v_i} (h_i) W^T \right) $$

$$ - \frac{1}{2} \text{Tr} \left( \Psi^{-1} W \sum_{i=1}^n E_{h_i|v_i} (h_i) h_i^T W^T \right) $$

$$ = -\frac{m \cdot n}{2} \log(2\pi) - \frac{n}{2} \log|\Psi| $$

$$ - \frac{1}{2} \text{Tr} \left( \Psi^{-1} \sum_{i=1}^n v_i v_i^T \right) + \frac{1}{2} \text{Tr} \left( \Psi^{-1} W \sum_{i=1}^n \mu_{h_i|v_i} v_i^T \right) $$

$$ + \frac{1}{2} \text{Tr} \left( \Psi^{-1} \sum_{i=1}^n v_i^T \mu_{h_i|v_i} W^T \right) $$

$$ - \frac{1}{2} \text{Tr} \left( \Psi^{-1} W \sum_{i=1}^n \left(\mu_{h_i|v_i} \mu_{h_i|v_i}^T + \Sigma_{h_i|v_i}\right) W^T \right). $$

Also the Kullback-Leibler divergence can be computed:

$$ D_{KL}(\mathcal{N}(\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \parallel \mathcal{N}(0, I)) $$

$$ = \frac{1}{2} \text{Tr}(\Sigma_{h_i|v_i}) + \frac{1}{2} \mu_{h_i|v_i}^T \mu_{h_i|v_i} - \frac{1}{2} n - \frac{1}{2} \ln |\Sigma_{h_i|v_i}|. \quad (122) $$
The full objective is
\[
\mathcal{F} = -\frac{m}{2} \log (2\pi) - \frac{n}{2} \log |\Psi| \tag{123}
\]
\[\begin{align*}
&- \frac{1}{2} \text{Tr} \left( \Psi^{-1} \sum_{i=1}^{n} v_i v_i^T \right) + \frac{1}{2} \text{Tr} \left( \Psi^{-1} W \sum_{i=1}^{n} \mu_{h_i|v_i} v_i^T \right) \\
&+ \frac{1}{2} \text{Tr} \left( \Psi^{-1} \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T W^T \right) \\
&- \frac{1}{2} \text{Tr} \left( \Psi^{-1} W \sum_{i=1}^{n} \left( \mu_{h_i|v_i}, \mu_{h_i|v_i}^T + \Sigma_{h_i|v_i} \right) W^T \right) \\
&- \frac{1}{2} \text{Tr} \left( \sum_{i=1}^{n} \Sigma_{h_i|v_i} \right) - \frac{1}{2} \sum_{i=1}^{n} \mu_{h_i|v_i}^T \mu_{h_i|v_i} + \frac{1}{2} l + \frac{1}{2} \sum_{i=1}^{n} \log |\Sigma_{h_i|v_i}| .
\end{align*}\]

This objective can be rewritten as
\[
\frac{2}{n} \mathcal{F} = -m \log (2\pi) - \log |\Psi| \tag{124}
\]
\[\begin{align*}
&- \text{Tr} \left( \Psi^{-1} \frac{1}{n} \sum_{i=1}^{n} \left( v_i - W \mu_{h_i|v_i} \right) \left( v_i - W \mu_{h_i|v_i} \right)^T \right) \\
&- \text{Tr} \left( \Psi^{-1} W \frac{1}{n} \sum_{i=1}^{n} \Sigma_{h_i|v_i} W^T \right) \\
&- \text{Tr} \left( \frac{1}{n} \sum_{i=1}^{n} \Sigma_{h_i|v_i} \right) - \frac{1}{n} \sum_{i=1}^{n} \mu_{h_i|v_i}^T \mu_{h_i|v_i} + \frac{l}{n} + \frac{1}{n} \sum_{i=1}^{n} \log |\Sigma_{h_i|v_i}| \\
&= -m \log (2\pi) - \log |\Psi| \\
&- \frac{1}{n} \sum_{i=1}^{n} \left( v_i - W \mu_{h_i|v_i} \right)^T \Psi^{-1} \left( v_i - W \mu_{h_i|v_i} \right) \\
&- \text{Tr} \left( \frac{1}{n} \sum_{i=1}^{n} \Sigma_{h_i|v_i} \left( W^T \Psi^{-1} W + I \right) \right) \\
&- \frac{1}{n} \sum_{i=1}^{n} \mu_{h_i|v_i}^T \mu_{h_i|v_i} + \frac{l}{n} + \frac{1}{n} \sum_{i=1}^{n} \log |\Sigma_{h_i|v_i}| .
\end{align*}\]

Now we can maximize $\mathcal{F}$. First we maximize it with respect to the parameters ($W, \Psi$) (M-step) and then with respect to the moments ($\mu_{h_i|v_i}, \Sigma_{h_i|v_i}$) of the variational distribution $Q$ (E-step).

First we maximize $\mathcal{F}$ with respect to ($W, \Psi$). Since the parameters ($W, \Psi$) only enter $\mathcal{E}$, we have only to maximize this term with respect to ($W, \Psi$). The derivative of $\mathcal{E}$ with respect to $\Psi$ is
\[
\nabla_{\Psi} \mathcal{E} = -\frac{n}{2} \Psi^{-1} \tag{125}
\]
\[\begin{align*}
&+ \frac{1}{2} \Psi^{-1} \left( \sum_{i=1}^{n} v_i v_i^T \right) \Psi^{-1} - \frac{1}{2} \Psi^{-1} W \sum_{i=1}^{n} \mu_{h_i|v_i} v_i^T \Psi^{-1} \\
&- \frac{1}{2} \Psi^{-1} \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T W^T \Psi^{-1} \\
&+ \frac{1}{2} \Psi^{-1} W \sum_{i=1}^{n} \Sigma_{h_i|v_i} \left( h_i \mu_{h_i|v_i}^T \right) W^T \Psi^{-1} .
\end{align*}\]
For the optimal $\Psi$ we have
\[
\nabla_{\Psi} \mathcal{E} = 0 ,
\]
which gives
\[
\Psi = \frac{1}{n} \left( \sum_{i=1}^{n} v_i v_i^T \right) - W \frac{1}{n} \sum_{i=1}^{n} \mu_{h_i|v_i} v_i^T - \frac{1}{n} \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T W^T + W \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) W^T .
\]

The derivative of $\mathcal{E}$ with respect to $W$ is
\[
\nabla_W \mathcal{E} = \frac{1}{2} \Psi^{-1} \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T - \frac{1}{2} \Psi^{-1} \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T - \frac{1}{2} \Psi^{-1} W \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) - \frac{1}{2} \Psi^{-1} W \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) = \Psi^{-1} \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T - \Psi^{-1} W \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) .
\]

For the optimal $W$ we have
\[
\nabla_W \mathcal{E} = 0 ,
\]
which gives
\[
W = \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T \left( \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) \right)^{-1} .
\]

We have defined
\[
C = \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T ,
\]
\[
U = \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) = \frac{1}{n} \sum_{i=1}^{n} v_i \mu_{h_i|v_i}^T ,
\]
\[
S = \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T)
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} \mu_{h_i|v_i} \mu_{h_i|v_i}^T + \frac{1}{n} \sum_{i=1}^{n} \Sigma_{h_i|v_i} .
\]

Using these definitions, the optimal parameters are
\[
\Psi = C - W U^T - U W^T + W S W^T ,
\]
\[
W = U S^{-1} .
\]
Using last equation we obtain:
\[ W U^T = U S^{-1} U^T \]  \hspace{1cm} (136)
\[ U W^T = U S^{-1} U^T \]  \hspace{1cm} (137)
\[ W S W^T = U S^{-1} U^T. \]  \hspace{1cm} (138)

Therefore we can reformulate the update of the optimal \( \Psi \) if the optimal \( W \) is given:
\[ \Psi = C - U S^{-1} U^T = C - W U^T. \]  \hspace{1cm} (139)

Next we maximize \( F \) with respect to the moments \( (\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \) of the variational distribution \( Q \). We have following derivatives:
\[ \nabla \mu_{h_i|v_i} \mathcal{E} = W^T \Psi^{-1} v_i - W^T \Psi^{-1} W \mu_{h_i|v_i}, \]  \hspace{1cm} (140)
\[ \nabla \Sigma_{h_i|v_i} \mathcal{E} = -\frac{1}{2} W^T \Psi^{-1} W \]  \hspace{1cm} (141)
\[ \nabla \mu_{h_i|v_i} D_{KL}(\mathcal{N}(\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \parallel \mathcal{N}(0, I)) = \mu_{h_i|v_i} \]  \hspace{1cm} (142)
\[ \nabla \Sigma_{h_i|v_i} D_{KL}(\mathcal{N}(\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \parallel \mathcal{N}(0, I)) = \frac{1}{2} I - \frac{1}{2} \Sigma_{h_i|v_i}^{-1}. \]  \hspace{1cm} (143)

We set the derivatives of the objective
\[ F = \mathcal{E} - \sum_{i=1}^{n} D_{KL}(\mathcal{N}(\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \parallel \mathcal{N}(0, I)) \]  \hspace{1cm} (144)
with respect to \( (\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \) to zero and obtain
\[ \mu_{h_i|v_i} = (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} v_i \]  \hspace{1cm} (145)
\[ \Sigma_{h_i|v_i} = \Sigma_{h|v} = (I + W^T \Psi^{-1} W)^{-1}. \]  \hspace{1cm} (146)

Without \( D_{KL} \) we obtain for \( \mu_{h_i|v_i} \) the least squares estimator
\[ \mu_{h_i|v_i} = (W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} v_i \]  \hspace{1cm} (147)
and the covariance \( \Sigma_{h_i|v_i} \) of the hidden units goes to zero:
\[ \Sigma_{h_i|v_i} \to 0. \]  \hspace{1cm} (148)

Next we introduce a variational parameter \( \xi_i \) for the \( i \)-th sample which only enters the prior. The Gaussian prior is parametrized by \( \xi_i \) which shifts the mean away from zero:
\[ p(h_i) = \mathcal{N}(\xi_i, I). \]  \hspace{1cm} (149)

The Kullback-Leibler divergence becomes
\[ D_{KL}(\mathcal{N}(\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \parallel \mathcal{N}(\xi_i, I)) \]  \hspace{1cm} (150)
\[ = \frac{1}{2} \text{Tr}(\Sigma_{h_i|v_i}) + \frac{1}{2} (\mu_{h_i|v_i} - \xi_i)^T (\mu_{h_i|v_i} - \xi_i) - \frac{1}{2} l - \frac{1}{2} \ln |\Sigma_{h_i|v_i}|. \]
Only the derivative with respect to $\mu_{h_i|v_i}$ changes:

$$
\nabla_{\mu_{h_i|v_i}} D_{KL}(\mathcal{N}(\mu_{h_i|v_i}, \Sigma_{h_i|v_i}) \parallel \mathcal{N}(0, I)) = \mu_{h_i|v_i} - \xi_i. 
$$

(151)

For the optimal moments that determine $Q(h_i)$ we obtain

$$
\mu_{h_i|v_i} = (I + W^{T} \Psi^{-1} W)^{-1} \left(W^{T} \Psi^{-1} v_i + \xi_i\right) 
$$

(152)

$$
\Sigma_{h_i|v_i} = \Sigma_{h|v} = (I + W^{T} \Psi^{-1} W)^{-1}.
$$

(153)

If a desired mean vector $\mu^q_i$ for sample $v_i$ is given, then the variational parameter $\xi_i$ can be chosen to be

$$
\xi_i = (W^{T} \Psi^{-1} W + I) \mu^q_i - W^{T} \Psi^{-1} v_i
$$

(154)

which enforces

$$
\mu_{h_i|v_i} = \mu_i^q.
$$

(155)

Using the mean $\mu_{h_i|v_i}$ of $Q(h_i)$ obtained by a zero mean Gaussian prior, we can choose $\mu_i^q$ using rectifying, dropout, and normalization. If the dropout probability is $d$, then we determine $\mu_i^q$ as follows:

$$
\Pr(\delta = 0) = d, \quad \Pr(\delta = 1) = 1 - d
$$

(156)

$$
[\mu_i^q]_j = \delta \max \left\{ 0, [\mu_{h_i|v_i}]_j \right\}
$$

(157)

$$
[\mu_i^q]_j = \frac{[\mu_{h_i|v_i}]_j}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} [\mu_{h_i|v_i}]_j^2}}.
$$

(158)

### S10. Maximum Likelihood Factor Analysis

We are given the data $\{v\} = \{v_1, \ldots, v_n\}$ which are assumed to be centered, which can be done by subtracting the mean $\mu$ from the data. The model is

$$
v = Wh + \epsilon,
$$

(159)

where

$$
h \sim \mathcal{N}(0, I), \quad \epsilon \sim \mathcal{N}(0, \Psi).
$$

(160)

The model contains the observations $v \in \mathbb{R}^m$, the noise $\epsilon \in \mathbb{R}^m$, the factors $h \in \mathbb{R}^l$, the factor loading matrix $W \in \mathbb{R}^{m \times l}$, and the noise covariance matrix $\Psi$, which is a diagonal matrix from $\mathbb{R}^{m \times m}$. The data variance is explained through a signal part $Wh$ and through a noise part $\epsilon$. The parameters of the model are $W$ and $\Psi$. From the model assumption it follows that

$$
v \mid h \sim \mathcal{N}(Wh, \Psi).
$$

(161)

If $h$ is given, then only the noise $\epsilon$ is a random variable. In general, the number $l$ of factors is smaller than the number $m$ of features of the observations: $m \geq l$. A diagonal $\Psi$ assumes that the measurements are independent and the noise parts of the components are mutually independent. Therefore, the observations are mutually independent if the factors are known (only the noise is the random variable). Thus, **correlations between observations can only be explained by factors.**

We will now consider the likelihood of the data. Let $E$ denote the expectation of the data (i.e. the factor distribution and the noise distribution is combined), then we obtain for the first two moments:

$$
E(v) = E(Wh + \epsilon) = WE(h) + E(\epsilon) = 0,
$$

(162)

$$
E(vv^{T}) = E((Wh + \epsilon)(Wh + \epsilon)^{T}) = WE(hh^{T})W^{T} + WE(h)E(\epsilon^{T}) + E(\epsilon)E(h^{T})W^{T} + E(\epsilon\epsilon^{T}) = W^{T} + \Psi.
$$
The variance can be computed as
\[ \text{var}(v) = E(vv^T) - (E(v))^2 = WW^T + \Psi. \] (163)

Therefore, the marginal distribution for \( v \) is
\[ v \sim \mathcal{N}(0, WW^T + \Psi). \] (164)

This means that the observations are Gaussian distributed. The log-likelihood of the data \( \{v\} \) under the model \((W, \Psi)\) is
\[
\log \prod_{i=1}^{n} \frac{1}{(2\pi)^{-m/2}} |WW^T + \Psi|^{-1/2} \exp \left( -\frac{1}{2} (v_i^T (WW^T + \Psi)^{-1} v_i) \right), \] (165)

where \(|.|\) denotes the absolute value of the determinant of a matrix.

To maximize the likelihood is difficult since a closed form for maximizing the likelihood does not exist. Therefore, we apply the EM-algorithm. We introduce a distribution \( Q_i(h_i) = p(h_i | v_i; W, \Psi) \) which estimates the hidden states, here the factors. Using
\[ Q_i(h_i) = p(h_i | v_i; W, \Psi) \] (166)
then
\[
\begin{align*}
\mu_{h_i|v_i} &= (v_i)^T (WW^T + \Psi)^{-1} W, \\
\Sigma_{h_i|v_i} &= I - WW^T (WW^T + \Psi)^{-1} W,
\end{align*}
\] (167)

where we used the fact that
\[
\begin{align*}
a &\sim \mathcal{N}(\mu_u, \Sigma_{aa}) \quad u \sim \mathcal{N}(\mu_u, \Sigma_{uu}) \quad \Sigma_{ua} = \text{Cov}(u, a) \quad \Sigma_{au} = \text{Cov}(a, u) \\
a | u &\sim \mathcal{N}(\mu_u + \Sigma_{au} \Sigma_{uu}^{-1} (u - \mu_u), \Sigma_{aa} - \Sigma_{au} \Sigma_{uu}^{-1} \Sigma_{ua})
\end{align*}
\] (168)

and
\[ E(hv) = W E(h h^T) = W. \] (169)

We obtain
\[
Q_i(h_i) = (2\pi)^{-m/2} |\Sigma_{h_i|v_i}|^{-1/2} \exp \left( -\frac{1}{2} (h_i - \mu_{h_i|v_i})^T \Sigma_{h_i|v_i}^{-1} (h_i - \mu_{h_i|v_i}) \right). \] (170)

The EM algorithm for maximum likelihood maximizes in the M-step a lower bound for the log-likelihood:
\[
\log (p(v_i | W, \Psi)) = \log \left( \int_{\mathbb{R}^l} \frac{Q_i(h_i) p(v_i, h_i | W, \Psi)}{Q_i(h_i)} \, dh_i \right) \geq \int_{\mathbb{R}^l} Q_i(h_i) \log \left( \frac{p(v_i, h_i | W, \Psi)}{Q_i(h_i)} \right) \, dh_i. \] (171)

Using the expectation
\[ E_{h_i|v_i}(f(h_i)) = \int_{\mathbb{R}^l} Q_i(h_i) f(h_i) \, dh_i \] (172)
and neglecting all terms which are independent of \( W \) and \( \Psi \), the M-step requires to maximize
\[
\log \mathcal{L} = -\frac{m}{2} \log (2\pi) - \frac{n}{2} \log |\Psi| - \frac{1}{2} \sum_{i=1}^{n} E_{h_i|v_i} \left( (v_i - Wh_i)^T \Psi^{-1} (v_i - Wh_i) \right). \] (173)
The optimality criteria are

\[
\frac{1}{n} \nabla_W \log \mathcal{L} = -\frac{1}{n} \sum_{i=1}^{n} \Psi^{-1} W E_{h_i|v_i} (h_i, h_i^T) + \frac{1}{n} \sum_{i=1}^{n} \Psi^{-1} v_i E_{h_i|v_i}^T (h_i) = 0 \tag{174}
\]

and

\[
\nabla_\Psi \log \mathcal{L} = -\frac{n}{2} \Psi^{-1} + \frac{1}{2} \sum_{i=1}^{n} E_{h_i|v_i} \left( \Psi^{-1} (v_i - W h_i) (v_i - W h_i)^T \Psi \right) = 0 . \tag{175}
\]

Solving above equations gives:

\[
W^{\text{new}} = \left( \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i}^T (h_i) \right) \left( \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i, h_i^T) \right)^{-1} \tag{176}
\]

and

\[
\Psi^{\text{new}} = \text{diag} \left( \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} \left( (v_i - W^{\text{new}} h_i) (v_i - W^{\text{new}} h_i)^T \right) \right) = \text{diag} \left( \frac{1}{n} \sum_{i=1}^{n} v_i v_i^T - \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) (W^{\text{new}})^T - \frac{1}{n} \sum_{i=1}^{n} W^{\text{new}} E_{h_i|v_i} (h_i) v_i^T + W^{\text{new}} \sum_{i=1}^{n} E_{h_i|v_i} (h_i, h_i^T) (W^{\text{new}})^T \right), \tag{177}
\]

where “\text{diag}” enforces a diagonal matrix by setting all non-diagonal elements to zero.

From Eq. (176) we obtain

\[
W^{\text{new}} \left( \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i, h_i^T) \right) = \left( \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i}^T (h_i) \right) \tag{178}
\]

and can replace the last term of Eq. (177) with the left hand side of above equation. This leads to the fact that one term \(\frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i) W^{\text{new}} v_i^T\) cancels in Eq. (177). We obtain

\[
\Psi^{\text{new}} = \frac{1}{n} \text{diag} \left( \sum_{i=1}^{n} v_i v_i^T - \sum_{i=1}^{n} v_i E_{h_i|v_i}^T (h_i) (W^{\text{new}})^T \right) . \tag{179}
\]

This leads to the following EM updates:

**E-step:**

\[
E_{h_i|v_i} (h_i) = \mu_{h_i|v_i}
\]

\[
E_{h_i|v_i} (h_i, h_i^T) = \mu_{h_i|v_i} \mu_{h_i|v_i}^T + \Sigma_{h_i|v_i}
\]

**M-step:**

\[
W^{\text{new}} = \left( \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i}^T (h_i) \right) \left( \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i, h_i^T) \right)^{-1}
\]

\[
\Psi^{\text{new}} = \frac{1}{n} \text{diag} \left( \sum_{i=1}^{n} v_i v_i^T - \sum_{i=1}^{n} v_i E_{h_i|v_i}^T (h_i) (W^{\text{new}})^T \right) . \tag{182}
\]
The EM algorithms can be reformulated as:

**E-step:**

\[
E_{h_i|v_i}(h_i) = \mu_{h_i|v_i}
\]

\[
E_{h_i|v_i}(h_i h_i^T) = \mu_{h_i|v_i} \mu_{h_i|v_i}^T + \Sigma_{h_i|v_i}
\]

**M-step:**

\[
C = \sum_{i=1}^n v_i v_i^T
\]

\[
U = \frac{1}{n} \sum_{i=1}^n v_i E_{h_i|v_i}(h_i)
\]

\[
S = \frac{1}{n} \sum_{i=1}^n E_{h_i|v_i}(h_i h_i^T)
\]

\[
W_{\text{new}} = U S^{-1}
\]

\[
\Psi_{\text{new}} = \text{diag} \left( C - U (W_{\text{new}})^T \right) = \text{diag} \left( C - U S^{-1} U^T \right).
\]

**Speed Ups.** To speed up the algorithm especially for \( m > l \) the matrix inversion lemma can be used:

\[
(W W^T + \Psi)^{-1} = \Psi^{-1} - \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1},
\]

where \( \Psi^{-1} \) can be evaluated very fast because it is a diagonal matrix.

Another speed up is obtained by

\[
\frac{1}{n} \sum_{i=1}^n v_i E_{h_i|v_i}(h_i) = \left( \frac{1}{n} \sum_{i=1}^n v_i (v_i)^T \right) (W W^T + \Psi)^{-1} W =
\]

\[
C (W W^T + \Psi)^{-1} W =
\]

\[
C \left( \Psi^{-1} W - \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} W \right) =
\]

\[
C \left( \bar{W} - \bar{W} (I + B)^{-1} B \right),
\]

where \( W = \Psi^{-1} W, B = W^T \Psi^{-1} W = W^T \bar{W}, \) and \( C \) is the empirical covariance matrix (matrix of second moments), which has to be computed only once.

We can also compute

\[
\frac{1}{n} \sum_{i=1}^n \Sigma_{h_i|v_i} = I - W^T (W W^T + \Psi)^{-1} W =
\]

\[
I - W^T \Psi^{-1} W + W^T \Psi^{-1} W (I + W^T \Psi^{-1} W)^{-1} W^T \Psi^{-1} W =
\]

\[
I - B + B (I + B)^{-1} B =
\]

\[
I - B + BB (I + B)^{-1} =
\]

\[
(I - B) (I + B) (I + B)^{-1} + BB (I + B)^{-1} =
\]

\[
(I + B - B - BB + BB) (I + B)^{-1} =
\]

\[
(I + B)^{-1}
\]
where we used \((I + B)^{-1} B = B (I + B)^{-1}\), and

\[
\frac{1}{n} \sum_{i=1}^{n} \mu_{h_i|v_i} \mu_{h_i|v_i}^T = \frac{1}{n} \sum_{i=1}^{n} v_i (v_i)^T \left( WW^T + \Psi \right)^{-1} W = W^T \left( WW^T + \Psi \right)^{-1} C \left( WW^T + \Psi \right)^{-1} W = \left( \hat{W} - \hat{W} (I + B)^{-1} B \right)^T C \left( \hat{W} - \hat{W} (I + B)^{-1} B \right).
\]

Using these equations the E-step and the M-step can be unified and all sums \(\sum_{i=1}^{n}\) are removed and the matrix \(C\) can be computed once at the beginning of the iterative procedure.

**MAP factor analysis.** Above EM algorithm can be generalized to a maximum a posteriori (MAP) method with posterior \(p(W, \Psi | \{v\})\) which is proportional to the product between the likelihood \(p(\{v\} | W, \Psi)\) and the prior \(p(W)\):

\[
p(W, \Psi | \{v\}) \propto p(\{v\} | W, \Psi) p(W),
\]

therefore up to a constant independent of the parameters the log-posterior is

\[
\log(p(W, \Psi | \{v\})) = \log(p(\{v\} | W, \Psi)) + \log(p(W)).
\]