

# Unsupervised Representation Learning: Autoencoders, Factorization, and Sparse Coding

---

**Ju Sun**

Computer Science & Engineering  
University of Minnesota, Twin Cities

March 17, 2020

We have talked about

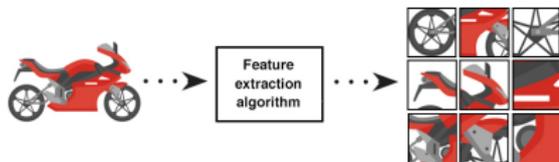
- Basic DNNs (multi-layer feedforward)
- Universal approximation theorems
- Numerical optimization and training DNNs

## **Models and applications**

- Unsupervised representation learning: autoencoders and variants
- DNNs for spatial data: CNNs
- DNNs for sequential data: RNNs, LSTM
- Generative models: variational Autoencoders and GAN
- Interactive models: reinforcement learning

involve modification and composition of the basic DNNs

# Feature engineering: old and new



**Feature engineering:** derive features for **efficient** learning

Credit: [Elgendy, 2020]

## Traditional learning pipeline



- feature extraction is “independent” of the learning models and tasks
- features are handcrafted and/or learned

## Modern learning pipeline



- end-to-end DNN learning

# Unsupervised representation learning

Learning feature/representation **without task information** (e.g., labels)  
(ICLR — International Conference on **Learning Representation**)

Why not jump into the end-to-end learning?

- **Historical:** Unsupervised representation learning key to the revival of deep learning (i.e., layerwise pretraining, [[Hinton et al., 2006](#), [Hinton, 2006](#)])

Science Contents - News - Careers - Journals -

SHARE REPORT

**Reducing the Dimensionality of Data with Neural Networks**

G. E. Hinton\*, R. R. Salakhutdinov  
+ See all authors and affiliations

Science | 26 Jul 2006  
Vol. 313, Issue 5786, pp. 504-507  
DOI: 10.1126/science.1127647

Article Figures & Data Info & Metrics eLetters PDF

Home | Neural Computation | List of Issues | Volume 18, No. 7 | A Fast Learning Algorithm for Deep Belief Nets

**NEURAL COMPUTATION**

**A Fast Learning Algorithm for Deep Belief Nets**

Geoffrey E. Hinton, Simon Osindero and Yee-Whye Teh

Printed Online May 07, 2006  
Epub Online 5/12/06  
© 2006 Massachusetts Institute of Technology

Neural Computation  
Volume 18 | Issue 7 | July 2006  
p1507-1554

Abstract Authors

Monthly 288pp per issue

Who chose how to use "complementary neurons" to eliminate the overabundance...

- **Practical:** Numerous advanced models built on top of the ideas in unsupervised representation learning (e.g., encoder-decoder networks)

PCA for linear data

Extensions of PCA for nonlinear data

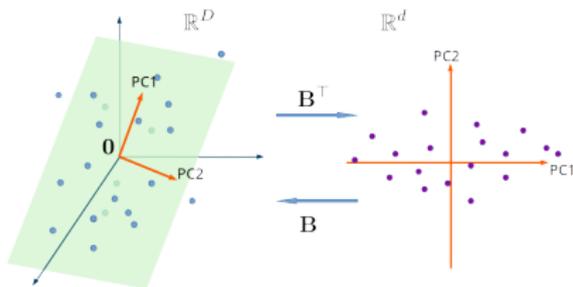
Application examples

Suggested reading

# PCA: the geometric picture

## Principal component analysis (PCA)

- Assume  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^D$  are zero-centered and write  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_m]$
- $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ , where  $\mathbf{U}$  spans the column space (i.e., range) of  $\mathbf{X}$
- Take top singular vectors  $\mathbf{B}$  from  $\mathbf{U}$ , and obtain  $\mathbf{B}^T\mathbf{X}$



PCA is effectively to identify the best-fit subspace to  $\mathbf{x}_1, \dots, \mathbf{x}_m$

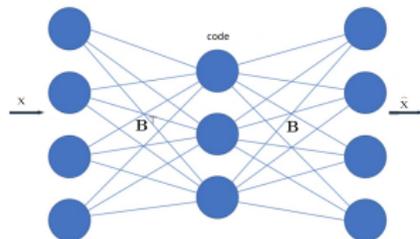
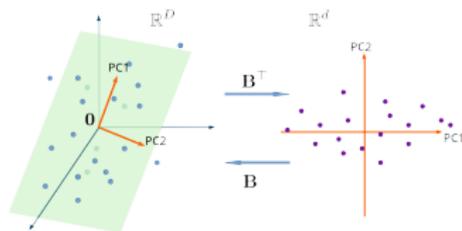
- $\mathbf{B}$  has orthonormal columns, i.e.,  $\mathbf{B}^T\mathbf{B} = \mathbf{I}$  ( $\mathbf{B}\mathbf{B}^T \neq \mathbf{I}$  when  $D \neq d$ )
- sample to representation:  
 $\mathbf{x} \mapsto \mathbf{x}' \doteq \mathbf{B}^T\mathbf{x}$  ( $\mathbb{R}^D \rightarrow \mathbb{R}^d$ , dimension reduction)
- representation to sample:  
 $\mathbf{x}' \mapsto \hat{\mathbf{x}} \doteq \mathbf{B}\mathbf{x}'$  ( $\mathbb{R}^d \rightarrow \mathbb{R}^D$ )
- $\hat{\mathbf{x}} = \mathbf{B}\mathbf{B}^T\mathbf{x} \approx \mathbf{x}$

# Autoencoders

story in digital communications ...



**autoencoder:** [Bourlard and Kamp, 1988, Hinton and Zemel, 1994]



– **Encoding:**

$$x \mapsto x' = B^T x$$

– **Decoding:**

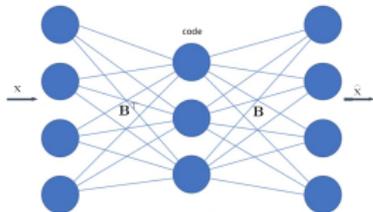
$$x' \mapsto BB^T x = \hat{x}$$

To find the basis  $B$ , solve ( $d \leq D$ )

$$\min_{B \in \mathbb{R}^{D \times d}} \sum_{i=1}^m \|x_i - BB^T x_i\|_2^2$$

# Autoencoders

autoencoder:



To find the basis  $B$ , solve

$$\min_{B \in \mathbb{R}^{D \times d}} \sum_{i=1}^m \|x_i - BB^T x_i\|_2^2$$

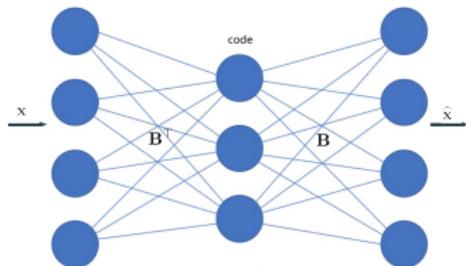
So the autoencoder is performing PCA!

One can even relax the weight tying:

$$\min_{B \in \mathbb{R}^{D \times d}, A \in \mathbb{R}^{d \times D}} \sum_{i=1}^m \|x_i - BA^T x_i\|_2^2,$$

which finds a basis (**not necessarily orthonormal**)  $B$  that spans the top singular space also [Baldi and Hornik, 1989], [Kawaguchi, 2016], [Lu and Kawaguchi, 2017].

# Factorization



To perform PCA,

$$\min_{B \in \mathbb{R}^{D \times d}} \sum_{i=1}^m \|\mathbf{x}_i - BB^T \mathbf{x}_i\|_2^2$$
$$\min_{B \in \mathbb{R}^{D \times d}, A \in \mathbb{R}^{d \times D}} \sum_{i=1}^m \|\mathbf{x}_i - BA^T \mathbf{x}_i\|_2^2,$$

But: the basis  $B$  and the representations/codes  $\mathbf{z}_i$ 's are all we care about

**Factorization:** (or autoencoder without encoder)

$$\min_{B \in \mathbb{R}^{D \times d}, \mathbf{Z} \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2.$$

All three formulations will find three **different**  $B$ 's that span the **same** principal subspace [Tan and Mayrovouniotis, 1995, Li et al., 2020b, Li et al., 2020a, Valavi et al., 2020]. They're all doing PCA!

# Sparse coding

**Factorization:** (or autoencoder without encoder)

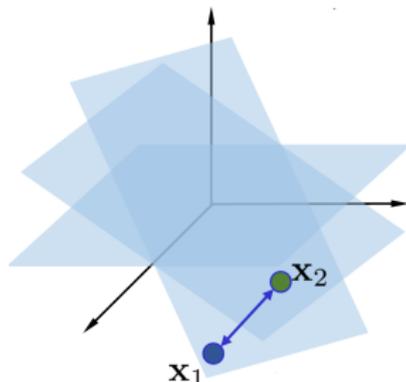
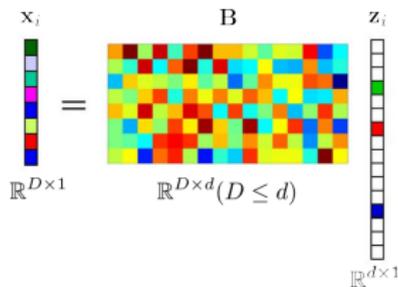
$$\min_{B \in \mathbb{R}^{D \times d}, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2.$$

What happens when we allow  $d \geq D$ ? Underdetermined even if  $B$  is known.

**Sparse coding:** assuming  $\mathbf{z}_i$ 's are sparse and  $d \geq D$

$$\min_{B \in \mathbb{R}^{D \times d}, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2 + \lambda \sum_{i=1}^m \Omega(\mathbf{z}_i)$$

where  $\Omega$  promotes sparsity, e.g.,  $\Omega = \|\cdot\|_1$ .



Letter | Published: 13 June 1996

## Emergence of simple-cell receptive field properties by learning a sparse code for natural images

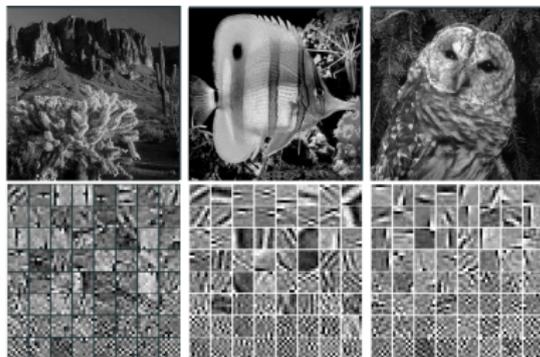
Bruno A. Olshausen & David J. Field

Nature 381, 607–609(1996) | [Cite this article](#)

5409 Accesses | 2901 Citations | 29 Altmetric | [Metrics](#)

### Abstract

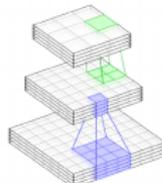
THE receptive fields of simple cells in mammalian primary visual cortex can be characterized as being spatially localized, oriented<sup>1-4</sup> and bandpass (selective to structure at different spatial scales), comparable to



denoising



super resol.



recognition

also known as (sparse) dictionary learning [Olshausen and Field, 1996, Mairal, 2014, Sun et al., 2017, Bai et al., 2018, Qu et al., 2019]

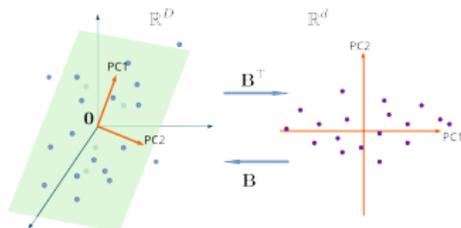
PCA for linear data

Extensions of PCA for nonlinear data

Application examples

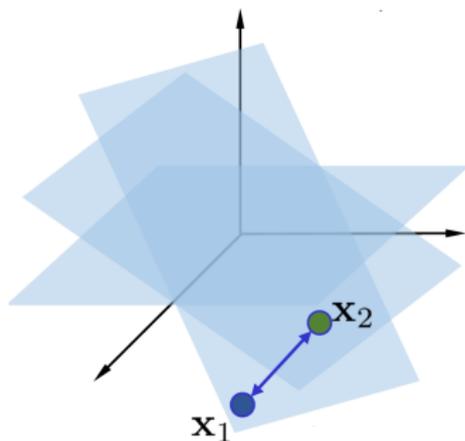
Suggested reading

# Quick summary of the linear models



PCA is effectively to identify  
the best-fit subspace to

$$\mathbf{x}_1, \dots, \mathbf{x}_m$$



–  $B$  from  $U$  of  $X = USV^T$

– autoencoder:

$$\min_{B \in \mathbb{R}^{D \times d}} \sum_{i=1}^m \|\mathbf{x}_i - BB^T \mathbf{x}_i\|_2^2$$

– autoencoder:

$$\min_{B \in \mathbb{R}^{D \times d}, A \in \mathbb{R}^{d \times D}} \sum_{i=1}^m \|\mathbf{x}_i - BA^T \mathbf{x}_i\|_2^2$$

– factorization:

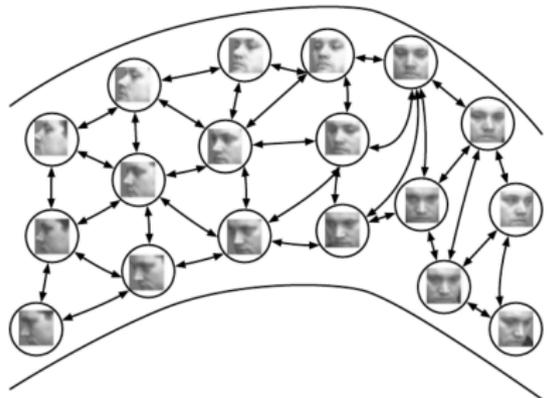
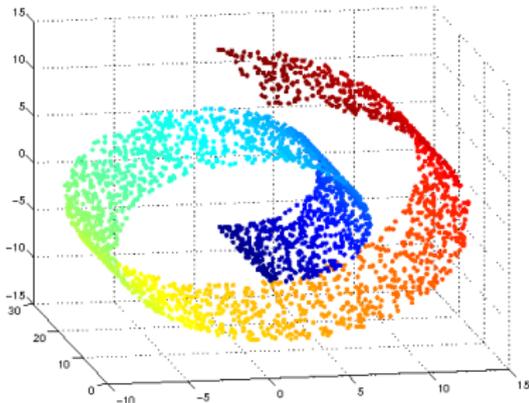
$$\min_{B \in \mathbb{R}^{D \times d}, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - Bz_i\|_2^2$$

– when  $d \geq D$ , sparse coding/dictionary learning

$$\min_{B \in \mathbb{R}^{D \times d}, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - Bz_i\|_2^2 + \lambda \sum_{i=1}^m \Omega(z_i)$$

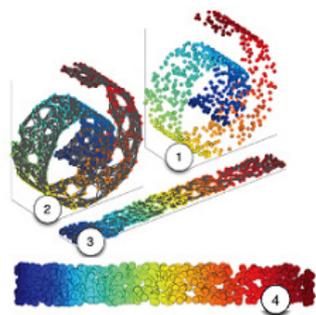
e.g.,  $\Omega = \|\cdot\|_1$

# What about nonlinear data?



- Manifold, but not mathematically (i.e., differential geometry sense) rigorous
- **Working hypothesis for high-dimensional data:** practical data lie (approximately) on union of **low-dimensional** “manifolds”. Why?
  - \* data generating processes often controlled by very few parameters

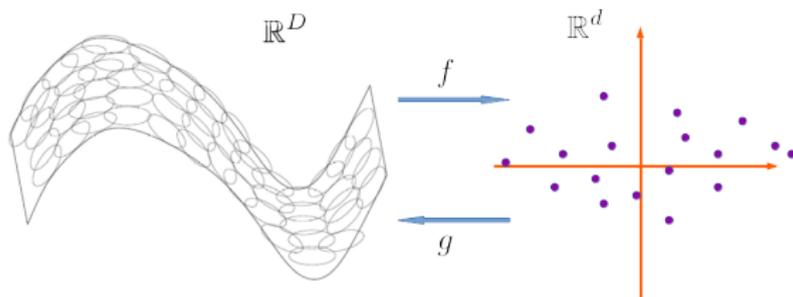
# Manifold learning



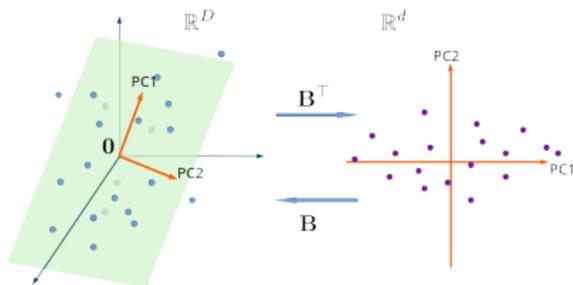
Classic methods (mostly for visualization): .e.g.,

- ISOMAP [Tenenbaum, 2000]
- Locally-Linear Embedding [Roweis, 2000]
- Laplacian eigenmap [Belkin and Niyogi, 2001]
- t-distributed stochastic neighbor embedding (t-SNE) [van der Maaten and Hinton, 2008]

Nonlinear dimension reduction and representation learning:

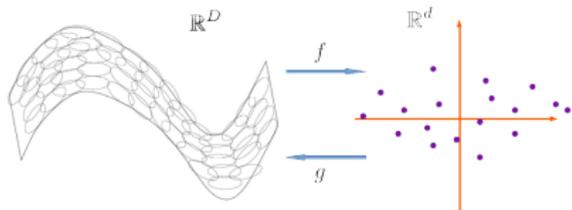


# From autoencoders to deep autoencoders



$$\min_{B \in \mathbb{R}^{D \times d}} \sum_{i=1}^m \|\mathbf{x}_i - BB^T \mathbf{x}_i\|_2^2$$
$$\min_{B \in \mathbb{R}^{D \times d}, A \in \mathbb{R}^{d \times D}} \sum_{i=1}^m \|\mathbf{x}_i - BA^T \mathbf{x}_i\|_2^2$$

nonlinear generalization of the linear mappings:



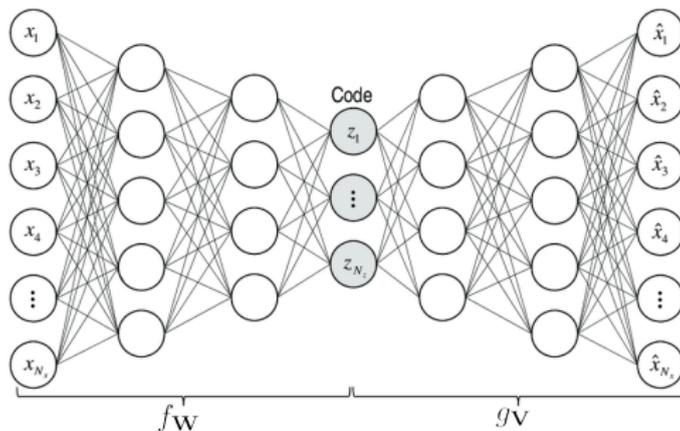
**deep autoencoders**

$$\min_{V, W} \sum_{i=1}^m \|\mathbf{x}_i - g_V \circ f_W(\mathbf{x}_i)\|_2^2$$

simply  $A^T \rightarrow f_W$  and  $B \rightarrow g_V$

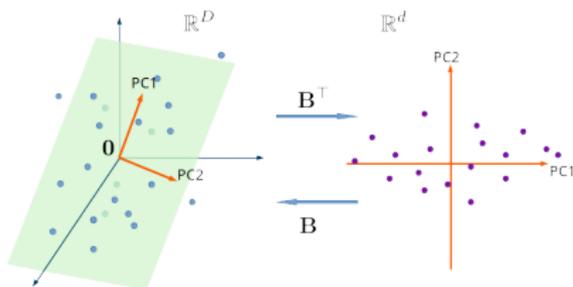
# Deep autoencoders

$$\min_{\mathbf{V}, \mathbf{W}} \sum_{i=1}^m \|\mathbf{x}_i - g\mathbf{v} \circ f\mathbf{w}(\mathbf{x}_i)\|_2^2$$



the landmark paper [[Hinton, 2006](#)]

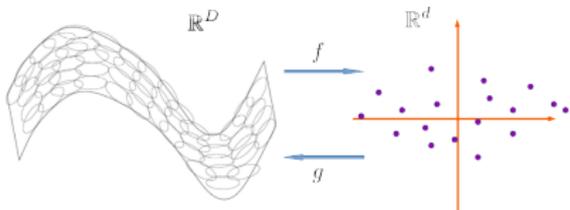
# From factorization to deep factorization



factorization

$$\min_{B \in \mathbb{R}^{D \times d}, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2$$

nonlinear generalization of the linear mappings:



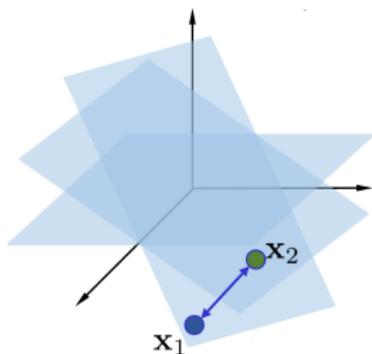
deep factorization

$$\min_{V, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - g_V(\mathbf{z}_i)\|_2^2$$

simply  $B \rightarrow g_V$

[Tan and Mayrovouniotis, 1995, Fan and Cheng, 2018, Bojanowski et al., 2017, Park et al., 2019, Li et al., 2020b], also known as **deep decoder**.

# From sparse coding to deep sparse coding



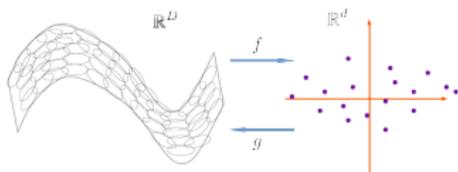
- when  $d \geq D$ , sparse coding/dictionary learning

$$\min_{B \in \mathbb{R}^{D \times d}, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - B\mathbf{z}_i\|_2^2 + \lambda \sum_{i=1}^m \Omega(\mathbf{z}_i)$$

e.g.,  $\Omega = \|\cdot\|_1$

nonlinear generalization of the linear mappings: ( $d \geq D$ )

## deep sparse coding/dictionary learning

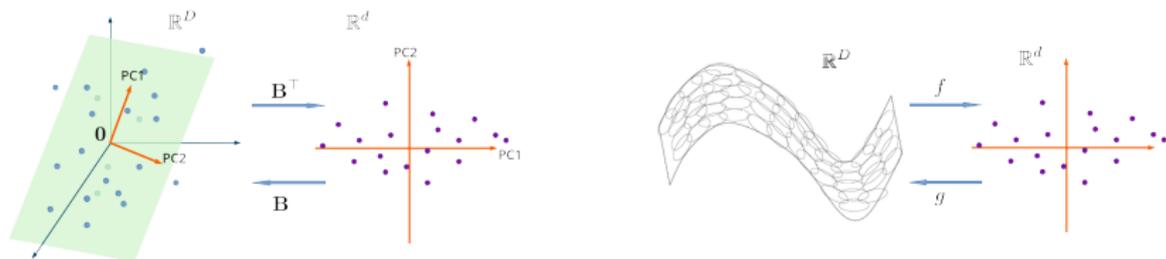


$$\min_{V, Z \in \mathbb{R}^{d \times m}} \sum_{i=1}^m \|\mathbf{x}_i - gV(\mathbf{z}_i)\|_2^2 + \lambda \sum_{i=1}^m \Omega(\mathbf{z}_i)$$

$$\min_{V, W} \sum_{i=1}^m \|\mathbf{x}_i - gV \circ fW(\mathbf{x}_i)\|_2^2 + \sum_{i=1}^m \Omega(fW(\mathbf{x}_i))$$

the 2nd also called **sparse autoencoder** [Ranzato et al., 2006].

# Quick summary of linear vs nonlinear models



	linear models	nonlinear models
autoencoder	$\min_B \sum_{i=1}^m \ell(\mathbf{x}_i, \mathbf{B}\mathbf{B}^T \mathbf{x}_i)$ $\min_{B,A} \sum_{i=1}^m \ell(\mathbf{x}_i, \mathbf{B}\mathbf{A}^T \mathbf{x}_i)$	$\min_{V,W} \sum_{i=1}^m \ell(\mathbf{x}_i, g_V \circ f_W(\mathbf{x}_i))$
factorization	$\min_{B,Z} \sum_{i=1}^m \ell(\mathbf{x}_i, \mathbf{B}\mathbf{z}_i)$	$\min_{V,Z} \sum_{i=1}^m \ell(\mathbf{x}_i, g_V(\mathbf{z}_i))$
sparse coding	$\min_{B,Z} \sum_{i=1}^m \ell(\mathbf{x}_i, \mathbf{B}\mathbf{z}_i)$ $+\lambda \sum_{i=1}^m \Omega(\mathbf{z}_i)$	$\min_{V,Z} \sum_{i=1}^m \ell(\mathbf{x}_i, g_V(\mathbf{z}_i))$ $+\lambda \sum_{i=1}^m \Omega(\mathbf{z}_i)$ $\min_{V,W} \sum_{i=1}^m \ell(\mathbf{x}_i, g_V \circ f_W(\mathbf{x}_i))$ $+\lambda \sum_{i=1}^m \Omega(f_W(\mathbf{x}_i))$

$\ell$  can be general loss functions other than  $\|\cdot\|_2$

$\Omega$  promotes sparsity, e.g.,  $\Omega = \|\cdot\|_1$

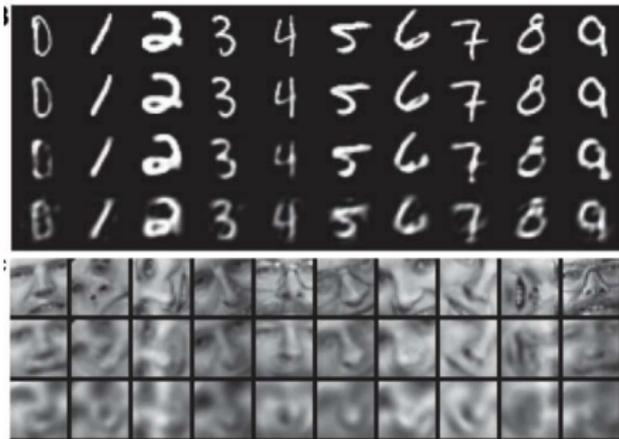
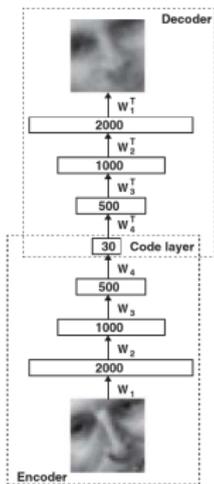
PCA for linear data

Extensions of PCA for nonlinear data

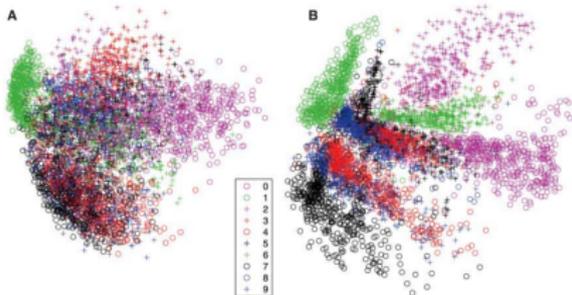
Application examples

Suggested reading

# Nonlinear dimension reduction



autoencoder vs. PCA vs. logistic PCA



[Hinton, 2006]

# Representation learning

## Traditional learning pipeline

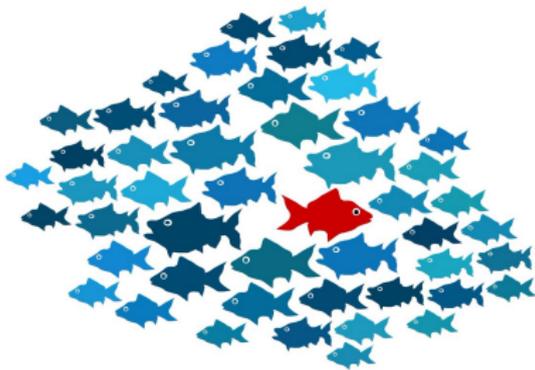


- feature extraction is “independent” of the learning models and tasks
- features are handcrafted and/or learned

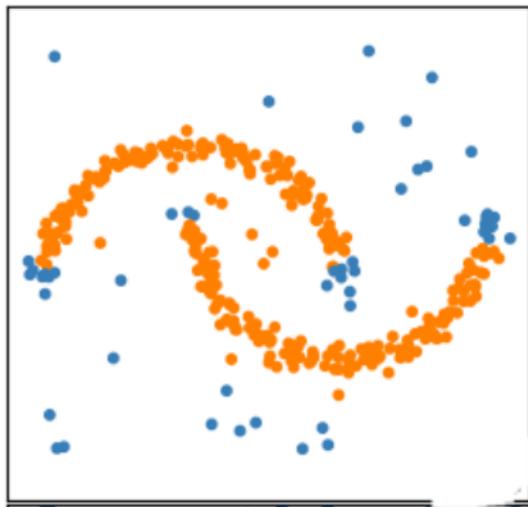
## Use the low-dimensional codes as features/representations

- task agnostic
- less overfitting
- semi-supervised (rich unlabeled data + little labeled data) learning

# Outlier detection

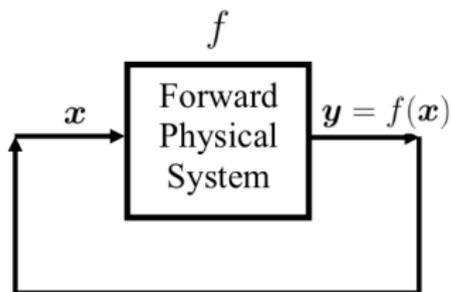


(Credit: towardsdatascience.com)



- idea: outliers don't obey the manifold assumption — the reconstruction error  $\ell(\mathbf{x}_i, g_V \circ f_W(\mathbf{x}_i))$  is large after autoencoder training
- for effective detection, better use  $\ell$  that penalizes large errors less harshly than  $\|\cdot\|_2^2$ , e.g.,  $\ell(\mathbf{x}_i, g_V \circ f_W(\mathbf{x}_i)) = \|\mathbf{x}_i - g_V \circ f_W(\mathbf{x}_i)\|_2$   
[Lai et al., 2019]

# Deep generative prior



Inverse Problem  $f^{-1}$

- **inverse problems:** given  $f$  and  $\mathbf{y} = f(\mathbf{x})$ , estimate  $\mathbf{x}$
- often ill-posed, i.e.,  $\mathbf{y}$  doesn't contain enough info for recovery
- regularized formulation:

$$\min_{\mathbf{x}} \ell(\mathbf{y}, f(\mathbf{x})) + \lambda \Omega(\mathbf{x})$$

where  $\Omega$  contains extra info about  $\mathbf{x}$

Suppose  $\mathbf{x}_1, \dots, \mathbf{x}_m$  come from the same manifold as  $\mathbf{x}$

- train a deep factorization model on  $\mathbf{x}_1, \dots, \mathbf{x}_m$ :

$$\min_{\mathbf{V}, \mathbf{Z}} \sum_{i=1}^m \ell(\mathbf{x}_i, g_{\mathbf{V}}(\mathbf{z}_i))$$

- $\mathbf{x} \approx g_{\mathbf{V}}(\mathbf{z})$  for a certain  $\mathbf{z}$  so:  $\min_{\mathbf{z}} \ell(\mathbf{y}, f \circ g_{\mathbf{V}}(\mathbf{z}))$ . Some recent work even uses random  $\mathbf{V}$ , i.e., without training

[Ulyanov et al., 2018, Bora and Dimakis, 2017]

## To be covered later

- convolutional encoder-decoder networks (i.e., segmentation, image processing, inverse problems)
- autoencoder sequence-to-sequence models (e.g., machine translation)
- variational autoencoders (generative models)

PCA for linear data

Extensions of PCA for nonlinear data

Application examples

Suggested reading

## Suggested reading

- Representation Learning: A Review and New Perspectives (Bengio, Y., Courville, A., and Vincent, P.) [[Bengio et al., 2013](#)]
- Chaps 13–15 of Deep Learning [[Goodfellow et al., 2017](#)].
- Rethink autoencoders: Robust manifold learning [[Li et al., 2020b](#)]

- [Bai et al., 2018] Bai, Y., Jiang, Q., and Sun, J. (2018). **Subgradient descent learns orthogonal dictionaries.** *arXiv:1810.10702*.
- [Baldi and Hornik, 1989] Baldi, P. and Hornik, K. (1989). **Neural networks and principal component analysis: Learning from examples without local minima.** *Neural Networks*, 2(1):53–58.
- [Belkin and Niyogi, 2001] Belkin, M. and Niyogi, P. (2001). **Laplacian eigenmaps and spectral techniques for embedding and clustering.** In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pages 585–591. MIT Press.
- [Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). **Representation learning: A review and new perspectives.** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.
- [Bojanowski et al., 2017] Bojanowski, P., Joulin, A., Lopez-Paz, D., and Szlam, A. (2017). **Optimizing the latent space of generative networks.** *arXiv:1707.05776*.

- [Bora and Dimakis, 2017] Bora, Ashish, A. J. E. P. and Dimakis, A. G. (2017). **Compressed sensing using generative models.** In *Proceedings of the 34th International Conference on Machine Learning*, volume 70.
- [Bourlard and Kamp, 1988] Bourlard, H. and Kamp, Y. (1988). **Auto-association by multilayer perceptrons and singular value decomposition.** *Biological Cybernetics*, 59(4-5):291–294.
- [Elgendy, 2020] Elgendy, M. (2020). **Deep Learning for Vision Systems.** MANNING PUBN.
- [Fan and Cheng, 2018] Fan, J. and Cheng, J. (2018). **Matrix completion by deep matrix factorization.** *Neural Networks*, 98:34–41.
- [Goodfellow et al., 2017] Goodfellow, I., Bengio, Y., and Courville, A. (2017). **Deep Learning.** The MIT Press.
- [Hinton, 2006] Hinton, G. E. (2006). **Reducing the dimensionality of data with neural networks.** *Science*, 313(5786):504–507.
- [Hinton et al., 2006] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). **A fast learning algorithm for deep belief nets.** *Neural Computation*, 18(7):1527–1554.

- [Hinton and Zemel, 1994] Hinton, G. E. and Zemel, R. S. (1994). **Autoencoders, minimum description length and helmholtz free energy.** In *Advances in neural information processing systems*, pages 3–10.
- [Kawaguchi, 2016] Kawaguchi, K. (2016). **Deep learning without poor local minima.** *arXiv:1605.07110*.
- [Lai et al., 2019] Lai, C.-H., Zou, D., and Lerman, G. (2019). **Robust subspace recovery layer for unsupervised anomaly detection.** *arXiv:1904.00152*.
- [Li et al., 2020a] Li, S., Li, Q., Zhu, Z., Tang, G., and Wakin, M. B. (2020a). **The global geometry of centralized and distributed low-rank matrix recovery without regularization.** *arXiv:2003.10981*.
- [Li et al., 2020b] Li, T., Mehta, R., Qian, Z., and Sun, J. (2020b). **Rethink autoencoders: Robust manifold learning.** Submitted to European Conference on Computer Vision.
- [Lu and Kawaguchi, 2017] Lu, H. and Kawaguchi, K. (2017). **Depth creates no bad local minima.** *arXiv:1702.08580*.
- [Mairal, 2014] Mairal, J. (2014). **Sparse modeling for image and vision processing.** *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283.

- [Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). **Emergence of simple-cell receptive field properties by learning a sparse code for natural images.** *Nature*, 381(6583):607–609.
- [Park et al., 2019] Park, J. J., Florence, P., Straub, J., Newcombe, R., and Lovegrove, S. (2019). **Deepspf: Learning continuous signed distance functions for shape representation.** pages 165–174. IEEE.
- [Qu et al., 2019] Qu, Q., Zhai, Y., Li, X., Zhang, Y., and Zhu, Z. (2019). **Analysis of the optimization landscapes for overcomplete representation learning.** *arXiv:1912.02427*.
- [Ranzato et al., 2006] Ranzato, M., Poultney, C. S., Chopra, S., and LeCun, Y. (2006). **Efficient learning of sparse representations with an energy-based model.** In *Advances in Neural Information Processing Systems*.
- [Roweis, 2000] Roweis, S. T. (2000). **Nonlinear dimensionality reduction by locally linear embedding.** *Science*, 290(5500):2323–2326.
- [Sun et al., 2017] Sun, J., Qu, Q., and Wright, J. (2017). **Complete dictionary recovery over the sphere i: Overview and the geometric picture.** *IEEE Transactions on Information Theory*, 63(2):853–884.

- [Tan and Mayrovouniotis, 1995] Tan, S. and Mayrovouniotis, M. L. (1995). **Reducing data dimensionality through optimizing neural network inputs.** *AIChE Journal*, 41(6):1471–1480.
- [Tenenbaum, 2000] Tenenbaum, J. B. (2000). **A global geometric framework for nonlinear dimensionality reduction.** *Science*, 290(5500):2319–2323.
- [Ulyanov et al., 2018] Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). **Deep image prior.** In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9446–9454.
- [Valavi et al., 2020] Valavi, H., Liu, S., and Ramadge, P. J. (2020). **The landscape of matrix factorization revisited.** *arXiv:2002.12795*.
- [van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). **Visualizing data using t-sne.** *Journal of Machine Learning Research*, 9:2579–2605.