

Od rozkładu Gaussa do autoenkoderów wariacyjnych

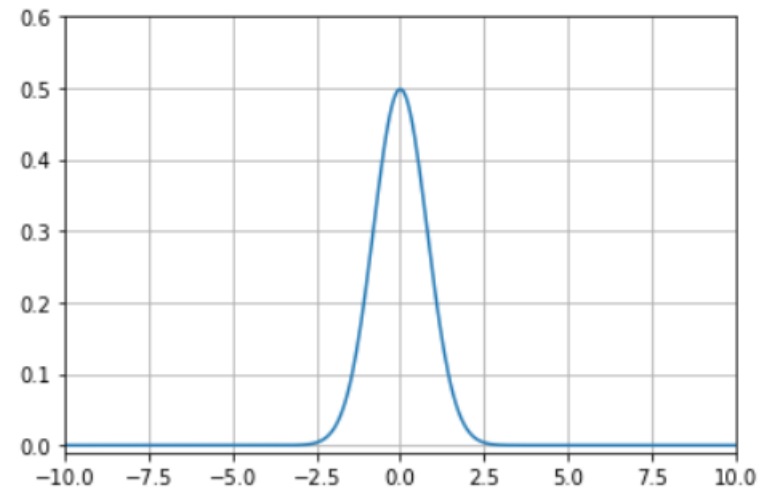
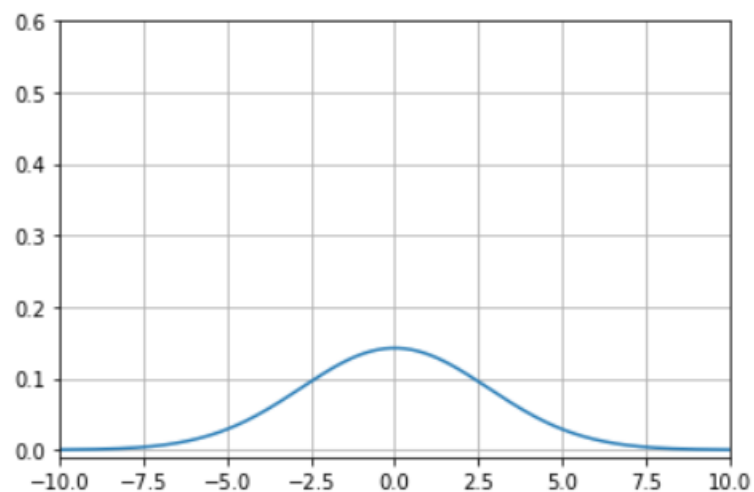
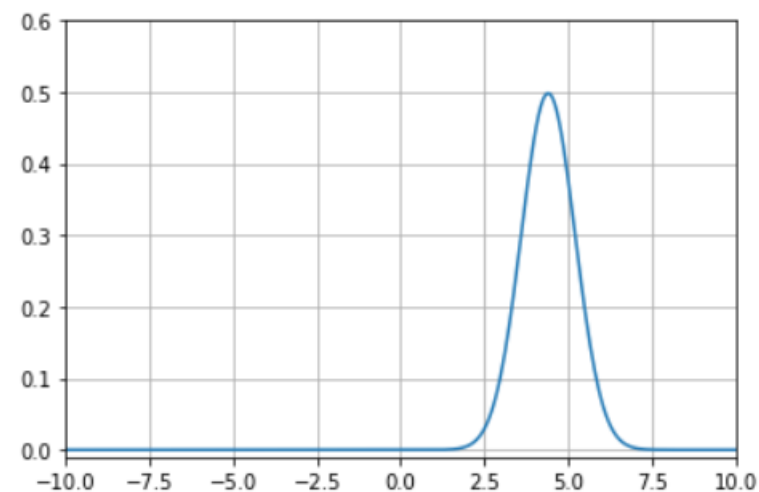
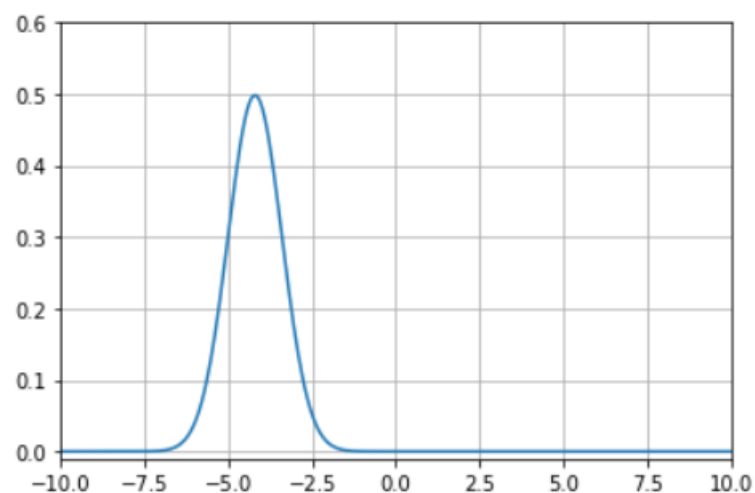
Arkadiusz Kwasigroch

Plan

- Estymacja parametrów rozkładu
- Regresja
- Regresja logistyczna
- PCA
- Autoenkodery
- Probabilistyczne PCA
- Autoenkodery wariacyjne

Rozkład Gaussa

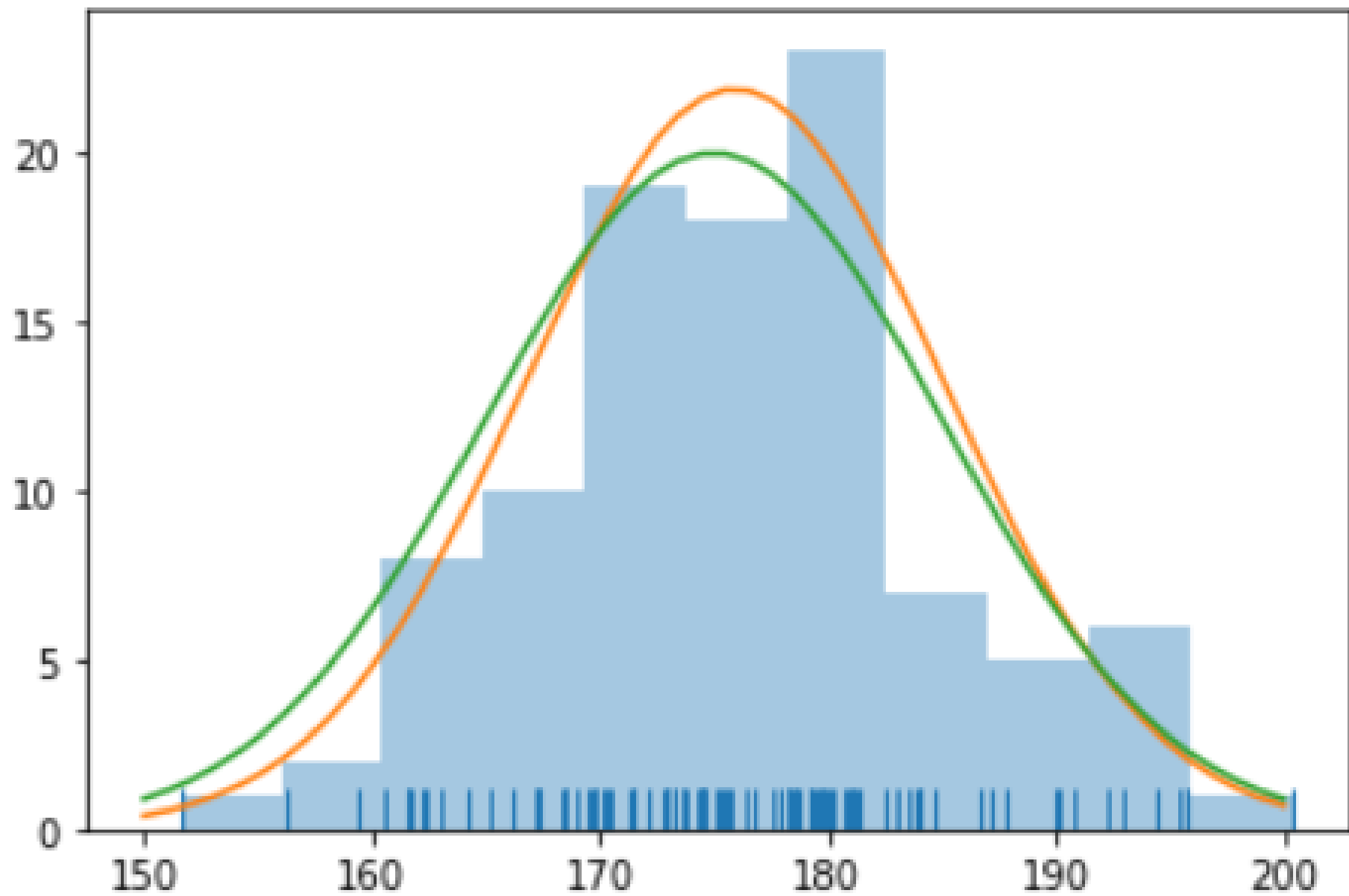
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$



W jaki sposób wyznaczyć parametry rozkładu

1. Dany jest wektor obserwacji \mathbf{X} (np. wzrost poszczególnych osób w grupie)
2. Zakładamy, że wzrost jest zmienną losową o rozkładzie normalnym (Gausa), których parametrów μ i σ nie znamy
3. Zakładamy, że prawdopodobieństwo pojawienia się osoby o określonym wzroście wynosi:

$$p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2)$$



W jaki sposób wyznaczyć parametry rozkładu

4. Szukamy łącznego prawdopodobieństwa wystąpienia wszystkich zdarzeń – wylosowania grupy osób

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3)p(x_4) \quad \longrightarrow \quad \begin{array}{l} \text{Tylko dla IID} \\ \text{Independent and} \\ \text{identically} \\ \text{distributed} \end{array}$$

Co w naszym przypadku prowadzi do następującego wzoru

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

W jaki sposób wyznaczyć parametry rozkładu

Otrzymana funkcja nazywa się funkcją wiarygodności (likelihood function)

Funkcja wiarygodności (wiarygodność) – w statystyce, funkcja **parametru modelu i próby losowej**, która jest proporcjonalna do **prawdopodobieństwa zaobserwowania próby o konkretnej postaci** przy różnych parametrach modelu. Wyraża „wiarygodność” wartości parametru w obliczu danych.

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

W jaki sposób wyznaczyć parametry rozkładu

- In statistics, the **likelihood function** (often simply called the **likelihood**) measures the **goodness of fit** of a statistical model to a sample of data for given values of the **unknown parameters**. It is formed from the **joint probability distribution** of the sample, but viewed and used as a **function of the parameters only**, thus treating the random variables as fixed at the observed values

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

W jaki sposób wyznaczyć parametry rozkładu

Chcemy aby nasze dopasowanie było jak najlepsze, dlatego funkcję wiarygodności maksymalizujemy. Procedura uzyskania parametrów, które maksymalizują tę funkcję nazywamy:

Maximum Likelihood Estimation (MLE)

$$\max_{\mu, \sigma^2} p(\mathbf{X} \mid \mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n \mid \mu, \sigma^2)$$

W jaki sposób wyznaczyć parametry rozkładu

Uzyskana forma funkcji wiarygodności jest podatna na niestabilność numeryczną. Niskie wartości prawdopodobieństwa mogą prowadzić do bardzo małych liczb bliskich zera. Aby temu zapobiec korzystamy z właściwości logarytmów:

$$\log(ab) = \log(a) + \log(b)$$

Logarytm jest monotoniczną funkcją rosnącą, więc maksymalizacja logarytmu danej funkcji jest równoznaczna maksymalizacji samej funkcji

$$\arg \max_x \log(f(x)) = \arg \max_x f(x)$$

W jaki sposób wyznaczyć parametry rozkładu

Wykorzystując opisane własności możemy zapisać:

$$\log p(\mathbf{X}|\mu, \sigma) = \sum_{n=1}^N \log \mathcal{N}(x_n|\mu, \sigma)$$

W jaki sposób wyznaczyć parametry rozkładu

Po podstawieniu równania rozkładu Gaussowskiego i uproszczeniu otrzymujemy:

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

Uzyskanie parametrów maksymalizujących równanie:

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

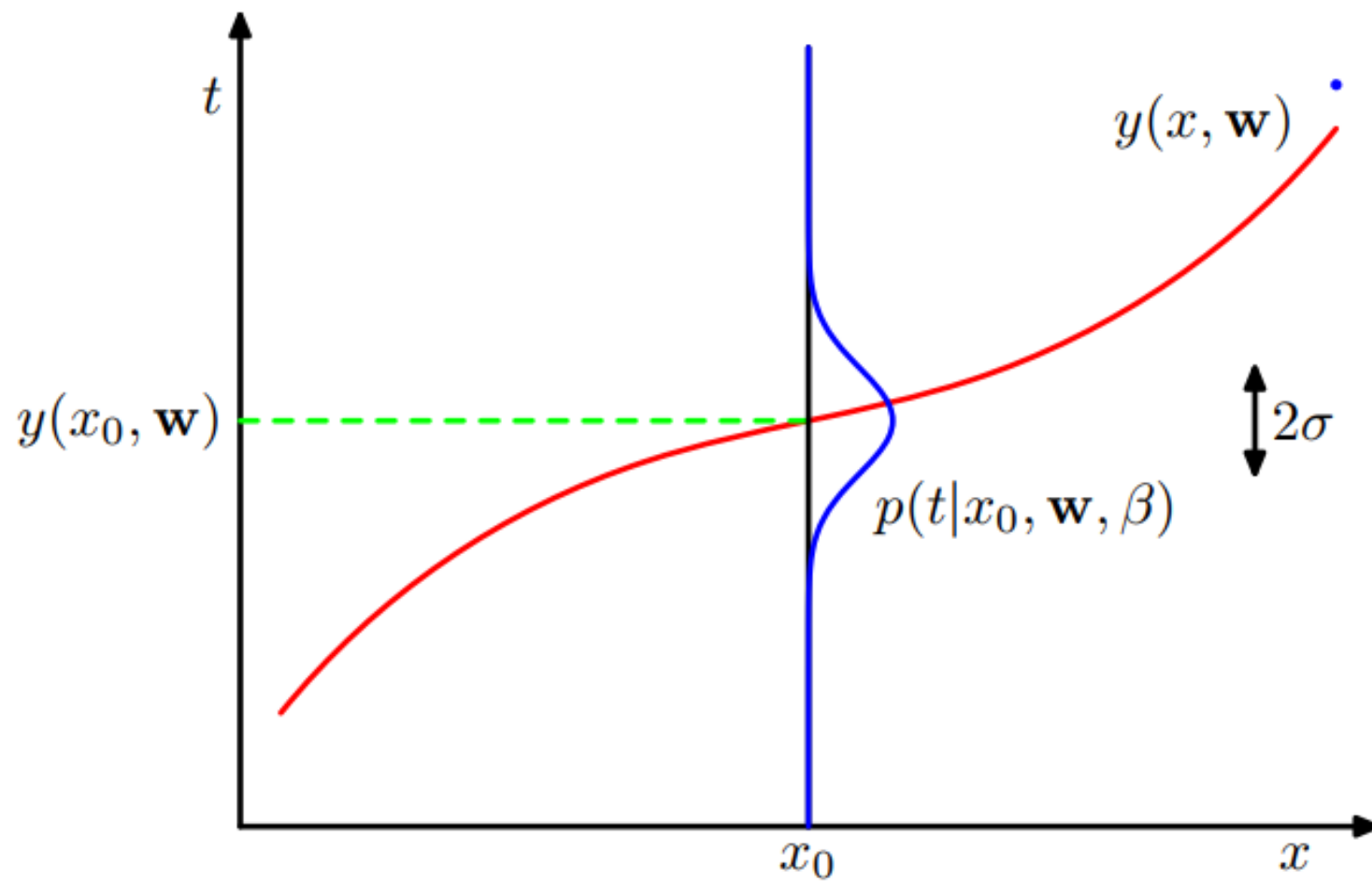
Regresja

Funkcja wiarygodności jest pojęciem ogólnym, stosowanym do dużego zakresu różnych algorytmów.

Regresję możemy przedstawić jako model probabilistyczny:

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N} \left(t|y(x, \mathbf{w}), \beta^{-1} \right)$$

Regresja



Regresja– funkcja wiarygodności

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | y(x_n, \mathbf{w}), \beta^{-1})$$

Podstawiając równanie na rozkład Gaussa oraz logarytmując dwie strony funkcji wiarygodności:

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\cancel{\frac{\beta}{2}} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 - \cancel{\frac{N}{2} \ln \beta} - \cancel{\frac{N}{2} \ln(2\pi)}.$$

$$\beta = \frac{1}{\sigma^2}$$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

Regresja – funkcja wiarygodności

Aby uzyskać parametr β maksymalizujący funkcję wiarygodności korzystamy z równania

$$\frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{\text{ML}}) - t_n\}^2$$

Regresja– ostateczny model

Ostatecznie otrzymujemy „rozkład predykcyjny” (*predictive distribution*)

$$p(t|x, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N} \left(t | y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1} \right)$$

Regresja logistyczna

Model probabilistyczny dwuklasowej regresji logistycznej wyrażony jest wzorami:

$$p(\mathcal{C}_1|\phi) = y(\phi) = \sigma(\mathbf{w}^T \phi)$$

$$p(\mathcal{C}_2|\phi) = 1 - p(\mathcal{C}_1|\phi)$$

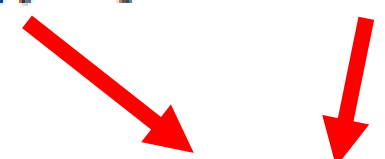
Gdzie:

$$\sigma(\cdot)$$

jest funkcją sigmoidalną

Regresja logistyczna – funkcja wiarygodności

Funkcja wiarygodności dla takiego modelu wygląda następująco

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^N y_n^{t_n} \{1 - y_n\}^{1-t_n}$$


The diagram consists of two red arrows pointing from the terms $y_n^{t_n}$ and $\{1 - y_n\}^{1-t_n}$ in the likelihood function to the sigmoid function $\sigma(\mathbf{w}^T \phi)$. This indicates that the predicted probability y_n is modeled by the sigmoid function applied to the linear combination of features.

$$\sigma(\mathbf{w}^T \phi)$$

Regresja logistyczna – funkcja wiarygodności

Po zlogarytmowaniu i uproszczeniu, otrzymujemy dobrze znaną funkcję nazywaną „**binary cross entropy**”

$$-\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

Funkcja wiarygodności - podsumowanie

$$\begin{aligned}\boldsymbol{\theta}_{\text{ML}} &= \arg \max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X}; \boldsymbol{\theta}) \\ &= \arg \max_{\boldsymbol{\theta}} \prod_{i=1}^m p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})\end{aligned}$$

Funkcja wiarygodności może być zastosowana do rozkładów warunkowych (uczenie nadzorowane)

$$\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta})$$

PCA - Principal Component Analysis

Algorytm PCA wykorzystywany jest do redukcji wymiarowości danych

Zastosowania:

- Kompresja danych
- Ekstrakcja cech
- Wizualizacja danych

PCA - wzór

$$J(\mathbf{W}, \mathbf{Z}) = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$$

$$z = W^T x$$

$$x^* = W z$$

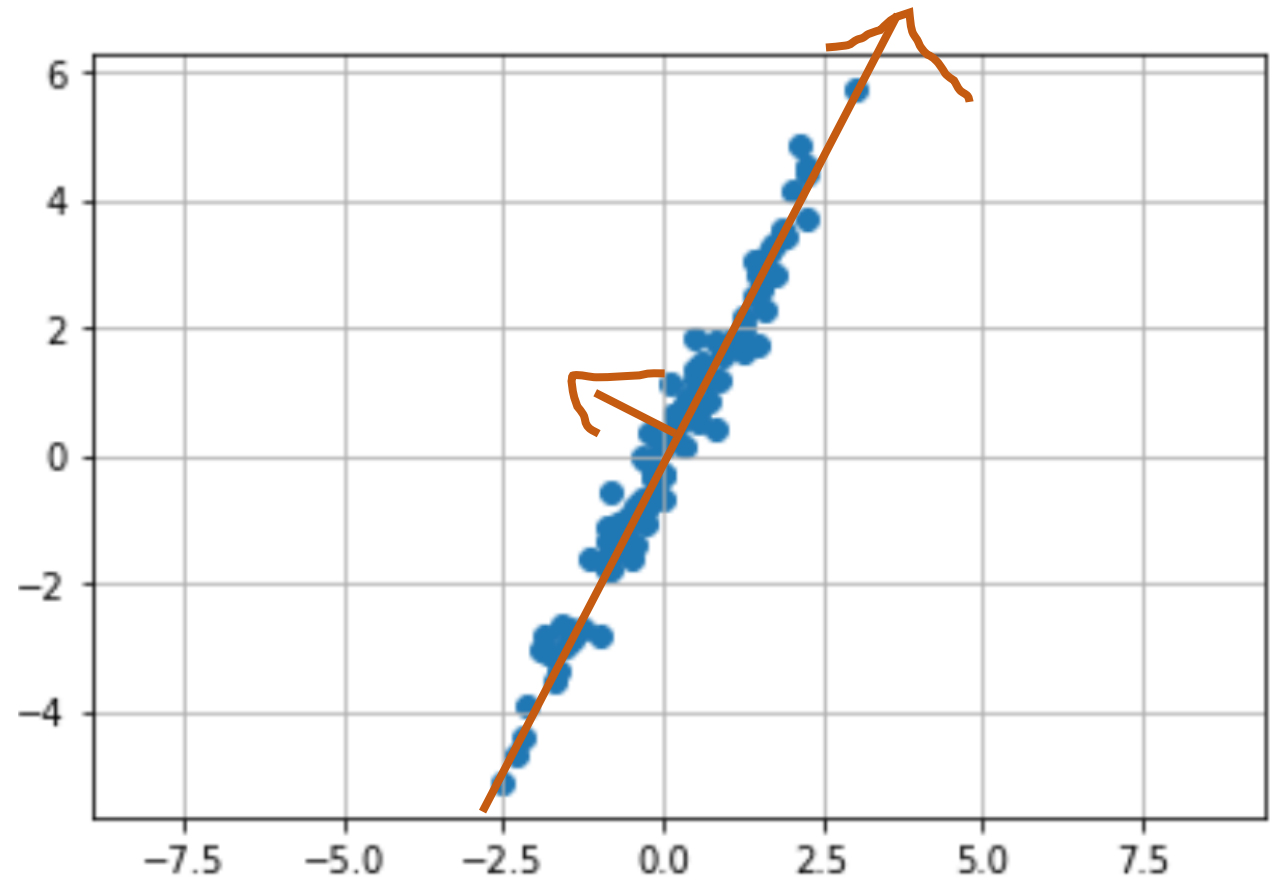
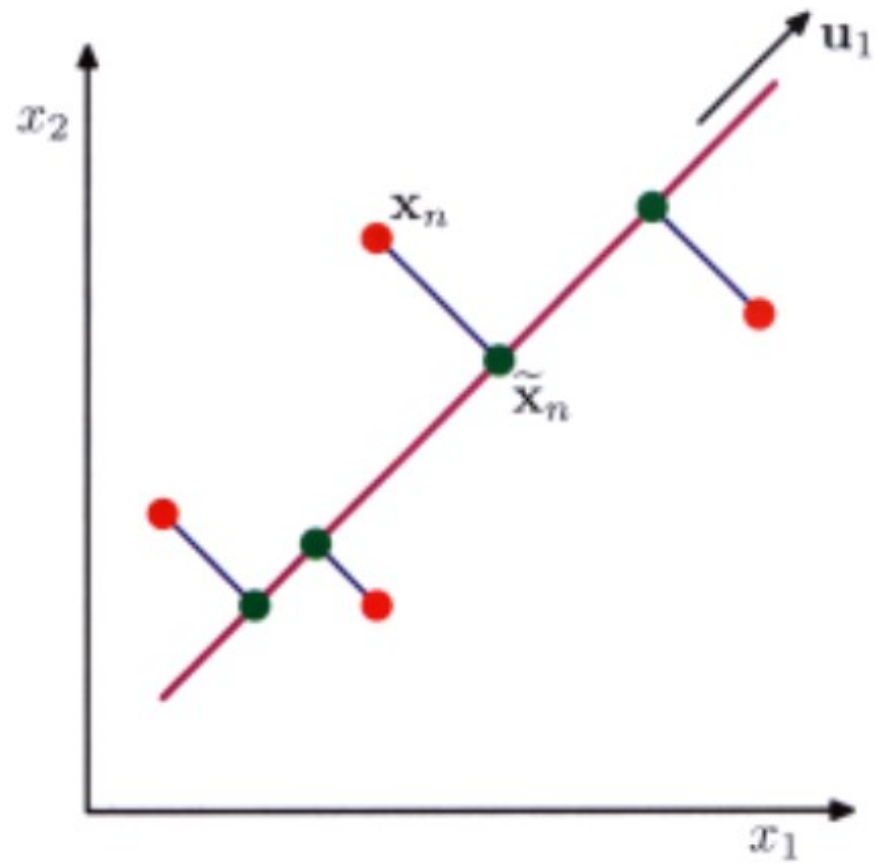


↑
Zmienna obserwowana
(*observed variable*)

↑
Zmienna ukryta
(*latent variable*)

↑
Rekonstrukcja

PCA - intuicja



PCA - rekonstrukcja

Original



$M = 1$



$M = 10$



$M = 50$

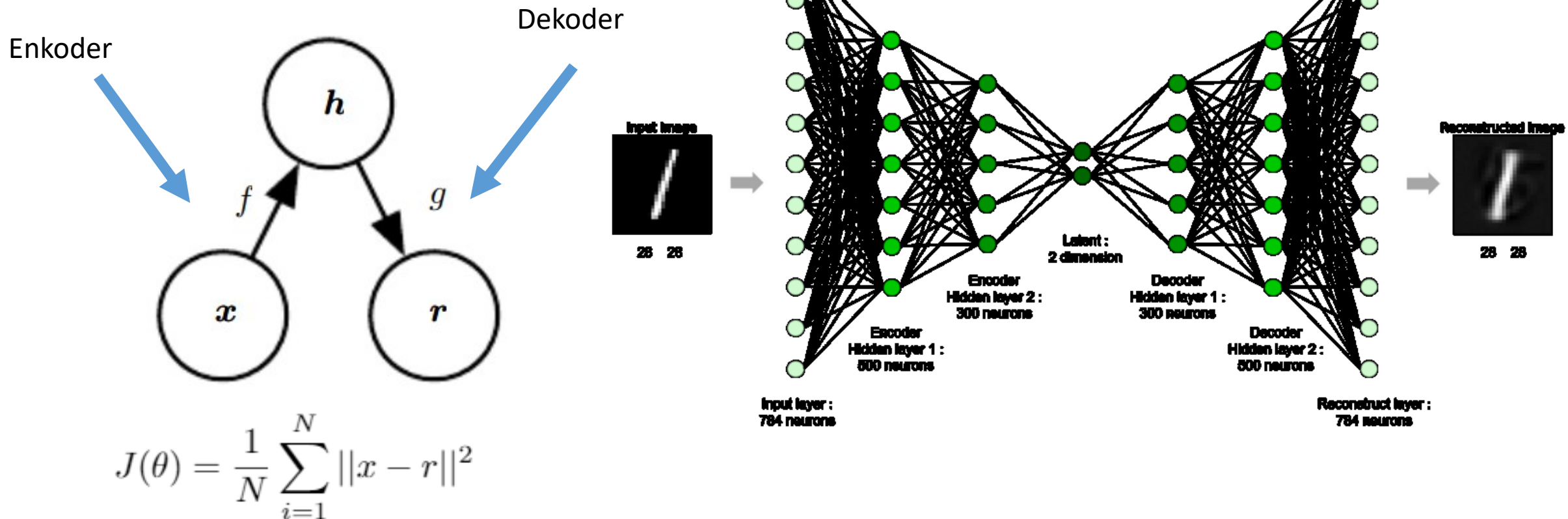


$M = 250$



Autoenkodery

Na podobnej zasadzie do PCA działają autoenkodery, które są oparte o sieci neuronowe



$$J(\theta) = \frac{1}{N} \sum_{i=1}^N ||x - r||^2$$

Autoenkodery - wykorzystanie

- Redukcja wymiarowości
- Wizualizacja
- Odszumianie (denoising autoencoders)
- Nienadzorowana ekstrakcja cech (nie potrzeba danych oetykietowanych)

Probabilistyczne PCA

Uogólnieniem algorytmu PCA, jest probabilistyczne PCA

Probabilistyczne PCA – budowa modelu

Definiujemy równanie na zmienną obserwowaną

$$x = \mathbf{W}\mathbf{z} + \mu + \epsilon$$

Ze względu na to, że ϵ jest zmienną losową, x również jest zmienną losową wyrażoną rozkładem

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I})$$

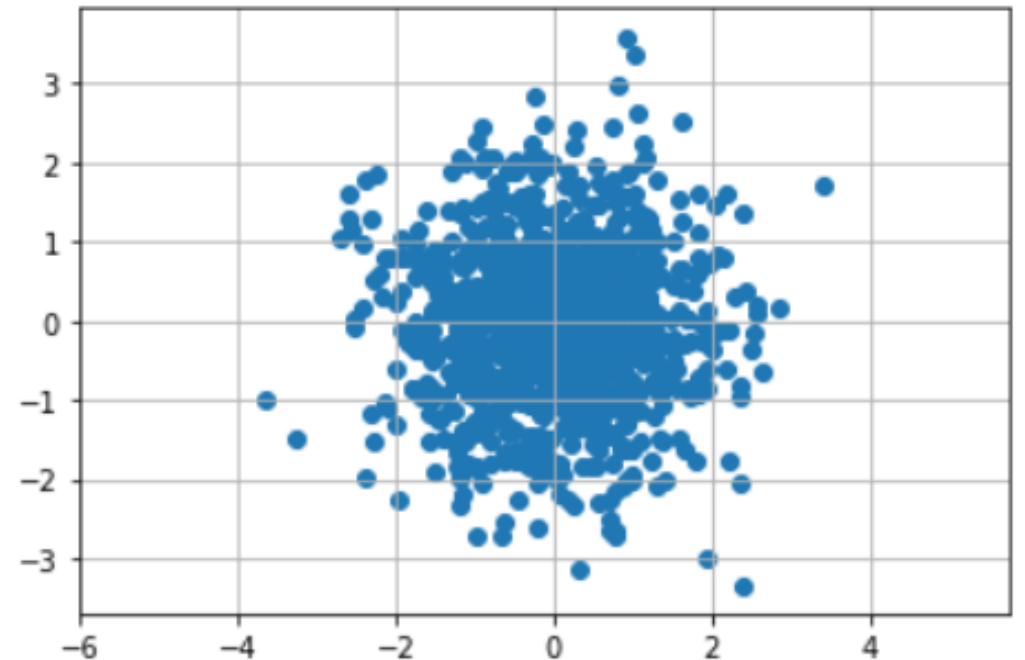


Parametry modelu

Probabilistyczne PCA – budowa modelu

Zakładamy rozkład zmiennej ukrytej z (latent variable) jako wielowymiarowy rozkład Gaussa:

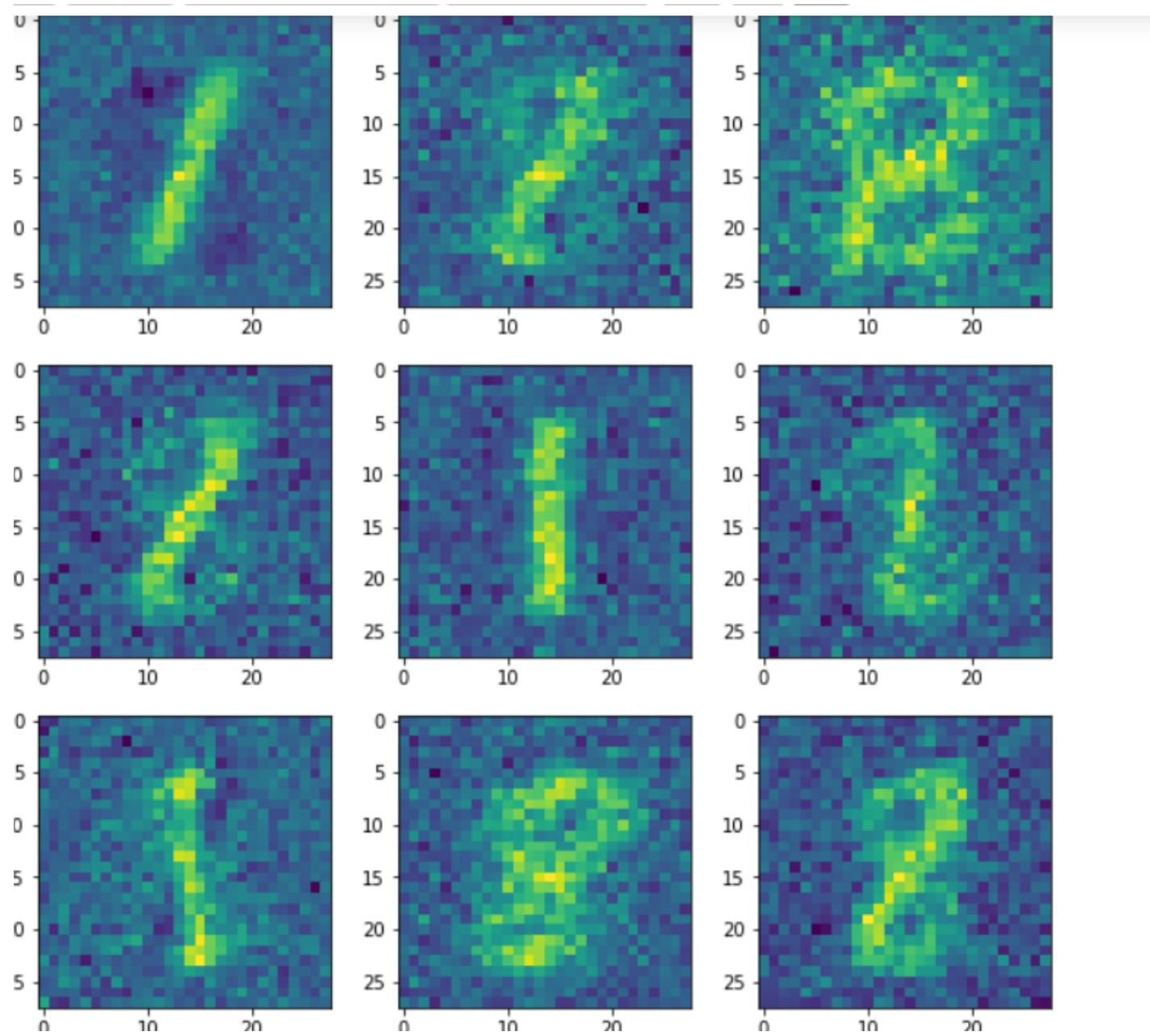
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$



Probabilistyczne PCA – model generatywny

Zdefiniowane prawdopodobieństwa umożliwiają nam generowanie przykładów z rozkładu $p(x)$, który jest rozkładem danych (data distribution)

1. Losujemy zmienną ukrytą z rozkładu $p(z)$
2. Losujemy zmienną obserwowaną z rozkładu $p(x|z)$ wykorzystując wylosowane z



Dwie zasady prawdopodobieństwa

The Rules of Probability

sum rule $p(X) = \sum_Y p(X, Y)$ (1.10)

product rule $p(X, Y) = p(Y|X)p(X).$ (1.11)

Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for \mathbf{x} and a conditional Gaussian distribution for \mathbf{y} given \mathbf{x} in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

the marginal distribution of \mathbf{y} and the conditional distribution of \mathbf{x} given \mathbf{y} are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (2.115)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (2.116)$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \quad (2.117)$$

Probabilistic PCA – zbiór rozkładów

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

Wynikające z budowy modelu

$$p(x) = \mathcal{N}(x|\mu, C)$$

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2}\mathbf{M})$$

Otrzymane z zasad sumy i
iloczynu prawdopodobieństwa

$$p(x, z) = p(x | z)p(z)$$
$$p(x) = \int p(x, z)dz = \int p(x | z)p(z)dz$$

Otrzymane z twierdzenia Bayesa

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

Probabilistic PCA – rozkład $p(x)$

W probabilistycznym PCA, możemy obliczyć funkcję wiarygodności.

Aby to zrobić należy wyznaczyć rozkład $p(x)$

Mając zdefiniowane rozkłady $p(z)$ oraz $p(x|z)$ możemy z łatwością obliczyć rozkład $p(x)$, wykorzystując zasady iloczynu i sumy prawdopodobieństwa

$$p(x, z) = p(x | z)p(z)$$

$$p(x) = \int p(x, z)dz = \int p(x | z)p(z)dz$$



Znane rozkłady Gaussa

Funkcja wiarygodności

Mając wyznaczone $p(\mathbf{x})$, możemy wyznaczyć funkcję wiarygodności dla posiadanych danych. Jak zazwyczaj, używa się logarytmu funkcji wiarygodności w celu ułatwienia obliczeń. Parametrami naszego modelu są \mathbf{W} , $\boldsymbol{\mu}$, i σ^2

$$\begin{aligned}\ln p(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) &= \sum_{n=1}^N \ln p(\mathbf{x}_n|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\ &= -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\mathbf{C}| - \frac{1}{2} \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x}_n - \boldsymbol{\mu})\end{aligned}$$

Maksymalizacja funkcji wiarygodności

Maksymalizację funkcji wiarygodności można dokonać na dwa sposoby

- Closed-form solution – obliczenie parametrów z równań

$$\boldsymbol{\mu} = \overline{\mathbf{X}}$$

$$\sigma_{\text{ML}}^2 = \frac{1}{D - M} \sum_{i=M+1}^D \lambda_i$$

$$\mathbf{W}_{\text{ML}} = \mathbf{U}_M (\mathbf{L}_M - \sigma^2 \mathbf{I})^{1/2} \mathbf{R}$$

- Expectation Maximization – algorytm iteracyjny wyznaczania parametrów modelu

Losowo wybieramy parametry modelu, następnie podczas każdej iteracji parametry są aktualizowane

PCA - przypadek szczególny PPCA

Przy $\sigma \rightarrow 0$ algorytm PPCA staje się algorytmem PCA

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z} | \mathbf{M}^{-1} \mathbf{W}^T (\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2} \mathbf{M})$$



Standardowe PCA

Algorytm EM dla PCA

$$\mathbf{\Omega} = (\mathbf{W}_{\text{old}}^T \mathbf{W}_{\text{old}})^{-1} \mathbf{W}_{\text{old}}^T \tilde{\mathbf{X}}$$

and the M step (12.56) takes the form

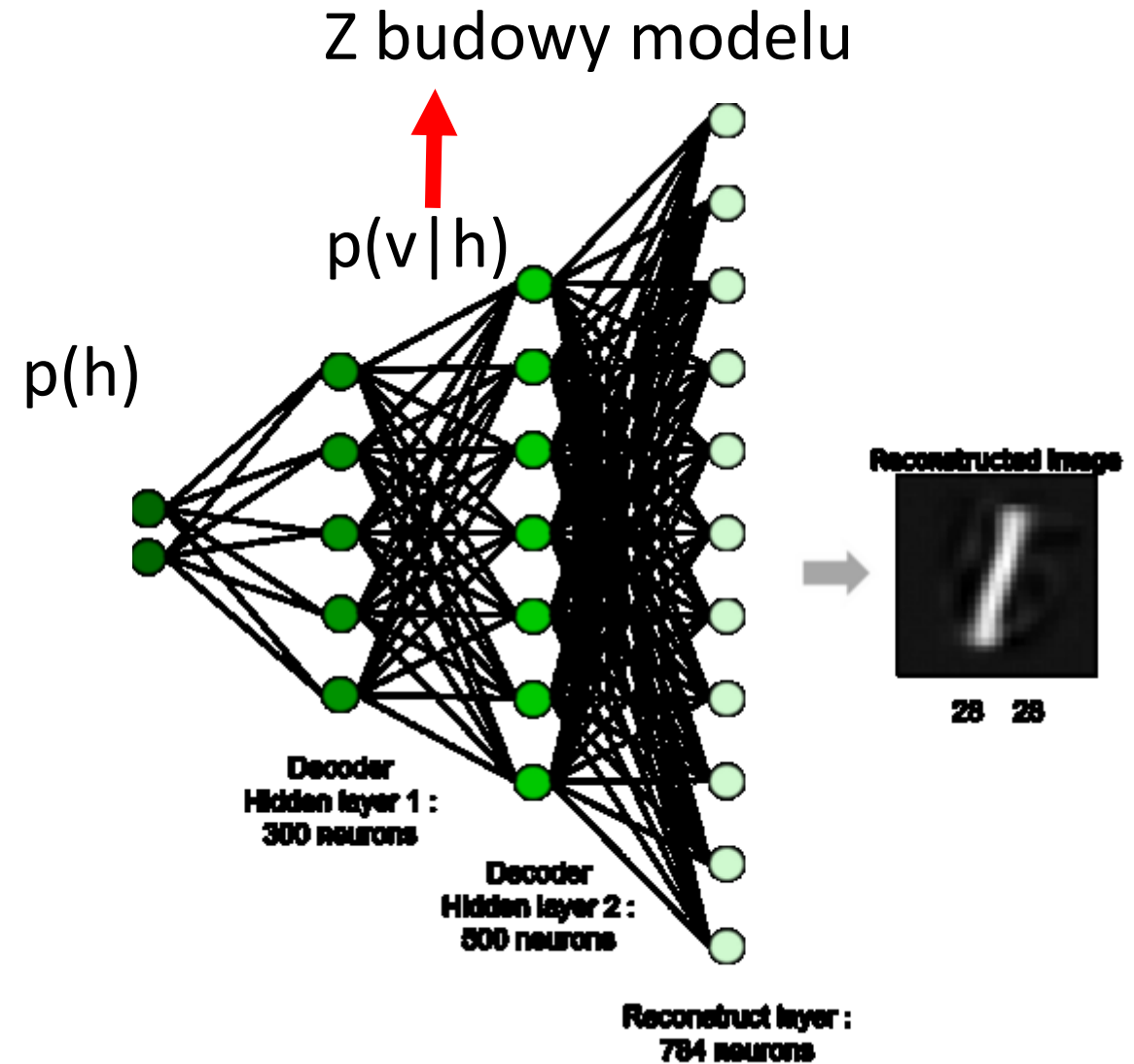
$$\mathbf{W}_{\text{new}} = \tilde{\mathbf{X}}^T \mathbf{\Omega}^T (\mathbf{\Omega} \mathbf{\Omega}^T)^{-1}.$$

Zalety PPCA

- We can derive an EM algorithm for PCA that is computationally efficient in situations where only a few leading eigenvectors are required and that avoids having to evaluate the data covariance matrix as an intermediate step.
- The combination of a probabilistic model and EM allows us to deal with missing values in the data set.
- Probabilistic PCA forms the basis for a Bayesian treatment of PCA in which the dimensionality of the principal subspace can be found automatically from the data.
- The existence of a likelihood function allows direct comparison with other probabilistic density models. By contrast, conventional PCA will assign a low reconstruction cost to data points that are close to the principal subspace even if they lie arbitrarily far from the training data.
- Probabilistic PCA can be used to model class-conditional densities and hence be applied to classification problems.
- The probabilistic PCA model can be run generatively to provide samples from the distribution.

Wariacyjny autoenkoder

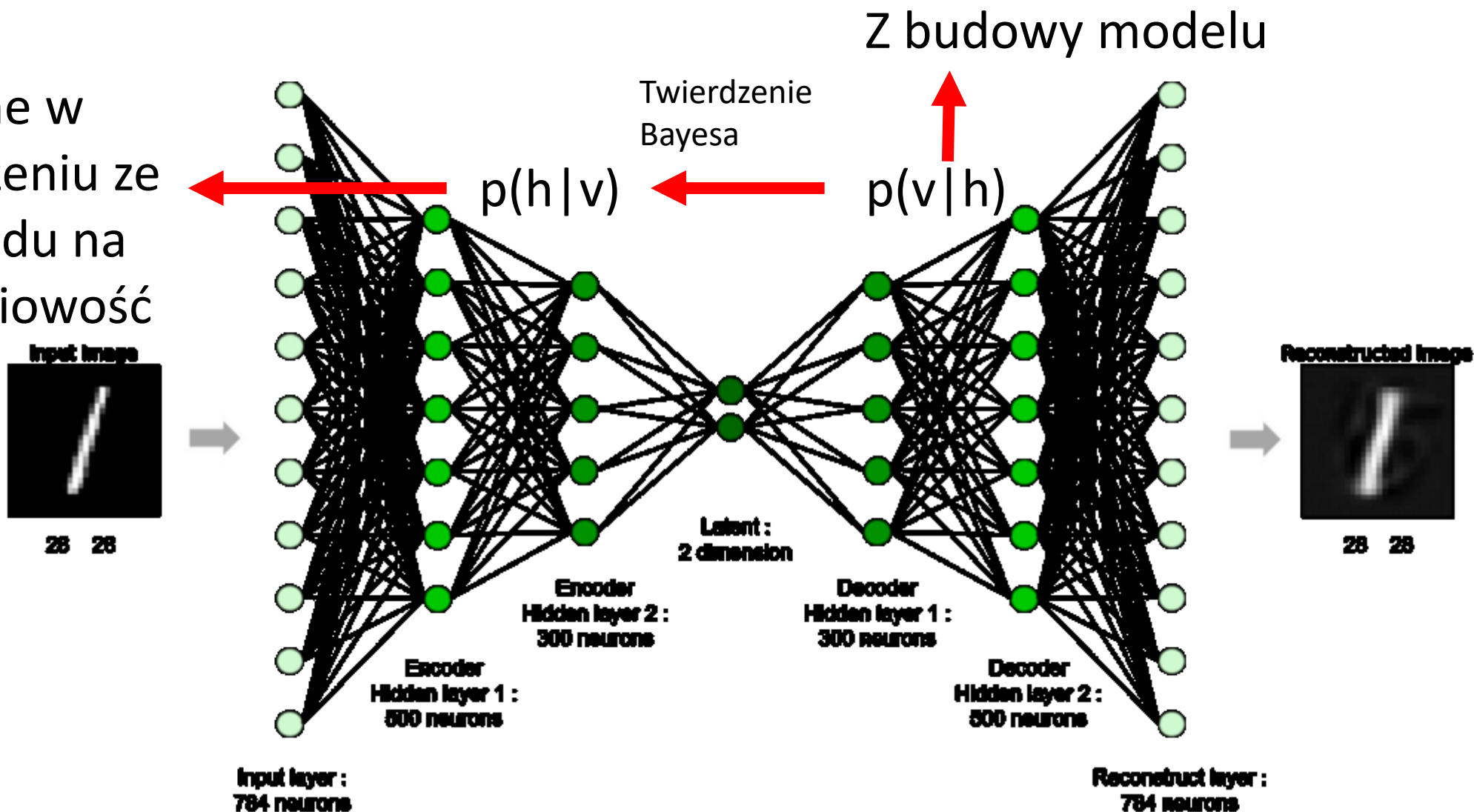
$$p(v|h) = d(\theta, h)$$



Wariacyjny autoenkoder

$$p(v|h) = d(\theta, h)$$

Trudne w
obliczeniu ze
względu na
nieliniowość



Uczenie autoenkodera wariacyjnego

Chcąc uczyć wariacyjny autoenkoder musimy przygotować funkcję wiarygodności:

$\log p(\mathbf{v})$

\mathbf{v} – visible

\mathbf{h} - hidden

Niestety, rozkład **$p(\mathbf{v})$** i rozkład **$p(\mathbf{h} | \mathbf{v})$** są ,intractable' i mogą zostać obliczone tylko numerycznie, co jest bardzo kosztowne

Approximate Inference

Negative Free Energy

Evidence Lower Bound (ELBO)

Takimi problemami w uczeniu maszynowym zajmuje się dziedzina
„Approximate Inference”

Zamiast funkcji wiarygodności

$\log p(\mathbf{v})$ maksymalizujemy funkcję :

$$\mathcal{L}(\mathbf{v}, \boldsymbol{\theta}, q) = \log p(\mathbf{v}; \boldsymbol{\theta}) - D_{\text{KL}}(q(\mathbf{h} | \mathbf{v}) || p(\mathbf{h} | \mathbf{v}; \boldsymbol{\theta}))$$



Trudne w
obliczeniu



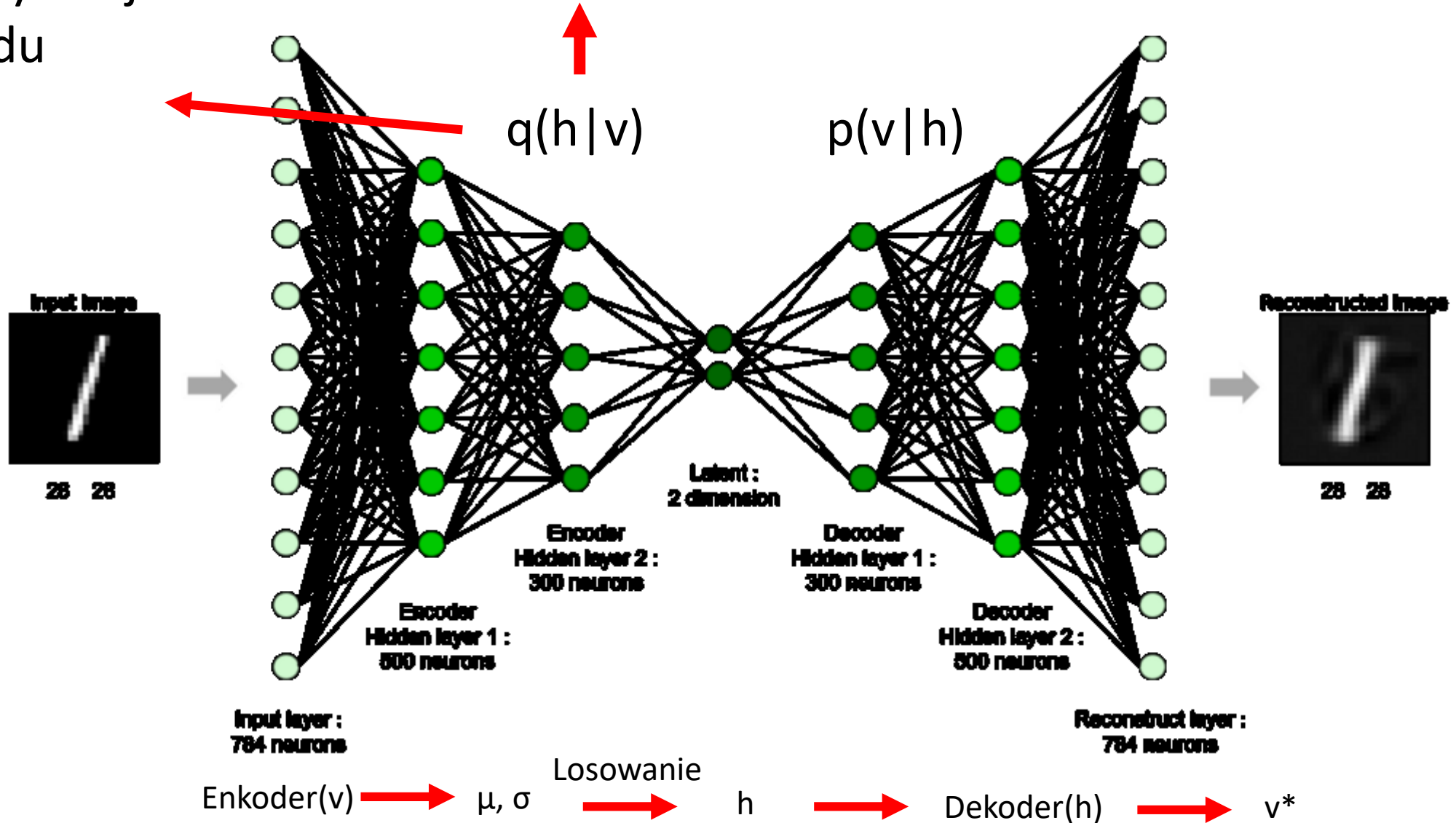
Parametryzowany
rozkład Gaussa



Trudne w
obliczeniu

Aproksymacja
rozkładu
 $p(h|v)$

$$q(h|v) = \mathcal{N}(z; \mu, \sigma^2 \mathbf{I})$$




Po kilku obliczeniach . . .


$$\begin{aligned}\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) &= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - D_{\text{KL}}(q(\boldsymbol{h} \mid \boldsymbol{v}) \parallel p(\boldsymbol{h} \mid \boldsymbol{v}; \boldsymbol{\theta})) \\ &= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{h} \sim q} \log \frac{q(\boldsymbol{h} \mid \boldsymbol{v})}{p(\boldsymbol{h} \mid \boldsymbol{v})} \\ &= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{h} \sim q} \log \frac{q(\boldsymbol{h} \mid \boldsymbol{v})}{\frac{p(\boldsymbol{h}, \boldsymbol{v}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})}} \\ &= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{h} \sim q} [\log q(\boldsymbol{h} \mid \boldsymbol{v}) - \log p(\boldsymbol{h}, \boldsymbol{v}; \boldsymbol{\theta}) + \log p(\boldsymbol{v}; \boldsymbol{\theta})] \\ &= - \mathbb{E}_{\boldsymbol{h} \sim q} [\log q(\boldsymbol{h} \mid \boldsymbol{v}) - \log p(\boldsymbol{h}, \boldsymbol{v}; \boldsymbol{\theta})] .\end{aligned}$$

Po kilku obliczeniach . . .


$$\mathcal{L}(\theta; v^{(i)}) = -D_{KL}(q_{\theta}(h|v^{(i)})||p_{\theta}(h)) + \mathbb{E}(\log p_{\theta}(v^{(i)}|h))$$



Rozkład Gaussa,
parametryzowany
przez enkoder



Przyjęty rozkład
Gaussa



Różnica pomiędzy przykładem v , a
jego odtworzeniem
W zależności od zadania:
- MSE
- binary cross entropy


$$p_{\theta}(h) = \mathcal{N}(h; 0, I)$$

Po kilku obliczeniach . . .

$$\mathcal{L}(\theta; v^{(i)}) = -D_{KL}(q_{\theta}(h|v^{(i)})||p_{\theta}(h)) + \mathbb{E}(\log p_{\theta}(v^{(i)} | h))$$

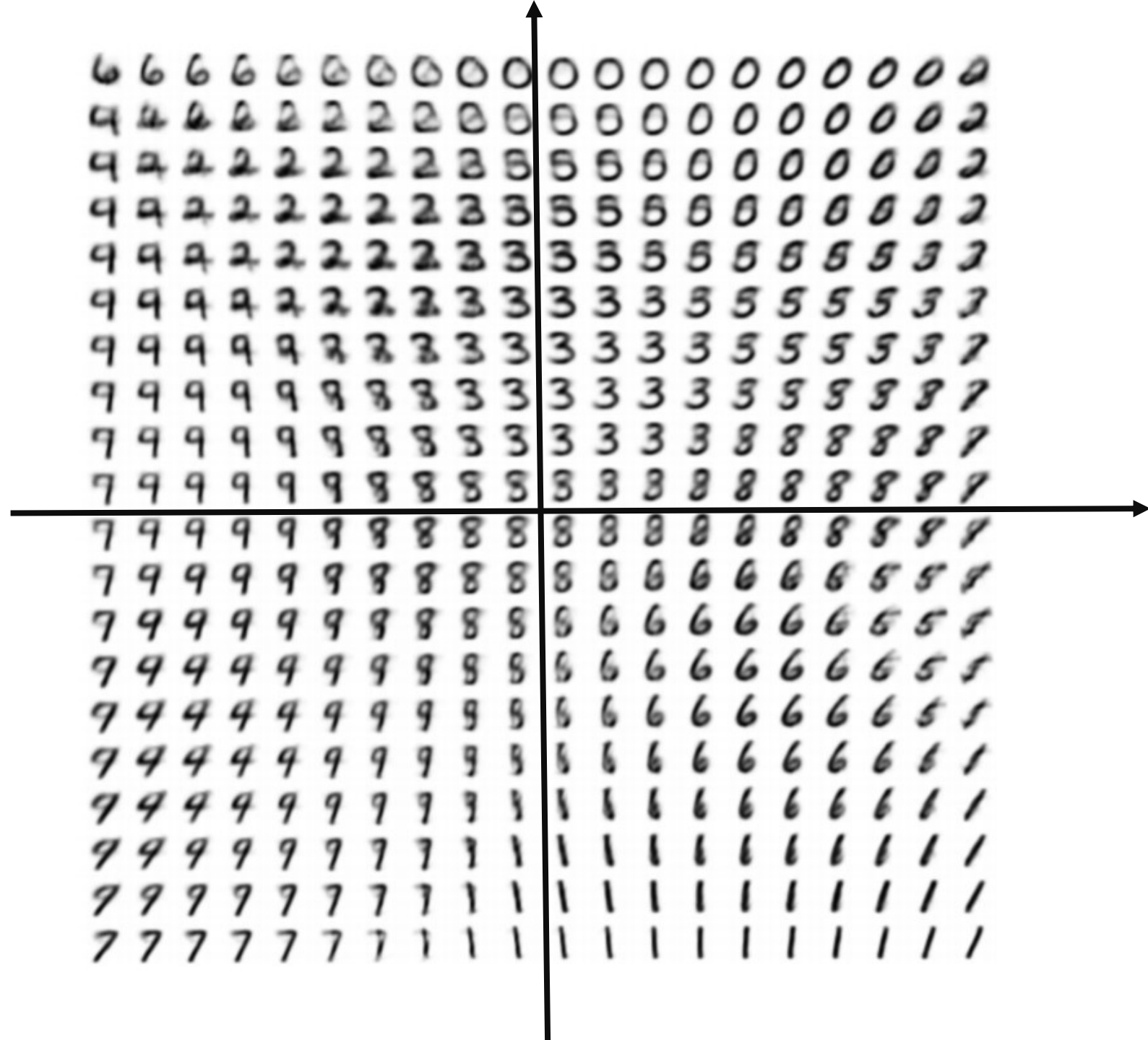
$$\begin{aligned} -D_{KL}((q_{\phi}(\mathbf{h})||p_{\theta}(\mathbf{h})) &= \int q_{\theta}(\mathbf{h}) (\log p_{\theta}(\mathbf{h}) - \log q_{\theta}(\mathbf{h})) d\mathbf{h} \\ &= \frac{1}{2} \sum_{j=1}^J (1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2) \end{aligned}$$

Po kilku obliczeniach . . .

$$\mathcal{L}(\boldsymbol{\theta} ; \mathbf{v}^{(i)}) \simeq \frac{1}{2} \sum_{j=1}^J \left(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2 \right) + \frac{1}{L} \sum_{l=1}^L \log p_{\boldsymbol{\theta}}(\mathbf{v}^{(i)} | \mathbf{h}^{(i,l)})$$



MSE lub binary cross-entropy



Zalety VAE względem AE

- Wypełnienie przestrzeni
- Disentangle of factors (rozwikłanie czynników)

Materialy

- Christopher Bishop, „**Pattern Recognition and Machine Learning**”
- Bengio Yoshua, Courville Aaron, Goodfellow Ian, „**Deep Learning**”
- Diederik P Kingma, Max Welling, „**Auto-Encoding Variational Bayes**”