UNIVERSITÄT
LEIPZIG

Dimensionsreduktion

INTERACTIVE VISUAL DATA MINING

ÜBERSICHT

Dimensionsreduktion

- Einleitung
- Factor Analysis
- Principal Component Analysis (PCA)
- Multi-Dimensional Scaling (MDS)
- T-distributed stochastic neighborhood embedding (t-SNE)
- Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP)

EINLEITUNG

- Ausgangssituation
  - Tabelle mit n Datenpunkten und m Attributen

ID	Attribut 1	Attribut 2	Attribut 3	…	Attribut m
id_1	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$		$a_{1,m}$
id_2	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$		$a_{2,m}$
…					
id_n	$a_{n,1}$	$a_{n,2}$	$a_{n,3}$		$a_{n,m}$

- Problem
  - Mehr Daten als Darstellungsfläche
- Lösung
  - Datenauswahl
    - Einschränkung der Attribute
    - Einschränkung der Attributwerte
    - Projektion
      Datenreduktion
    - Dimensionsreduktion („Dimensionality Reduction“)
  - Clustering

EINLEITUNG

- Projektion
  - Wähle $d \in \{2, 3\}$ Attribute
  - Führt in der Regel zu vielen „gleichen“ Datenpunkten → Overplotting
  - Probleme
    - Welche Attribute werden betrachtet?
    - Berücksichtigt die Werte und die Verteilung der Werte der Datenpunkte nicht

- Dimensionsreduktion
  - Ersetze gegebene Attribute
    $$A = \{A_1, \dots, A_m\}$$
    durch eine Menge von weniger Attributen
    $$D = \{D_1, \dots, D_d\}$$
    $$d \ll m$$
    - Meist: $d \in \{2, 3\}$
  - Alternativen:
    - $D \subseteq A$: entspricht Projektion
    - $D \cap A = \emptyset$
      - Die neuen Attribute werden aus den alten berechnet
    - Kombination aus beidem

EINLEITUNG

- Verfahren:
  - Factor Analysis
  - Principal Component Analysis (PCA)
  - Multi-Dimensional Scaling (MDS)

- Andere Möglichkeiten
  - Self-Organizing Maps
  - Visuelle Metaphern, um die Dimensionen anzuordnen

- **Achtung!**
- Für alle Verfahren gilt:
  - Information, die nicht erhoben wurde, wird durch eine Analyse nicht hinzugefügt
  - Beware of „garbage in, garbage out“ Vorsicht vor „Müll hinein, Müll heraus“:
    - Das Verfahren findet immer Faktoren
    - Hintergrund-/Domänen-Wissen ist notwendig, um zu entscheiden, ob die Faktoren einen Wert haben

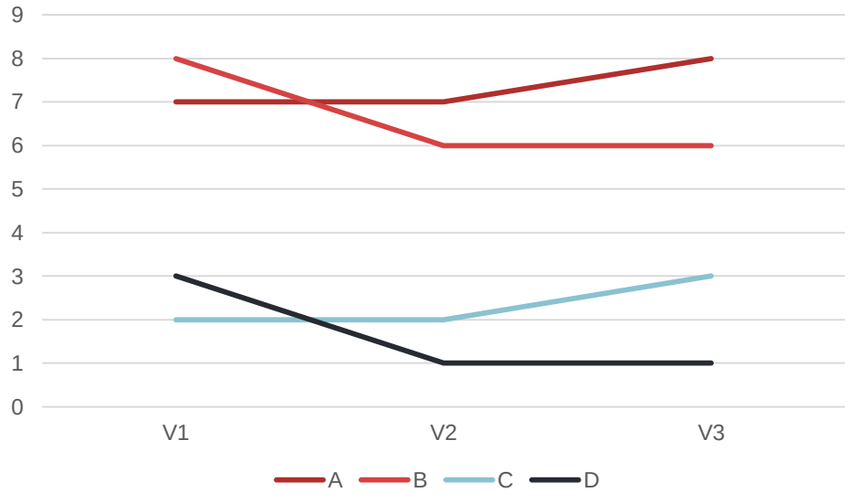
FACTOR ANALYSIS

- Arten
  - Typ Q
    - Zusammenfassung der Datenpunkte
    - Aufwändige Berechnung
    - Meist wird stattdessen Clustering verwendet
  - Typ R
    - Finde „verborgene Dimensionen“ (Gruppen von korrelierten Variablen)
- Ziele:
  - Identifikation von Strukturen
  - Zusammenfassung der Daten
  - Reduktion der Daten
- Eigenschaften:
  - Gruppen sind disjunkt
  - Gruppen können von ihren Mitgliedern repräsentiert werden

FACTOR ANALYSIS

- Typ Q Faktor-Analyse – Clustering

	V_1	V_2	V_3
A	7	7	8
B	8	6	6
C	2	2	3
D	3	1	1



- Faktor-Analyse
  - Gruppe 1: A, C
  - Gruppe 2: B, D
- Clustering
  - Gruppe 1: A, B
  - Gruppe 2: C, D

FACTOR ANALYSIS

- Anzahl der Datenpunkte ≫ Anzahl der Variablen
  - Minimum 50 – 100 Datenpunkte
  - #Datenpunkte = 5-10 · #Variablen
  - #Datenpunkte = 20 · #Variablen
- Verwendung von möglichst wenig Variablen
- Verwendung von möglichst vielen Datenpunkten

- Anwendbarkeit
  - Bartlett‘s Test auf Sphärizität
    - Nullhypothese: die Korrelationsmatrix ist gleich der Einheitsmatrix
    - Signifikanz: $p < 0{,}05$
    - Nullhypothese wird abgelehnt → Faktoranalyse möglich
    - Voraussetzung: multivariate Normalverteilung
    - $X^2 = -\left(n - 1 - \frac{2 \cdot m + 5}{6}\right) \log(\det(R))$
    - R: Korrelationsmatrix

FACTOR ANALYSIS

Anwendbarkeit: Measure of Sampling Adequacy (MSA)

- Gesamt und für jede der m Variablen
- Variablen mit kleineren Werten werden von der Faktoranalyse ausgenommen
- $\forall 1 \leq j \leq m$:

$$MSA_j := \frac{\sum_{k \neq j} r_{jk}^2}{\sum_{k \neq j} r_{jk}^2 + \sum_{k \neq j} p_{jk}^2}$$

- r_{jk}: Korrelation zwischen j und k
- p_{jk}: partielle Korrelation zwischen j und k

- Auswertung
  - $< 0{,}5$: Variable ungeeignet
  - $0{,}6 < MSA_j \leq 0{,}8$: Variable brauchbar
  - $> 0{,}8$: Variable gut geeignet

- Partielle Korrelation:
  - $p_{jk,l} = \frac{r_{jk} - r_{jl} \cdot r_{kl}}{\sqrt{\left(1 - r_{jl}^2\right) \cdot \left(1 - r_{kl}^2\right)}}$

FACTOR ANALYSIS

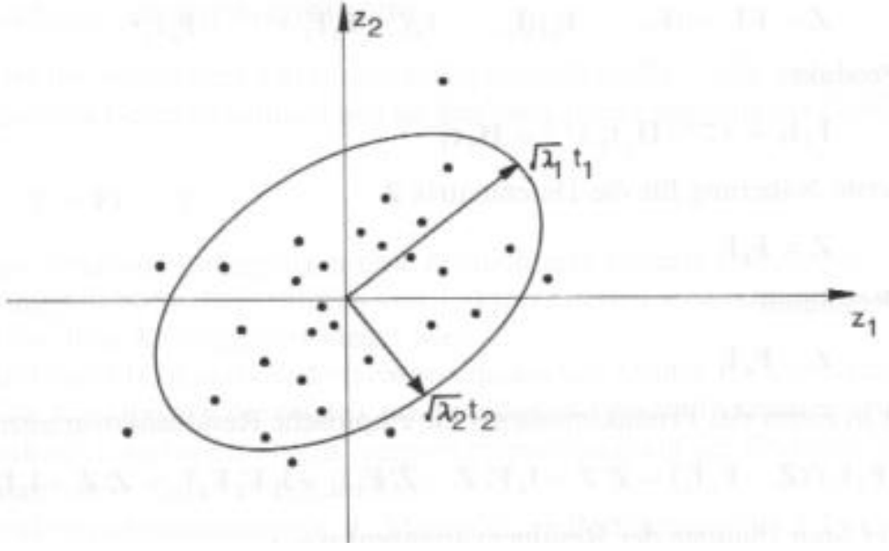
- Varianz
  - Gemeinsame Varianz mehrerer Variablen
  - Spezifische Varianz einer Variablen
  - Fehler-Varianz
- Component Analysis / Principal Component Analysis
  - Betrachtet gesamte Varianz der Variablen
- Common Factor Analysis
  - Betrachtet nur die gemeinsame Varianz der Variablen
  - Spezifische und Fehler-Varianz werden nicht betrachtet

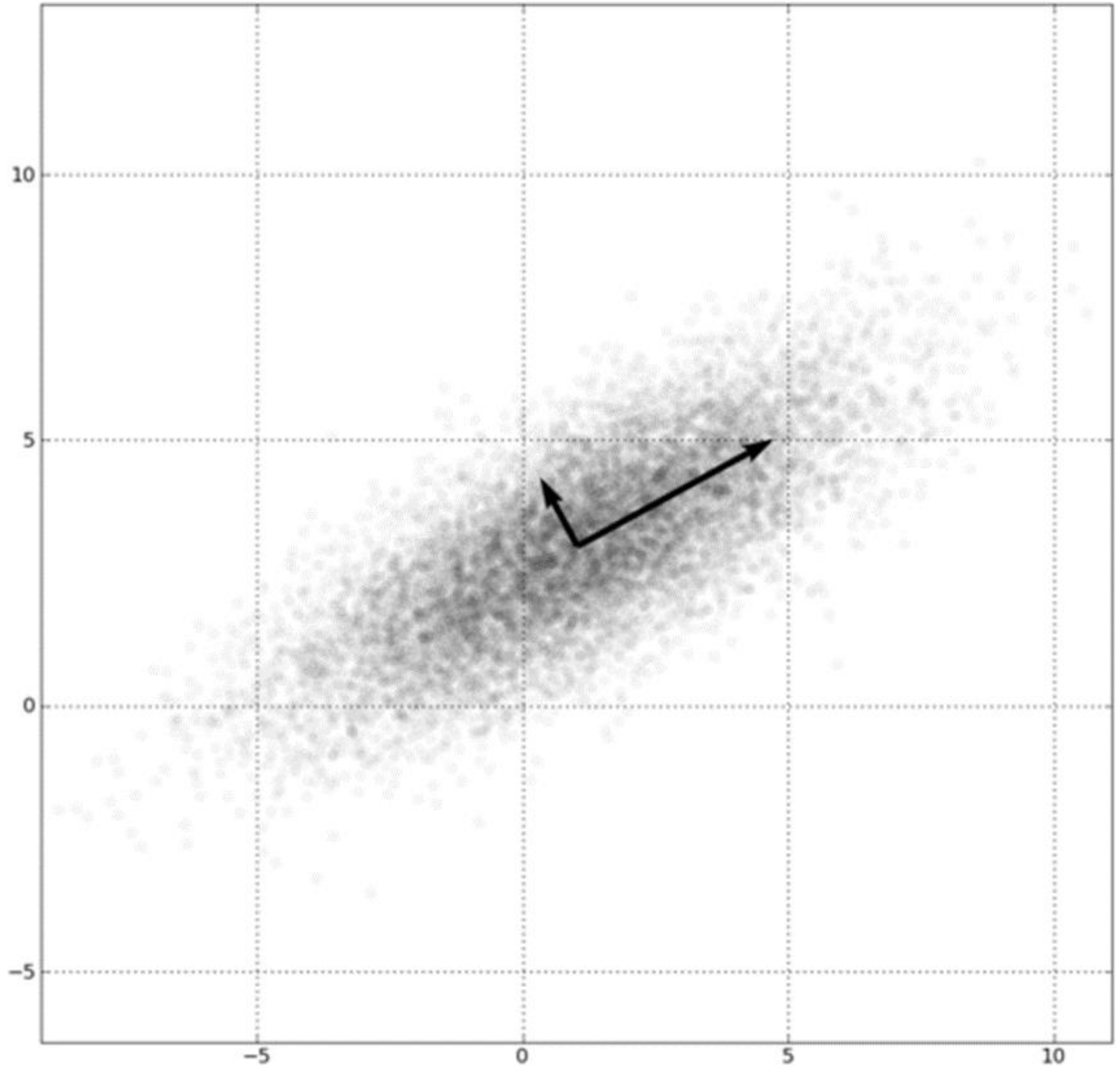
FACTOR ANALYSIS

- Anzahl der Faktoren
  - Fest: zum Beispiel in der Visualisierung
  - Betrachtete Varianz
    - Summiere die von den Faktoren berücksichtigte Varianz bis ein Schwellwert überschritten wird
      - Üblich: 95%
  - Betrachte nur Faktoren mit Eigenwerten ≥ 1
    - Gut geeignet für 20-50 Variablen
    - $m < 20$: tendenziell zu wenige Faktoren
    - $m > 50$: tendenziell zu viele Faktoren
- Nachbearbeitung Rotation
  - Zeilen: Variablen
  - Spalten: Faktoren
  - Einträge: Beitrag der Variablen zu einem Faktor
  - VARIMAX
    - Spalten-orientiert
    - Optimiere Beiträge der Variablen zu einem Faktor, so dass sie nahe 1 oder nahe 0 sind

PRINCIPAL COMPONENT ANALYSIS (PCA)

- 30 Beobachtungen
- t_1: Richtung der größten Variation (Diffusion, Abweichung)
- t_2: Richtung der zweitgrößte Variation (Diffusion, Abweichung)
- …





PRINCIPAL COMPONENT ANALYSIS (PCA)

- Gegeben: m möglicherweise korrelierte beobachtete Variablen $x_1, \ldots, x_m$
- Ergebnis: $p < m$ unkorrelierte Variablen (principal components)
- Anforderung: p soll so klein wie möglich sein
- Visualisierung: $p \in \{2, 3\}$
  Alternativ: Wähle p, so dass eine möglichst große Varianz abgedeckt wird und verwende Scatterplot-Matrizen

6.3 PRINCIPAL COMPONENT ANALYSIS (PCA)

- Andere Namen
  - Karhunen-Loève Transformation (KLT, Bildverarbeitung)
  - Empirical orthogonal functions (Meteorologie, Geophysik)
  - Hotelling Transformation
  - Proper orthogonal decomposition (POD)
  - Deutsch: Hauptkomponentenanalyse
- Ziele
  - Berechne eine reduzierte Menge von Dimensionen
    - Orthogonal
    - Linear
  - Ordne die Dimensionen absteigend bezüglich der Varianz
- Varianz
  - Maß für den Informationsgehalt einer Variable
- Mathematisch
  - Suche eine neue Basis des Vektorraumes

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Konstruktion
  - Gegeben: $X = (X_1, \dots, X_m),\ X_j = \left(x_{1,j}, \dots, x_{n,j}\right)^T$
  - Mittelwert: $\mu = (\mu_1, \dots, \mu_m),\ \mu_j = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{i,j}$
  - Zentrierte Werte: $Y = \left(y_{ij}\right),\ y_{ij} = x_{ij} - \mu_j$
  - Kovarianz-Matrix: $C = \frac{1}{n-1} \cdot Y^T Y$
    - Symmetrisch → spektrale Zerlegung möglich
  - Spektrale Zerlegung: $C = U \Lambda U^T,\ U^T U = I_m$
  - Eigenwert Matrix: $\Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_m \end{pmatrix}$
  - Eigenvektor Matrix: $U = (u_1, \dots, u_m)$
  - Ordne die Eigenwerte absteigend: $i > j \rightarrow \lambda_i > \lambda_j$

PRINCIPAL COMPONENT ANALYSIS (PCA)

- Eigenschaften:
  - λ_1, u_1: größte Varianz
  - λ_2, u_2: zweitgrößte Varianz
  - …
  - Die ersten p Eigenwerte beschreiben einen Großteil der Varianz
  - Principal components mit einer Varianz nahe 0 können verwendet werden, um Ausreißer zu identifizieren
- Wird auch häufig in der Scientific Visualization verwendet

- Einschränkungen:
  - Das Ergebnis ist abhängig von der Skalierung der einzelnen Variablen (Attribute)
  - Ausreißer haben einen großen Einfluss auf das Ergebnis
  - Lineare Methode
  - Interpretation der Basis (Ergebnisse)

MULTI-DIMENSIONAL SCALING (MDS)

- Gegeben: eine Tabelle mit Ähnlichkeiten
- Gesucht: Abbildung der Datenpunkte von einem Koordinatensystem in ein anderes
- Bedingungen
  - Monotonie: die Distanz zwischen zwei Datenpunkten im Zielsystem ist kleiner, wenn die Distanz der Datenpunkte im Originalsystem kleiner ist
  - Minimum: die Dimension des Zielsystems soll so klein wie möglich sein
- MDS: Menge von Systemen

MULTI-DIMENSIONAL SCALING (MDS)

- Gegeben: $X = (x_1, \dots, x_r),\ x_i \in \mathbb{R}^m$
- Gesucht: $Y = (y_1, \dots, y_n),\ y_i \in \mathbb{R}^n,\ n \ll m$
- Bedingung: $\|y_i - y_j\| \approx t_{ij}, \forall\ i, j$
- In der Visualisierung: $n = 2$ oder $n = 3$
- Minimierung der Kostenfunktion: $\min_{x_1, \dots x_r} \sum_{i<j} \left(\|y_i - y_j\| - t_{ij}\right)^2$

MULTI-DIMENSIONAL SCALING (MDS)

- Metric Multi-Dimensional Scaling
  - Gegeben: Distanzmatrix (t_{ij})
  - $A = (a_{ik}),\ a_{ik} = -\frac{1}{2} {t_{ik}}^2$
  - $B = (b_{ij}),\ b_{ij} = a_{ij} - \frac{1}{m}\sum_{k=1}^{m} a_{ik} - \frac{1}{m}\sum_{l=1}^{m} a_{lj} - \frac{1}{m^2}\sum_{l=1}^{m}\sum_{k=1}^{m} a_{lk}$
  - Bestimme die Eigenwerte λ_i und die zugehörigen Eigenvektoren e_i von B mit $\sum_{j=1}^{m} {\gamma_{ij}}^2 = \lambda_i$
  - Wähle die m Eigenvektoren mit den größten Eigenwerten
  - Vergleichbar mit der PCA

T-DISTRIBUTED STOCHASTIC NEIGHBORHOOD EMBEDDING (T-SNE)

- t-SNE ist eine weitere Methode zur Dimensionsreduktion
- Daten werden in eine 2 dimensionale Ebene eingebettet
- T-SNE versucht die lokale Verteilung der Daten zu bewahren
- Häufig im Bereich des Maschinellen Lernens eingesetzt

METHODE

- Berechne für jede Distanz zwischen Punkt i und j eine abhängige Wahrscheinlichkeit, welche die Ähnlichkeit repräsentiert
- Platziere einen Gausschen Kernel über dem Punkt i und berechne die Wahrscheinlichkeiten aller Nachbarn.
- Wandle die abhängigen Wahrscheinlichkeiten in Wahrscheinlichkeiten um
- Für jeden Punkt i muss die Breite des Kernels angepasst werden

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$$

METHODE

- Platziere jeden Punkt in eine 2D Ebene
- Berechne für jeden Punkt eine Wahrscheinlichkeit basierend auf der Student t-Verteilung
  - Ein Freiheitsgrad
- Optimiere die Einbettung durch einen Gradient Walk, bis die Wahrscheinlichkeiten “passen”
- Kullback-Leibler Divergenz als Metrik für die Platzierung

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

METHODE

- t-SNE hat mehrere Parameter
  - Perplexity: Wie viele Datenpunkte sollen im Gauss Kernel liegen
  - Maximale Iterationen
  - theta: Parameter für die Barnes-Hut Optimierung
  - Anzahl der Dimensionen

METHODE

- Die Komplexität ist für Laufzeit und Speicher ist n^2
- Optimierung des 2. Schrittes mit einem Quadtree
- Nur für Datensätze mit weniger als 10.000 Datenpunkten
- Erzeugt keine deterministischen Ergebnisse

- Demo

- https://distill.pub/2016/misread-tsne/

UMAP

- UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
- Ähnlich zu t-SNE, hat aber entscheidende Unterschiede
  - Berechnet auch bedingte Wahrscheinlichkeiten, diese werden aber nicht normalisiert
  - Kann andere Distanzmetriken als Euklid nutzen
  - Normalisiert auch nicht während der Einbettung
  - Nutzt nearest neighbor statt perplexity

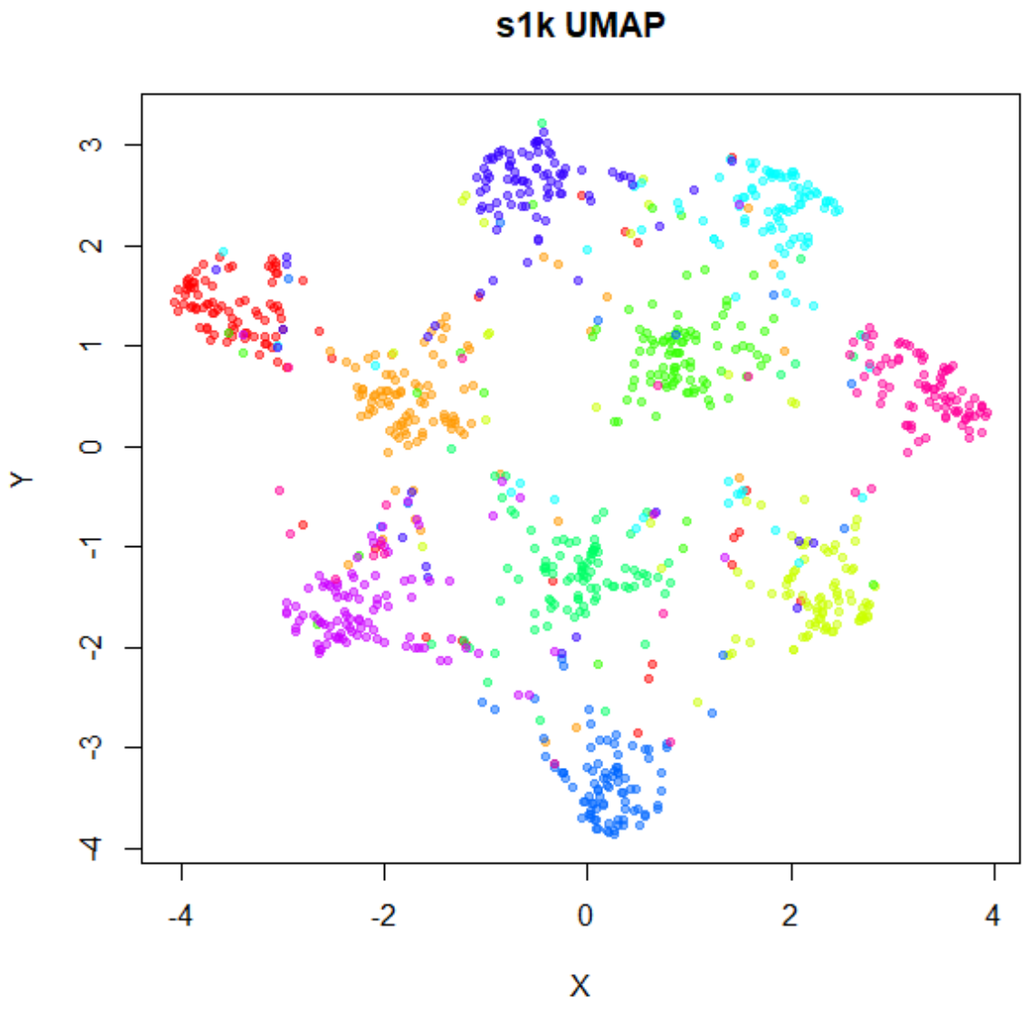
UMAP

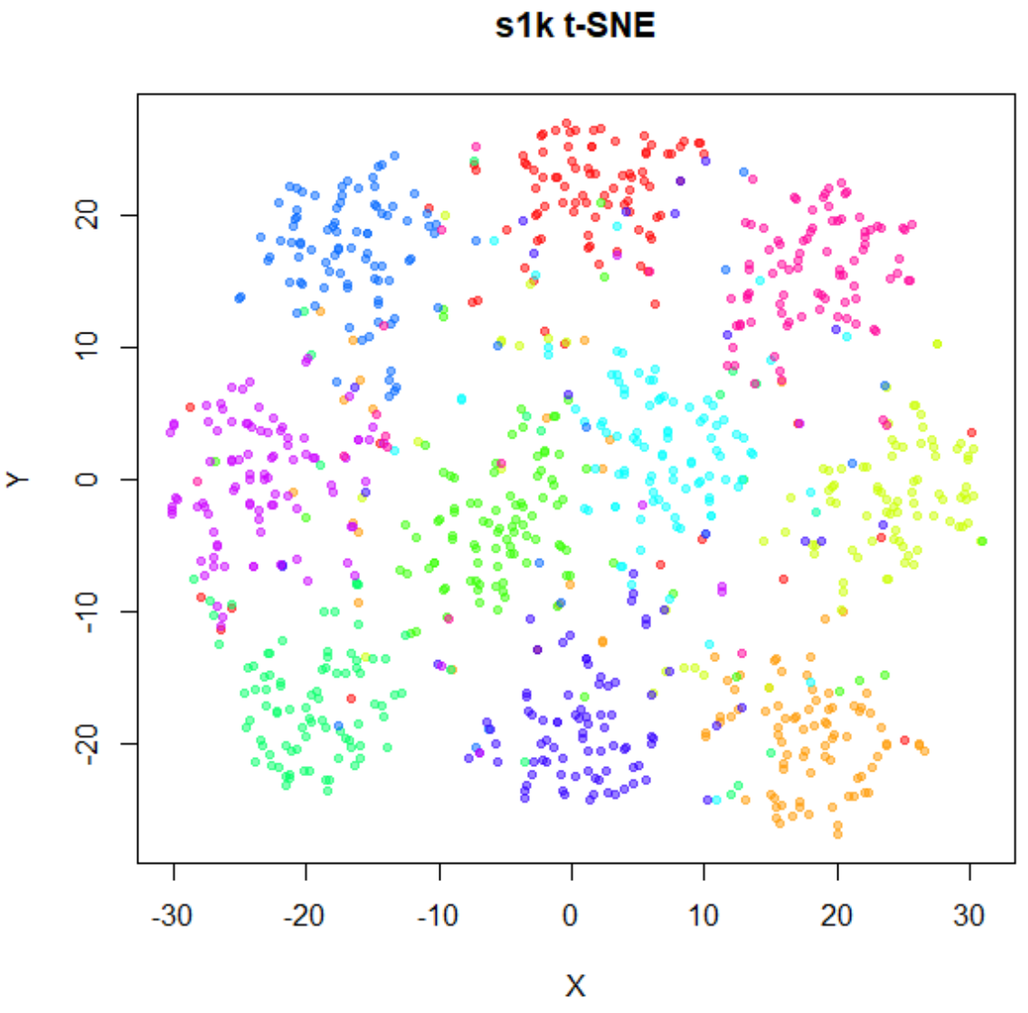
- Nutzt bei der Einbettung eine andere Initialisierung, Verteilungsfunktion sowie Vergleichsmaß
  - Graph Laplacian
  - Binary Cross Entropy
- Nutzt den Stochastic Gradient Descent statt dem regulären
  - Performanzvorteil

UMAP

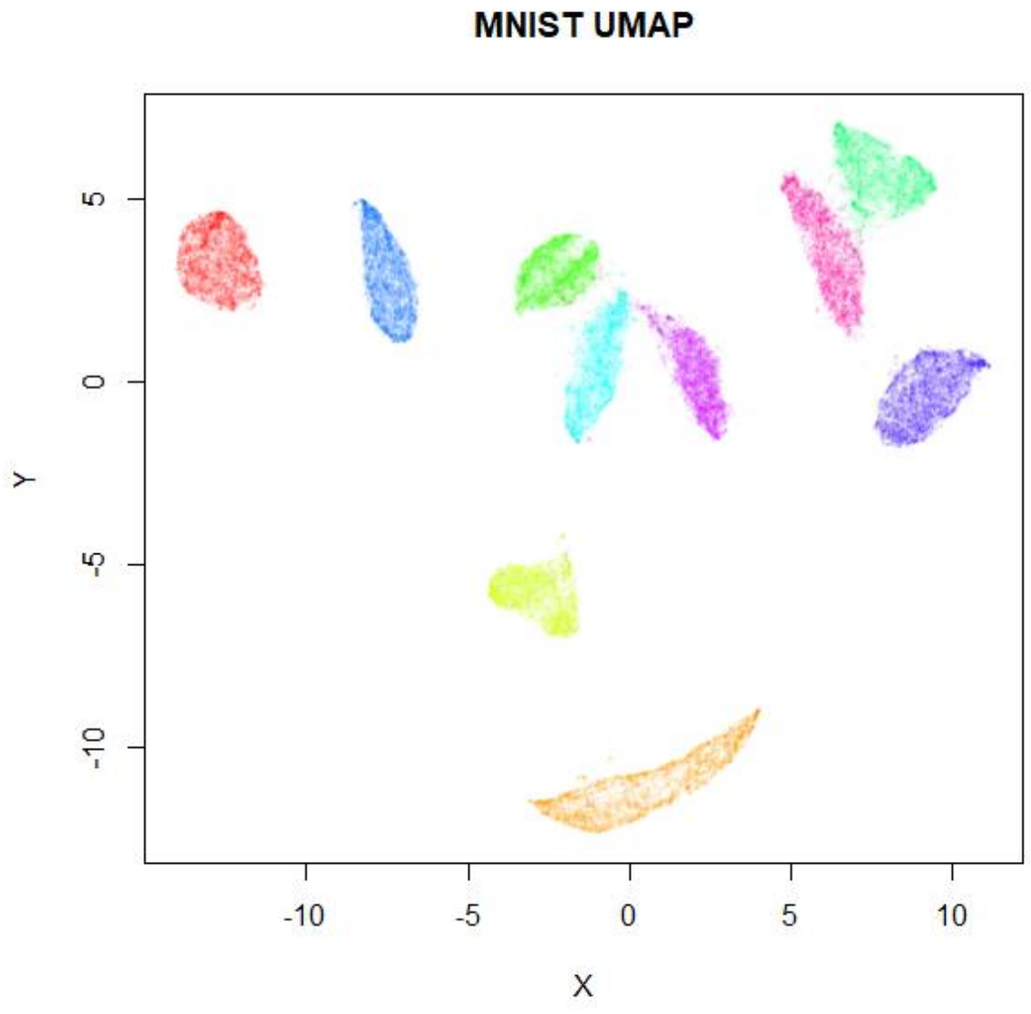
- Kann durch diese Veränderungen wesentlich mehr Datenpunkte verarbeiten
  - Keine Logarithmen
  - Keine Normalisierung
  - SGD bei der Einbettung
- Konserviert auch globale Strukturen in der Einbettung

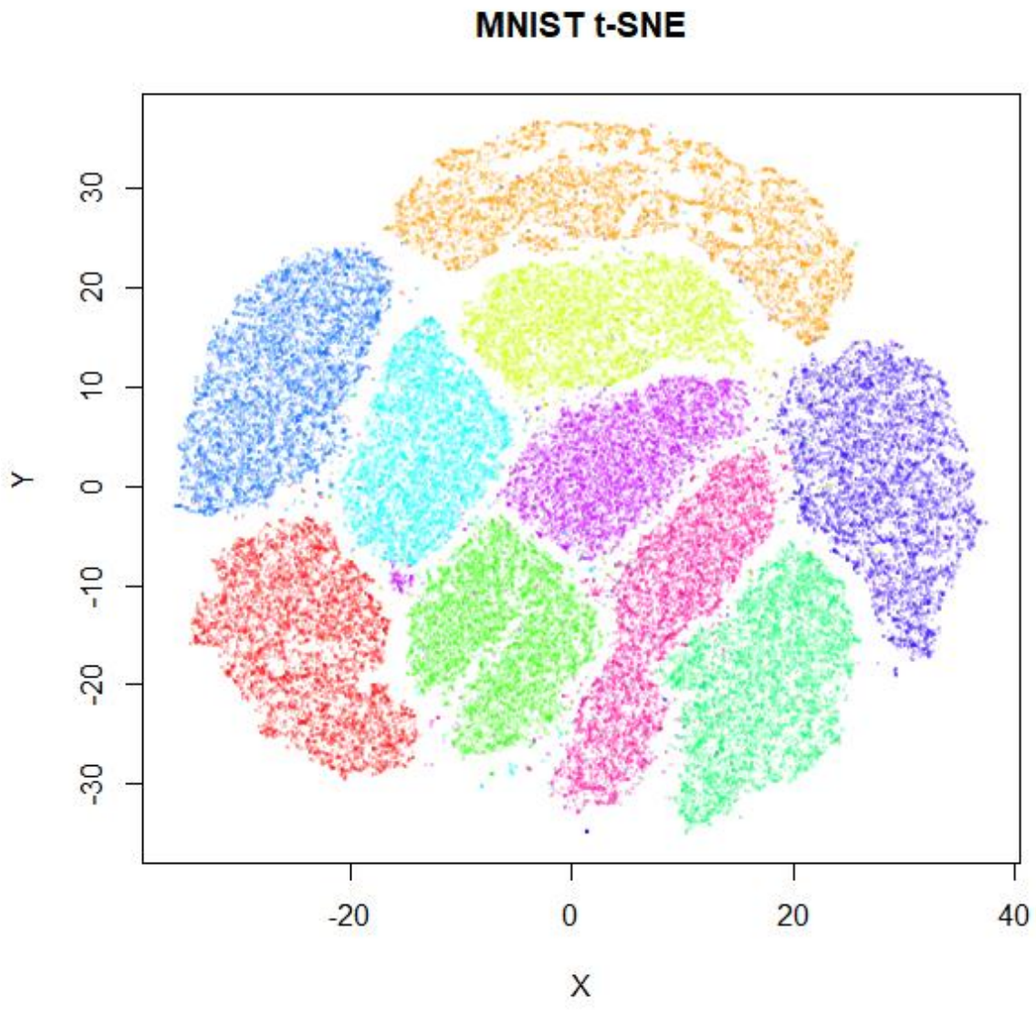
UMAP


s1k UMAP
Y
X
3
2
1
0
-1
-2
-3
-4
-4
-2
0
2
4


s1k t-SNE
Y
X
20
10
0
-10
-20
-30
-20
-10
0
10
20
30

UMAP


MNIST UMAP
Y
X
5
0
-5
-10
-10
-5
0
5
10


MNIST t-SNE
Y
X
30
20
10
0
-10
-20
-30
-20
0
20
40