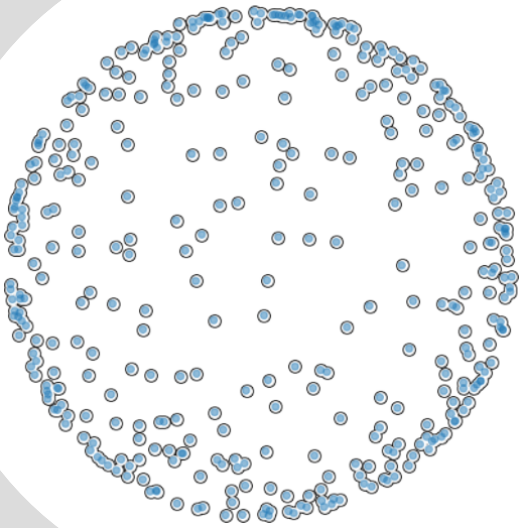
WiSe 2024/25

Machine Learning for Data Science

Lecture by G. Montavon

FREIE UNIVERSITÄT BERLIN



Lecture 8c

Sequence Models

Outline

Motivations

Sequence models

- Problem formulation
- Least squares regression
- Tikhonov regularization
- Total variation regularization

Further topics

- Multi-dimensional sequences
- Two-component models

Part 1

Motivations

Motivation: Time Series Analysis

- Particular types of sequences that are found in many applications.

Browse Through: **38** Data Sets

Table View List View

Default Task
Classification (27)
Regression (23)
Clustering (7)
Other (0)

Attribute Type
Categorical (0)
Numerical (37)
Mixed (1)

Data Type - Undo
Multivariate (137)
Univariate (5)
Sequential (21)
Time-Series (38)
Text (10)
Domain-Theory (2)
Other (3)

Area
Life Sciences (3)
Physical Sciences (5)
CS / Engineering (22)
Social Sciences (1)
Business (1)
Game (0)
Other (6)

 # Attributes - Undo
Less than 10 (27)
10 to 100 (38)
Greater than 100 (21)

 # Instances - Undo
Less than 100 (3)
100 to 1000 (8)
Greater than 1000 (38)

Format Type
Matrix (31)
Non-Matrix (7)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
AI4I 2020 Predictive Maintenance Dataset	Multivariate, Time-Series	Classification, Regression, Causal-Discovery	Real	10000	14	2020
Air Quality	Multivariate, Time-Series	Regression	Real	9358	15	2016
Air quality	Multivariate, Time-Series	Regression	Real	9358	15	2016
Appliances energy prediction	Multivariate, Time-Series	Regression	Real	19735	29	2017
Australian Sign Language signs	Multivariate, Time-Series	Classification	Categorical, Real	6650	15	1999
Australian Sign Language signs (High Quality)	Multivariate, Time-Series	Classification	Real	2565	22	2002
Beijing Multi-Site Air-Quality Data	Multivariate, Time-Series	Regression	Integer, Real	420768	18	2019
Beijing PM2.5 Data	Multivariate, Time-Series	Regression	Integer, Real	43824	13	2017
BitcoinHeistRansomwareAddressDataset	Multivariate, Time-Series	Classification, Clustering	Integer, Real	2916697	10	2020
BLE RSSI Dataset for Indoor localization and Navigation	Multivariate, Sequential, Time-Series	Classification, Clustering	Integer	6611	15	2018

Motivation: Analysis of Geology Data

- We want to better understand the structure of the earth surface.
- Organic deposits form complex three-dimensional volumes due to the complex earth dynamics (sedimentation, tectonics, alluvion, tides, etc.)
- These volumes exhibit some regularities, e.g. mostly continuous with few abrupt transition between facies.
- These volumes can only be partly observed (drilling is expensive, sensing from the surface only provides partial information).

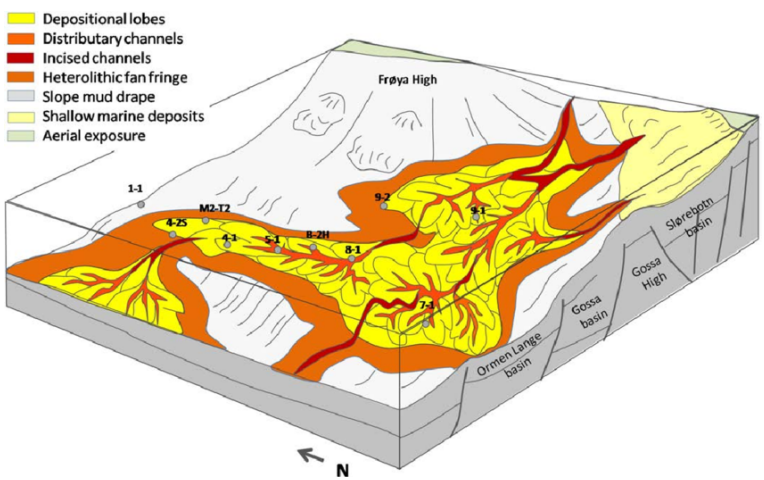


Image source: DOI:10.1144/SP403.7

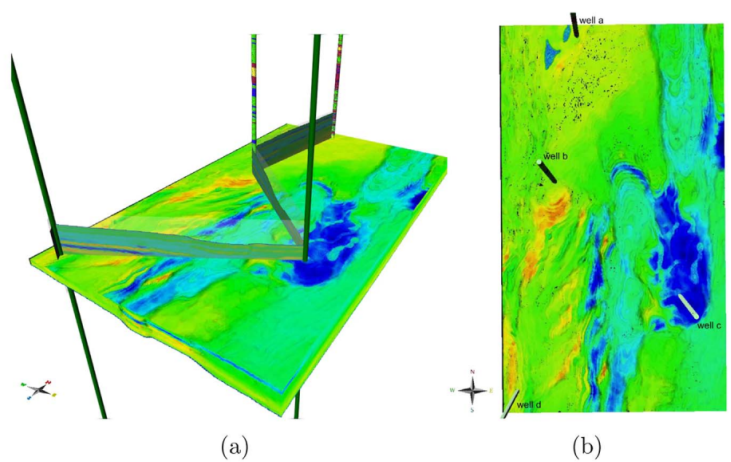


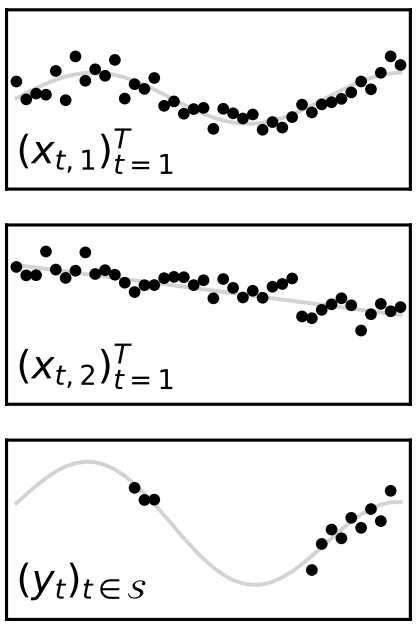
Image source: DOI:10.1016/j.cageo.2017.05.004

Part 2

Sequence Models

Sequence Data

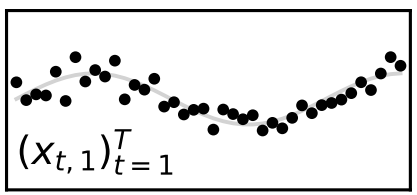
- Let
$$X = (\boldsymbol{x}_t)_{t=1}^T$$
with $\boldsymbol{x}_t \in \mathbb{R}^d$ be a multivariate input sequence.
- Let
$$\boldsymbol{y} = (y_t)_{t=1}^T$$
be a univariate sequence we would like to predict.
- We only have a few observations (i.e. a couple of indices $t \in \mathcal{S}$) of the latter sequence, and we denote these observations by $\boldsymbol{y}_\mathcal{S}$, i.e.
$$\boldsymbol{y}_\mathcal{S} = (y_t)_{t \in \mathcal{S}}$$

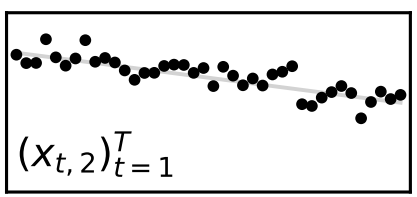


Sequence Data: Two Possible Views

View 1: Non-iid. dataset

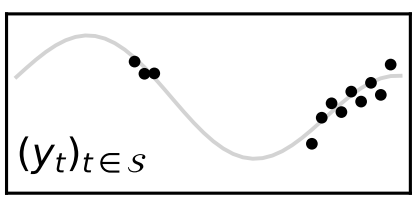
- Each index t corresponds to one data point (i.e. the dataset size is $N = T$).
- Some instances t are labeled, other are unlabeled $\Rightarrow$ it is a *semi-labeled dataset*.
- Dataset is non-iid. (nearby points in the sequence are correlated).
- Each data point comes with its index t as meta-data or additional feature.





View 2: Single-point dataset ($N = 1$)

- The sequence is a unique data point that is high-dimensional ($d \times T$ input dimensions).
- The model built on this unique data point should be highly constrained (e.g. share parameters across different time steps).



Sequence Models

Simple approach:

- Apply least square regression (Lecture 7a), i.e. assuming centered data, minimize:

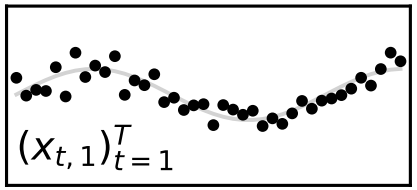
$$\mathcal{E}(\boldsymbol{w}) = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} (y_t - \boldsymbol{w}^\top \boldsymbol{x}_t)^2$$

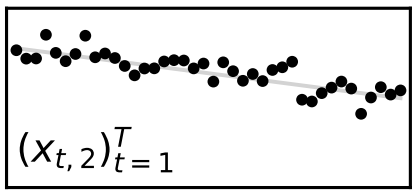
- This gives us the model:

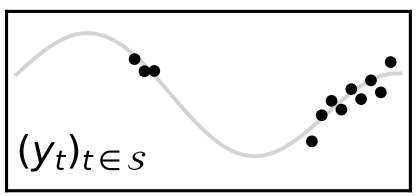
$$\boldsymbol{w} = C_{xx}^{-1} C_{xy}$$

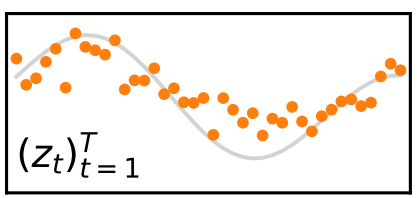
and the predictions:

$$\forall_{t=1}^{T} : \; z_t = \boldsymbol{x}_t^\top \boldsymbol{w}$$









Observations:

- This simple approach does not take advantage of the correlations between adjacent elements of the sequence.
- Although it is clear that the time series y_t is continuous, the prediction of that time series is noisy.

Tikhonov Regularization

Tikhonov Regularization [5]:

- Add to some least square objective of interest an additive term that penalizes too complex solutions:

$$\mathcal{E}(\boldsymbol{\theta}) = \|A\boldsymbol{\theta} - \boldsymbol{b}\|^2 + \|\Gamma\boldsymbol{\theta}\|^2$$

  The matrix Γ is provided by the user and implements the desired regularization/smoothness properties.

Andrey Nikolayevich Tikhonov (1906–1993)

Special case: Ridge regression

$$\begin{aligned}\mathcal{E}(\boldsymbol{w}) &= \frac{1}{N}\|X^\top \boldsymbol{w} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2 \\ &= \frac{1}{N}\|X^\top \boldsymbol{w} - \boldsymbol{y}\|^2 + \|\underbrace{\sqrt{\lambda}I}_{\Gamma}\boldsymbol{w}\|^2\end{aligned}$$

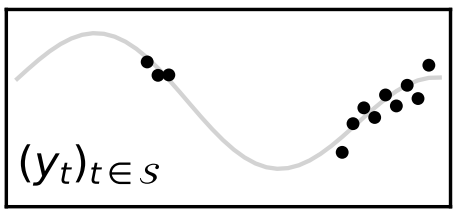
Note:

- Tikhonov regularization can apply to the parameters $\boldsymbol{w}$ of a linear model, but also to the predicted sequence itself.

Tikhonov Regularization

Idea:

- Build from $\boldsymbol{y}_{\mathcal{S}}$ a surrogate sequence $\boldsymbol{z}$ that matches the data where it is available and that penalizes local variations of that sequence:



$$\mathcal{E}(\boldsymbol{z}) = \sum_{t\in\mathcal{S}} (z_t - y_t)^2 + \lambda \sum_{t=1}^{T-1} (z_{t+1} - z_t)^2$$

- This can be seen as a Tikhonov regularization scheme

$$\mathcal{E}(\boldsymbol{z}) = \|M\boldsymbol{z} - \boldsymbol{y}_{\mathcal{S}}\|^2 + \|\Gamma\boldsymbol{z}\|^2$$

where M is a matrix that selects indices $t \in \mathcal{S}$ and where Γ is the Toeplitz matrix:

$$\Gamma = \sqrt{\lambda} \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}$$

Tikhonov Regularization

- Recall that the objective we would like to optimize is:

$$\mathcal{E}(\boldsymbol{z}) = \|M\boldsymbol{z} - \boldsymbol{y}_{\mathcal{S}}\|^2 + \|\Gamma\boldsymbol{z}\|^2$$

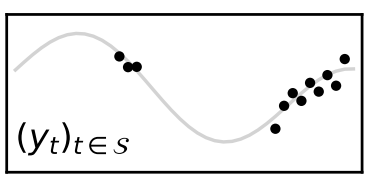
- Minimizing $\mathcal{E}(\boldsymbol{z})$ can achieved by first observing that the objective is convex with $\boldsymbol{z}$ and we find the minimum at $\partial\mathcal{E}/\partial\boldsymbol{z} = \mathbf{0}$. The derivative has the form:

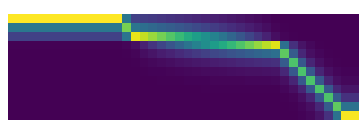
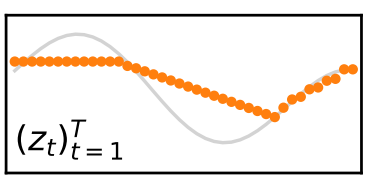
$$\frac{\partial\mathcal{E}}{\partial\boldsymbol{z}} = 2M^\top(M\boldsymbol{z} - \boldsymbol{y}_{\mathcal{S}}) + 2\Gamma^\top\Gamma\boldsymbol{z}$$

- Equating it to zero and isolating $\boldsymbol{z}$, we get the closed form:

$$\boldsymbol{z} = \underbrace{(M^\top M + \Gamma^\top\Gamma)^{-1}M^\top}_{W}\boldsymbol{y}_{\mathcal{S}}$$

Example:

- For 

we get 
$W^\top$
and 

Tikhonov Regularization: Choosing Γ

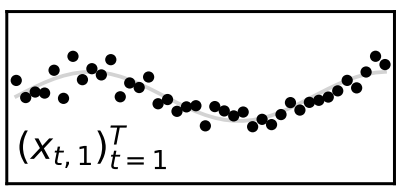
Linear regression:

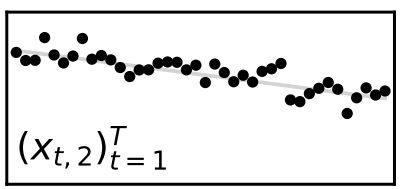
Minimizing slope:

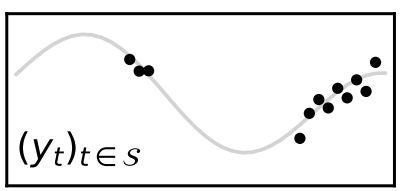
$$\begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix}$$

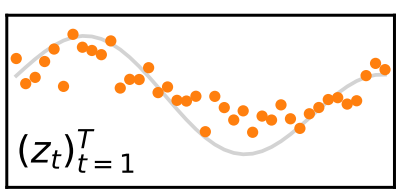
Minimizing curvature:

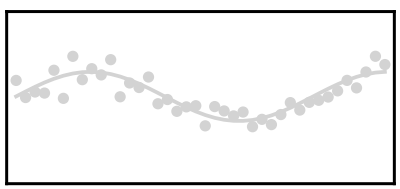
$$\begin{pmatrix} -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \end{pmatrix}$$

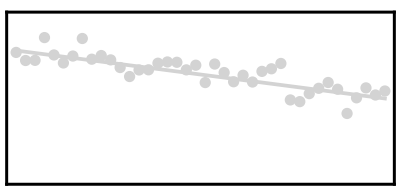


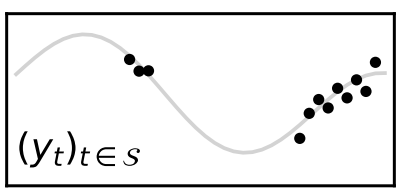


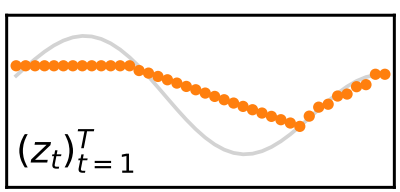


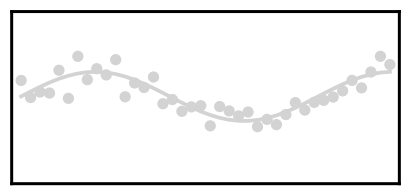


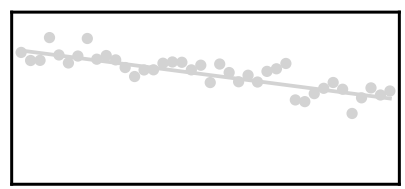


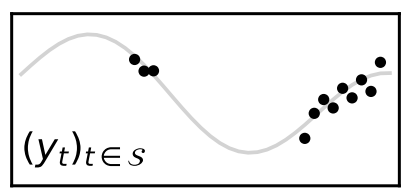


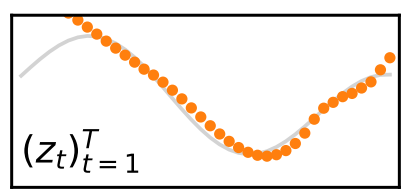












Combining Regularization and Regression

Idea:

- Combine Tikhonov regularization and the original regression model in a single objective.

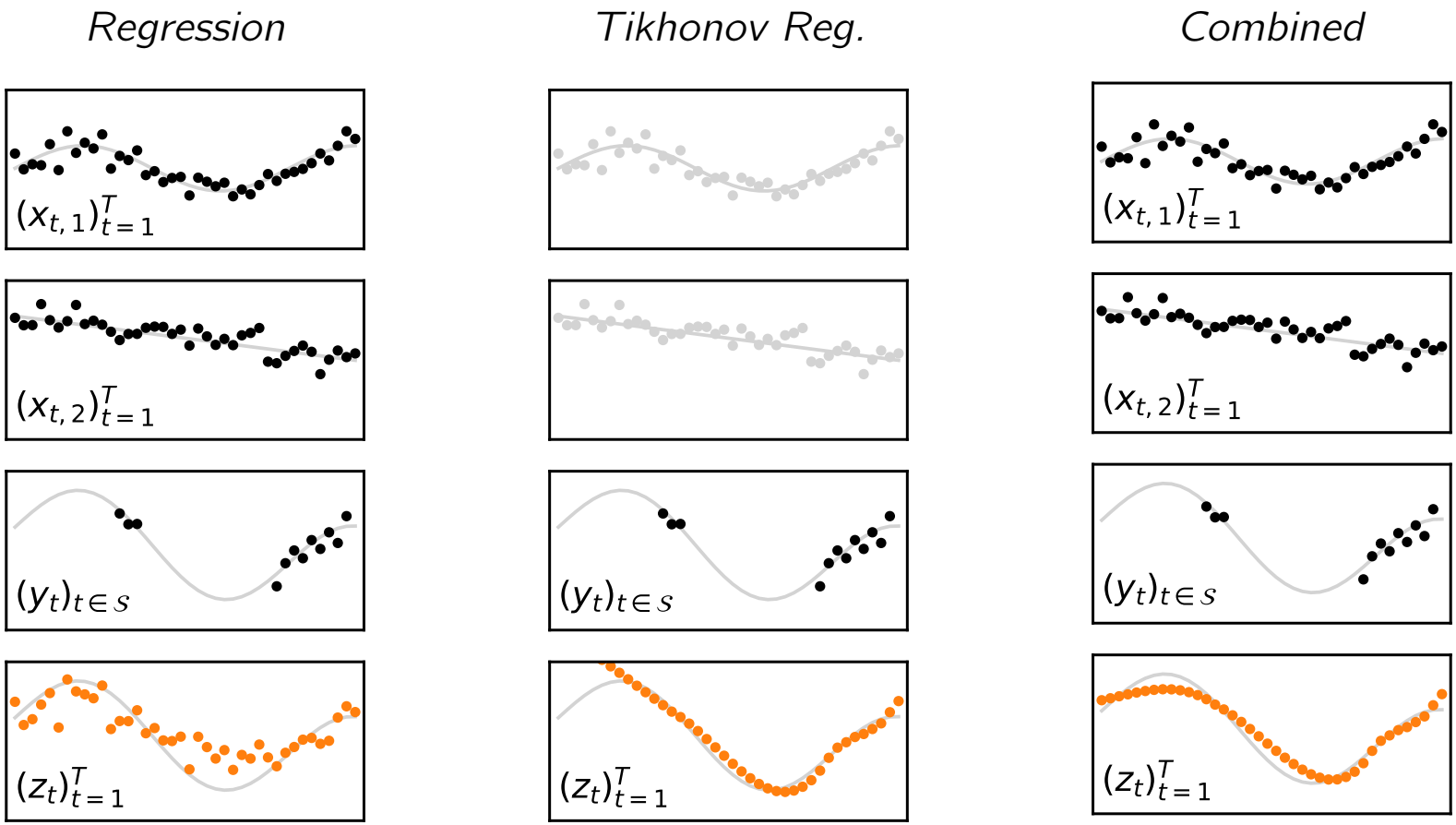
$$\mathcal{E}(\boldsymbol{z}, \boldsymbol{w}) = \|M\boldsymbol{z} - \boldsymbol{y}_{\mathcal{S}}\|^2 + \|\Gamma\boldsymbol{z}\|^2 + \eta \sum_{t=1}^{T} (z_t - \boldsymbol{w}^\top \boldsymbol{x}_t)^2$$

- The objective is non-convex, however, we can find a local optimum by iteratively solving the problem in closed form for $\boldsymbol{z}$ and $\boldsymbol{w}$, treating the other variable as constant. This gives us:

$$\begin{aligned} \boldsymbol{w} &\leftarrow C_{xx}^{-1} C_{xz} \\ \boldsymbol{z} &\leftarrow (M^\top M + \Gamma^\top \Gamma + \eta I)^{-1} (M^\top \boldsymbol{y}_{\mathcal{S}} + \eta X^\top \boldsymbol{w}) \end{aligned}$$

- Each step necessarily improves the objective. The model converges to some local optimum.

Combining Regularization and Regression



Observation:

- The combination of both approaches achieves the best results.

Limits of Tikhonov Regularization

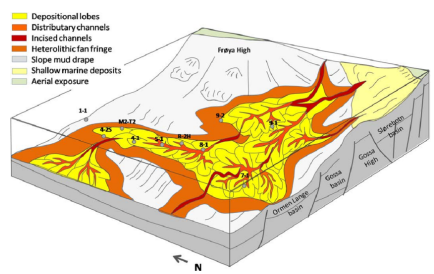
- Recall that the Tikhonov term is given by:

$$\|\Gamma \boldsymbol{z}\|^2 = \sum_j (\Gamma_j^\top \boldsymbol{z})^2$$

- We find that *any* violation of the smoothness property (e.g. $\Gamma_j^\top \boldsymbol{z} = z_{t+1} - z_t$ large) will incur a *very* large (quadratic) penalty in $\|\Gamma \boldsymbol{z}\|^2$.

Note:

- In practice, it is frequent that the sequence exhibits non-smooth behavior at few specific locations (e.g. transition between two geological facies, edge in an image, event in time).



Idea:

- Change the regularization term to allow for outliers.

Total Variation Regularization

Approach:

- Replace the squared L2-norm of the Tikhonov regularizer by a L1-norm.

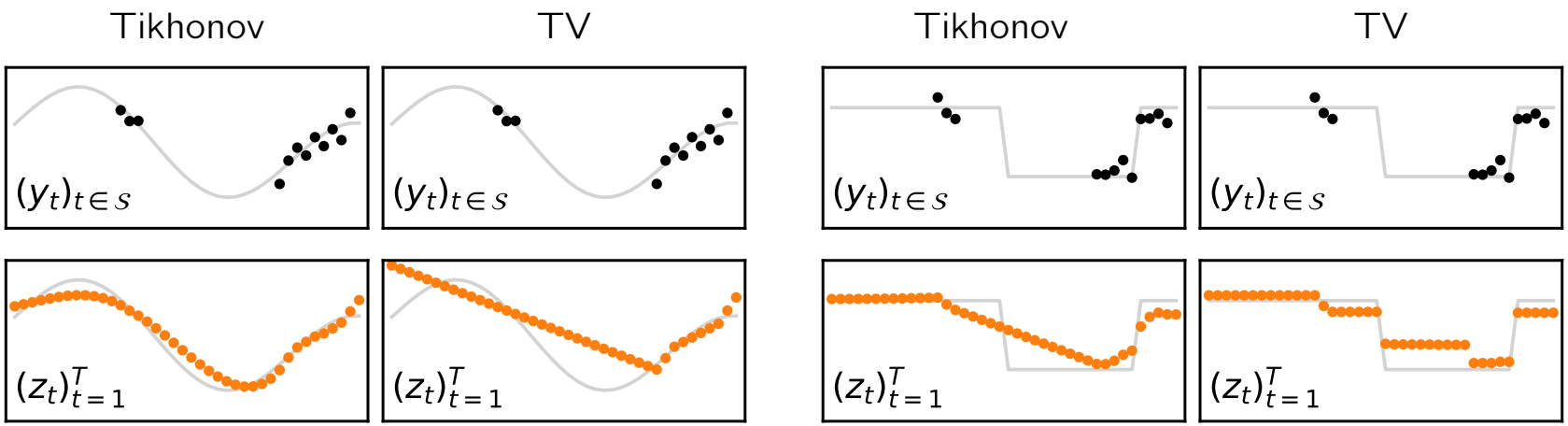
$$\arg\min_{\boldsymbol{z}} \frac{1}{2}\|M\boldsymbol{z} - \boldsymbol{y}\|^2 + \|\Gamma\boldsymbol{z}\|_1 + \frac{\eta}{2}\|X^\top \boldsymbol{w} - \boldsymbol{z}\|^2$$

- This is known as the *Total Variation* regularizer, and it rewards solutions that satisfy the smoothness condition almost everywhere except for a few locations in the sequence.
- In signal processing, the approach is known as 'Total Variation Denoising' [4].

Tikhonov vs. Total Variation Regularization

Scenario 1:
Low curvature and no abrupt transitions

Scenario 2:
Low slope and few abrupt transitions



Observation:

- Whether one should use Tikhonov ($\|\Gamma \boldsymbol{z}\|^2$) or Total Variation ($\|\Gamma \boldsymbol{z}\|_1$) is problem-specific.

Part 3

Further Topics

Multi-Dimensional Sequences

- Assume $\boldsymbol{z}$ is a two-dimensional predicted sequence of size $T \times T'$ (e.g. an synthesized image), and elements of that sequence are given using the indexing notation $z_{i,j}$.
- Then, one can define the new Tikhonov regularization scheme

$$\begin{aligned}\Omega(\boldsymbol{z}) &= \sum_{i=1}^{T-1}\sum_{j=1}^{T'}(z_{i+1,j} - z_{i,j})^2 + \sum_{i=1}^{T}\sum_{j=1}^{T'-1}(z_{i,j+1} - z_{i,j})^2 \\ &= \left\| \boldsymbol{z} * \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\|^2 + \left\| \boldsymbol{z} * \begin{bmatrix} 1 & -1 \end{bmatrix} \right\|^2\end{aligned}$$

Remarks:

- Other filters can be used to detect high curvature instead of high slope.
- To get the total variation penalty term, one can replace the squaring operations by absolute values.

Advanced Application: Style Transfer

Idea:

- Synthesize new images subject to a total variation penalty, in order to favor more asthetic images.

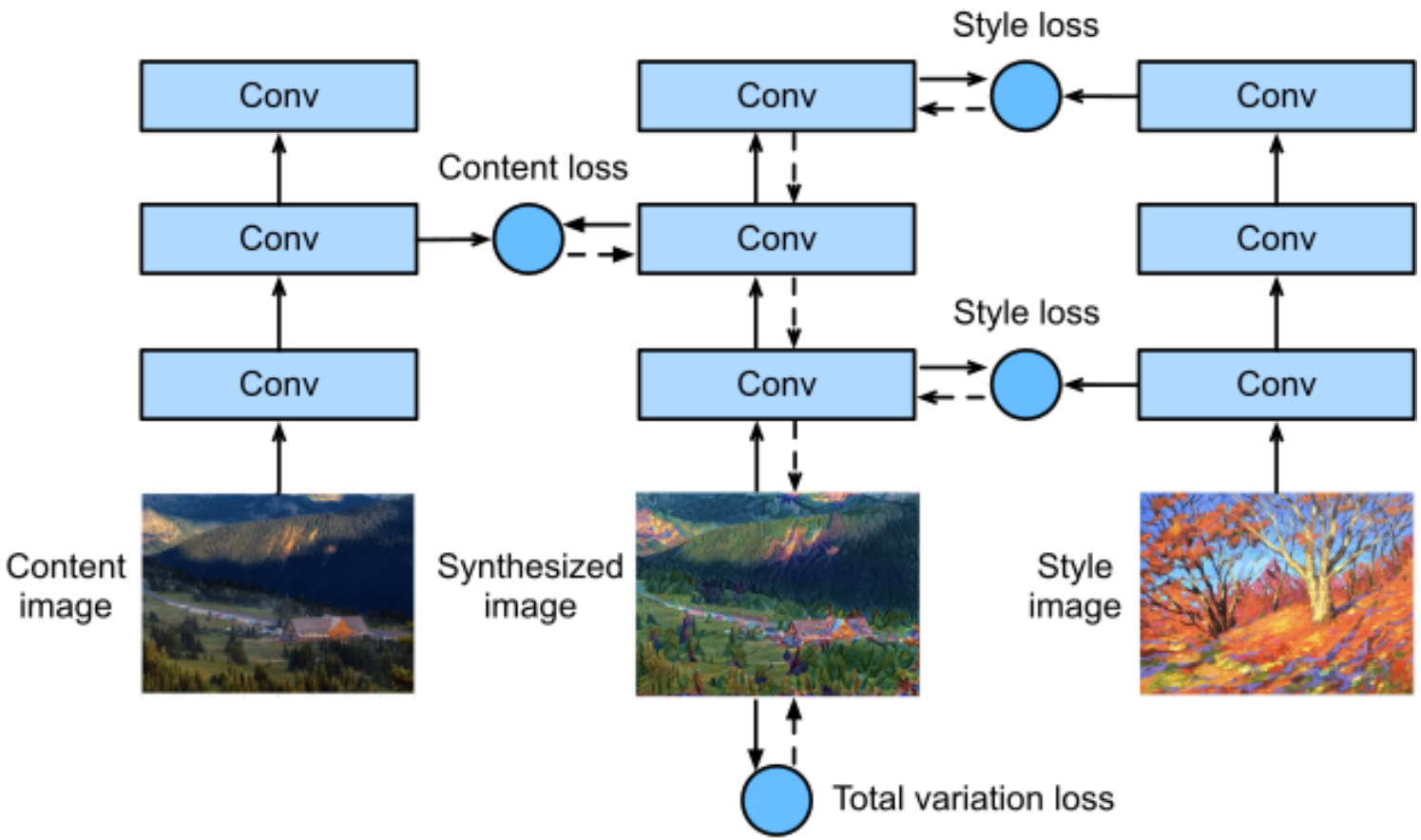
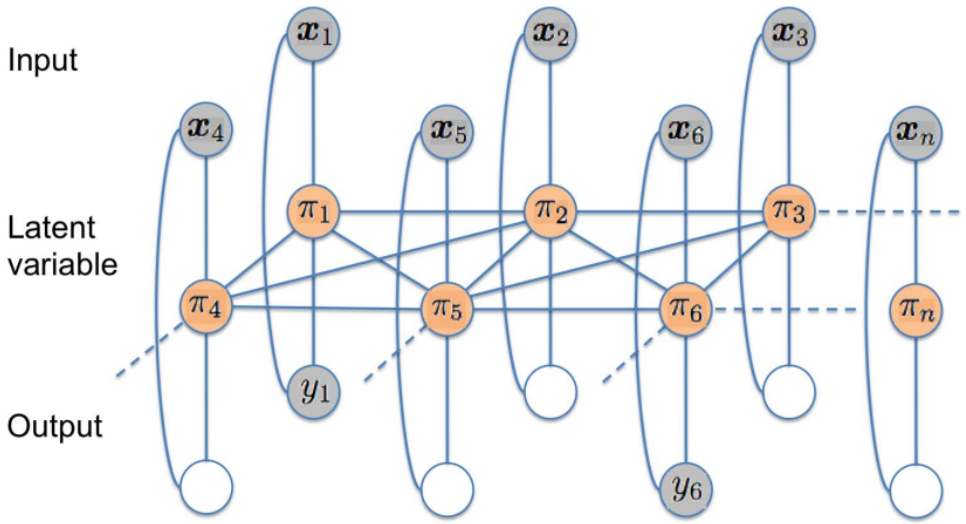
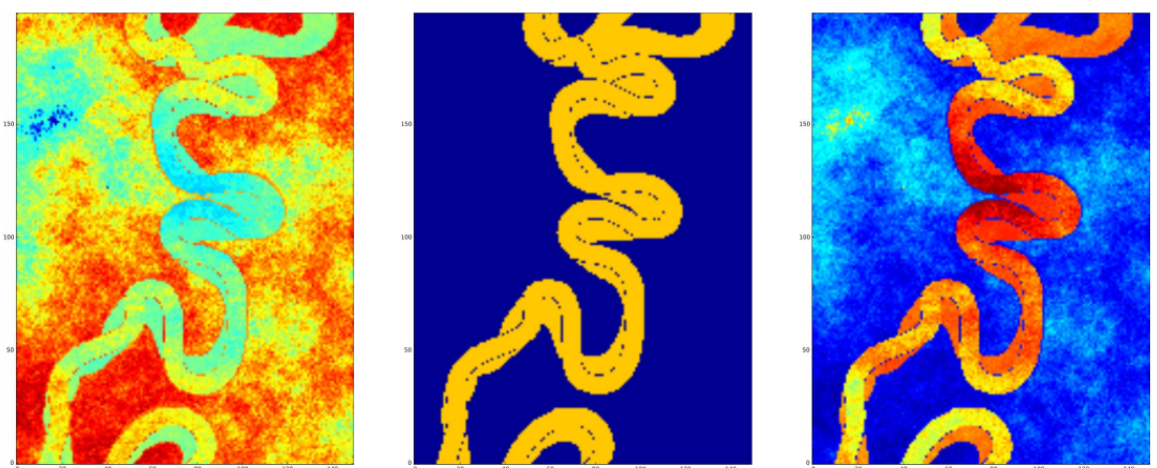


Image source: `https://d2l.ai/chapter_computer-vision/neural-style.html`

Advanced Application: Porosity Prediction [3, 1]



Note: Not based on a TV minimization formulation but on some probabilistic latent representation that also imposes limited changes in the facies representation.



(a) Impedance (input) (b) Facies (latent state) (c) Porosity (output)

Image source: doi:10.1109/TNNLS.2017.2700429

Two-Component Models

Motivation:

- Knowing that there is one abrupt change in the sequence we would like to identify it precisely (e.g. the exact index C where it occurs).

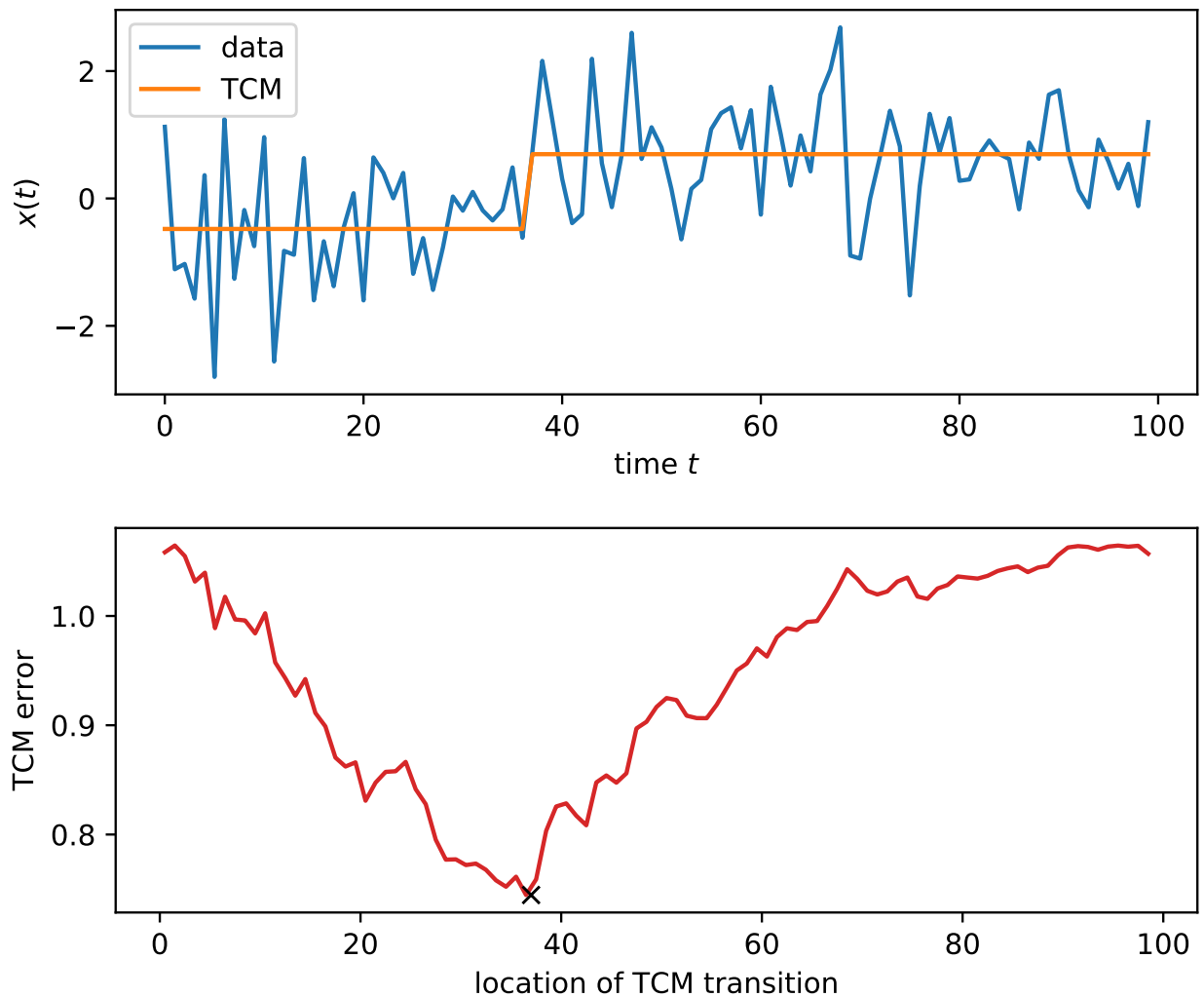
Approach:

- Build two separate models before and after the transition C, and measure the total error of the two models (cf. [2]).
- We then look for the index C and resulting model parameters that minimize that error.

  Example:
  $$\arg\min_{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, C} \left[\sum_{t=1}^{C} \|\boldsymbol{y}_t - \boldsymbol{\mu}_1\|^2 + \sum_{t=C+1}^{T} \|\boldsymbol{y}_t - \boldsymbol{\mu}_2\|^2 \right]$$
- If the number of time steps T is not too large, we can find the true optimum by search exhaustively for every change point C, and then for each C find the optimal model parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Two-Component Models

One-Dimensional Example:



Two-Component Models

Connection to K-means:

- The simple TCM above can be seen as a special case of k-means clustering solution with two clusters (and centroids $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$), and where cluster membership is constrained by ordering of data points.
- The TCM inherits the same decomposition as K-means, in particular, a decomposition of the total data dispersion in (1) the dispersion that is explained by the variance within components, and (2) the dispersion that is explained by the two components themselves.

Summary

Summary

- Many real-world datasets come in the form of sequences (e.g. images, volumes, time series).
- Sequential data often exhibits strong correlation between adjacent elements of the sequence. These correlations can be exploited by a ML algorithm to improve accuracy.
- Tikhonov regularization is an effective way of enforcing smoothness properties for the predicted sequence.
- Total variation regularization allows for a few smoothness violations in the sequence (e.g. edge in an image, change point in a time series).
- There are many more approaches for sequence modeling. One such approach is the two-component model. The latter has less flexibility but enables higher interpretability/reproducibility.

References

N. Görnitz, L. A. Lima, L. E. Varella, K. Müller, and S. Nakajima.

Transductive regression for data with latent dependence structure.
IEEE Trans. Neural Networks Learn. Syst., 29(7):2743–2756, 2018.

R. Killick and I. A. Eckley.

changepoint: An R package for changepoint analysis.
Journal of Statistical Software, 58(3):1–19, 2014.

L. A. Lima, N. Görnitz, L. E. Varella, M. M. B. R. Vellasco, K. Müller, and S. Nakajima.

Porosity estimation by semi-supervised learning with sparsely available labeled samples.
Comput. Geosci., 106:33–48, 2017.

L. I. Rudin, S. Osher, and E. Fatemi.

Nonlinear total variation based noise removal algorithms.
Physica D: Nonlinear Phenomena, 60(1-4):259–268, Nov. 1992.

A. N. Tikhonov.

Solution of incorrectly formulated problems and the regularization method.
Soviet Math. Dokl, 1963.