STATG019 – Selected Topics in Statistics 2018

Lecture 5

Validation principles for time series and probabilistic prediction tasks

Dr Franz J. Király

What is a time series?

Usual mathematical object model:

Time series taking values in $\mathcal{X}$ at times in $\mathcal{T} \subseteq \mathbb{R}$

is a collection of $\mathcal{X}$-valued random variables X_t

i.e., $(X_t; t \in \mathcal{T})$ where X_t t.v.in $\mathcal{X}$

Usual data storage object model:

$((x_1, t_1), \dots, (x_T, t_T)) \in$ seq$(\mathcal{X} \times \mathcal{T})$ “sequences with values in $(\mathcal{X} \times \mathcal{T})$”

t_i are time stamps at which x_i are observed

Mathematical notation will use the *mathematical* object model

(i.e., as usual in statistics, all data are random variables)

However, we will also use the data storage model as domain

i.e., $X = (X_t; t \in \mathcal{T})$ takes values in seq$(\mathcal{X} \times \mathcal{T})$

X_t is identified with an entry of the *ordered* tuple X

(i.e., also carries knowledge of its own time stamp)

Time series, sequences & Co.

Usual mathematical object model:

Time series taking values in $\mathcal{X}$ at times in $\mathcal{T} \subseteq \mathbb{R}$

is a collection of $\mathcal{X}$-valued random variables X_t

i.e., $(X_t; t \in \mathcal{T})$ where X_t t.v.in $\mathcal{X}$

Time series carry multiple layers of structural information:

complex

$X = (X_t; t \in \mathcal{T})$ t.v.in $\mathrm{seq}(\mathcal{X} \times \mathcal{T}) := \mathrm{series}(\mathcal{X})$

„forget time stamps“

„consider only time stamps“

$\mathrm{seq}(\mathcal{X})$ „sequence“

$\mathrm{seq}(\mathcal{T})$

„forget ordering“

$\mathrm{bag}(\mathcal{X}) = \mathcal{P}(\mathcal{X})$

„bag“ „set“ „empirical distribution“

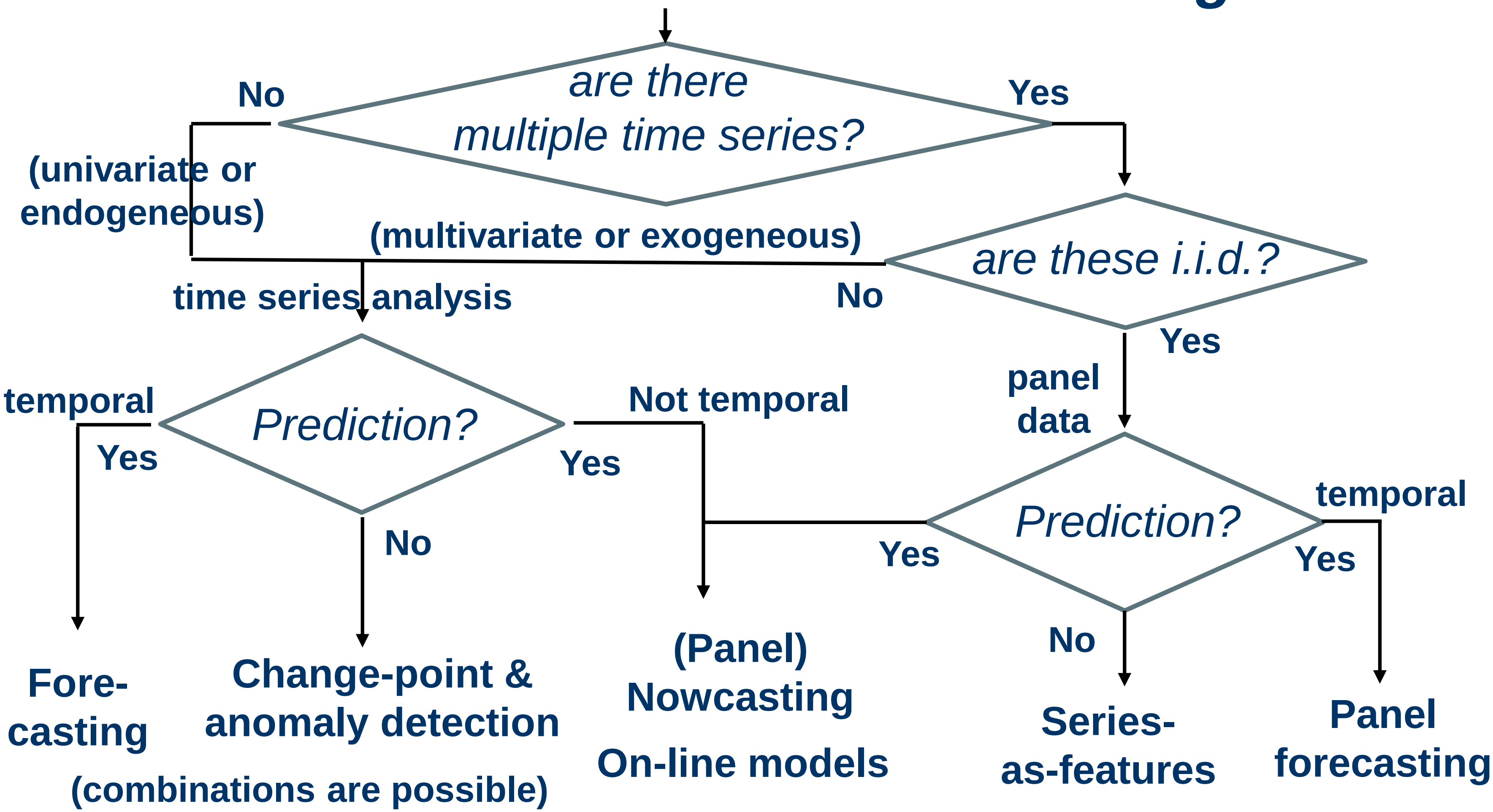
simple

Occam‘s razor: if ordering is irrelevant, don‘t use in model, etc

perhaps all information is contained in earliest time stamp?

Time series related modelling tasks

Crucial distinction: what is the scientific goal?



„complexity“ and validation depends on the task/setting!

The i.i.d. assumption divide

Two major „classes“ of assumptions/settings/tasks

(A) i.i.d. (panel) samples are available

series-as-features, or panel modelling tasks

It is *crucial* to make use of the i.i.d.-ness assumption!

models also trained on other time series will be better

validatory guarantees obtained from sample will be stronger

(B) no i.i.d. assumption can be made

in essence: one object, observed at subsequent time points

Alternative assumptions lead to *difficulties*, much is open

models using i.i.d. strategies perform badly

validatory guarantees are weak and rely on assumptions

which in addition require checking!

Model validation
in the i.i.d. settings

Time series tasks with i.i.d. data: overview

i.i.d. on-line learning: data is temporally revealed but actually i.i.d.

Algorithm interface has fit, predict/trafo, and *update* for new data

Validation is as in the i.i.d. case – performance statistics get updated

Below: time-series-within-dataset „panel dataset“

	***energy use** (time series)*	***user type** (categorical)*
1		*residential*
2		*factory*

Series-as-features-task: series appear only as features

Sub-cases: supervised prediction, unsupervised learning

Related fields: functional data analysis, kernels, signal processing

Validation almost exactly like the „standard feature type“ case

Panel data modelling: prediction/labelling task *within* series

Sub-cases: panel forecasting, panel anomaly/change-point d.

More technical but „easier“ than non-panel forecasting etc

Validation „along the i.i.d. axis“ easily generalizes tabular i.i.d case

Supervised learning with series-at-features

Observations $(X_1, Y_1), \dots, (X_N, Y_N) \underset{\text{i.i.d.}}{\sim} (X, Y)$

„primitive features“

t.v.in $(\text{series}(\mathcal{X}) \times \mathcal{X}') \times \mathcal{Y}$

	energy use *(time series)*	user type *(categorical)*
1		*residential*
2		*factory*

Estimate/learn

prediction f t.v.in $[(\text{series}(\mathcal{X}) \times \mathcal{X}') \to \mathcal{Y}]$

Such that the expected generalization error

$\varepsilon(f) := \mathbb{E}\left[L\left(f(X), Y\right)\right]$ is small

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is *one of the usual losses*

Validation: use conditionally i.i.d. test sample $L\left(f(X_i^*), Y_i^*\right)$

(just as in the „normal“ supervised learning case)

Validation caveat: there is a hierarchy of baselines implied by

complex

$X = (X_t; t \in \mathcal{T})$ t.v.in $\text{seq}(\mathcal{X} \times \mathcal{T}) := \text{series}(\mathcal{X})$

„forget time stamps“

$\text{seq}(\mathcal{X})$ „sequence“

„consider only time stamps“

$\text{seq}(\mathcal{T})$

„forget ordering“

$\text{bag}(\mathcal{X}) = \mathcal{P}(\mathcal{X})$

simple

Baselines in the series-as-features setting

Observations $(X_1, Y_1), \dots, (X_N, Y_N) \underset{\text{i.i.d.}}{\sim} (X, Y)$

t.v.in $(\text{series}(\mathcal{X}) \times \mathcal{X}') \times \mathcal{Y}$

Estimate/learn

a prediction functional f t.v.in $[(\text{series}(\mathcal{X}) \times \mathcal{X}') \to \mathcal{Y}]$

Validation caveat: there is a hierarchy of baselines/methods

in order of „complexity“ (later methods should improve on earlier ones)

simple → complex

0. methods not using the series feature at all t.v.in $[\mathcal{X}' \to \mathcal{Y}]$

1a. methods using only simple summaries of time stamps

1b. methods using only simple summaries of the „bag“/set

2. methods using only the sequence of time stamps

3. methods using only the bag/set t.v.in $[\text{bag}(\mathcal{X}) \times \mathcal{X}' \to \mathcal{Y}]$

4. methods using only the sequence t.v.in $[\text{seq}(\mathcal{X}) \times \mathcal{X}' \to \mathcal{Y}]$

5. genuine series methods, not of the above type

The panel data learning setting

Observations $(X_1, Z_1), \dots, (X_N, Z_N) \underset{\text{i.i.d.}}{\sim} (X, Z)$ t.v.in series$(\mathcal{X}) \times \mathcal{Z}$

Estimate/learn a decision rule f t.v.in $[\text{series}(\mathcal{X}) \times \mathcal{Z} \to \mathcal{S}]$

$\mathcal{S}$ is a set of decisions depending on the task

e.g., **(i)** f t.v.in $[\text{series}(\mathcal{X}) \times \mathcal{Z} \to \text{series}(\mathcal{X})]$ „panel forecasting"

$(x_{\le\tau}, z) \mapsto \widehat{x}_{>\tau}$ where $(x_{\le\tau}$ is "past", z "features", $\widehat{x}_{>\tau}$ "future"

(ii) f t.v.in $[\text{series}(\mathcal{X}) \times \mathcal{Z} \to \text{series}(\{\text{yes}, \text{no}\})]$

„anomaly detection on panel data" e.g., heart rate monitor

Such that the expected generalization error

$\varepsilon(f) := \mathbb{E}\left[L\left(f(X, Z), Y\right)\right]$ is small with Y ground truth

and $L : \mathcal{S} \times \mathcal{Y} \to \mathbb{R}$ an appropriate loss

Validation: use conditionally i.i.d. *test* sample $L\left(f(X_i^*, Z_i^*), Y_i^*\right)$

Task specific challenges are usually: defining loss, ground truth signal

Model validation in genuinely temporal settings

Time series forecasting setting

All data: $X = (X_t \, ; t \in \mathcal{T})$ with $\mathcal{T} \subseteq \mathbb{R}$ t.v.in series$(\mathcal{X})$

Observations: $X_{\le \tau} := (X_t \, ; \, t \in \mathcal{T}, t \le \tau)$

Predict/forecast: $X_{>\tau} := (X_t \, ; \, t \in \mathcal{T}, t > \tau)$

by $\widehat{X}_{>\tau} := \left(\widehat{X}_t \, ; \, t \in \mathcal{T}, t > \tau\right)$

Such that: the look-forward forecast errors

$\varepsilon(t) := \mathbb{E}\left[L\left(\widehat{X}_t, X_t\right)\right]$ are small, $L : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

Task variant 1: cut-off forecasting τ is *fixed* „present“

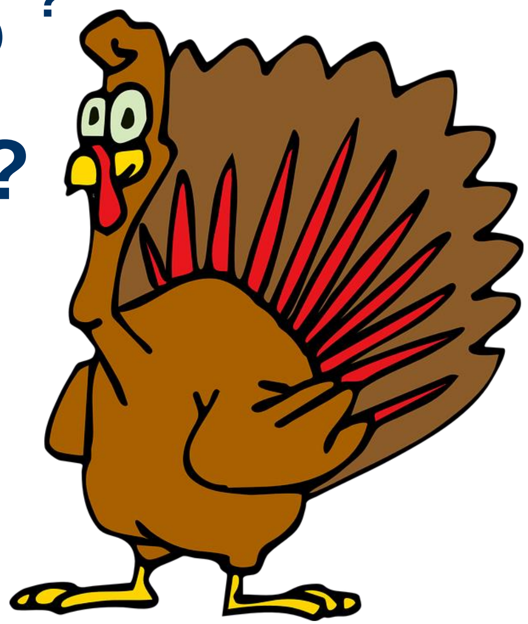
$\widehat{X}_{>\tau}$ and $X_{>\tau}$ independent conditional $\widehat{X}_{\le \tau}$ „don‘t use the future“

Task variant 2: sliding window forecasting τ *moves forward*

$\widehat{X}_{\tau}$ and $X_{>\tau}$ independent conditional $\widehat{X}_{\le \tau}$, for all τ

„don‘t use the future of *a given prediction time point*“

Bertrand Russell‘s turkey



The turkey wants to forecast whether
it is still going to be alive tomorrow.

Starting with the first day of November,

$$X_1 = \text{yes}, X_2 = \text{yes}, X_3 = \text{yes}, \ \dots,$$
$$X_{\text{day before the fourth Thursday of November}} = \text{yes}$$

Assuming the Turkey is an American,
what is the most sensible prediction for the next day?

More generally, what to do if

$$X_1, X_2, \dots, X_\tau \underset{\text{i.i.d.}}{\sim} X \text{ but } X_{\tau+1} \text{ may be completely different?}$$

There‘s nothing one *can* do! … in general.

i.e., generalizability *assumptions* are needed.

Ways out of the paradox

More generally, what to do if

$X_1, X_2, \ldots, X_\tau \underset{\text{i.i.d.}}{\sim} X$ but $X_{\tau+1}$ may be completely different?

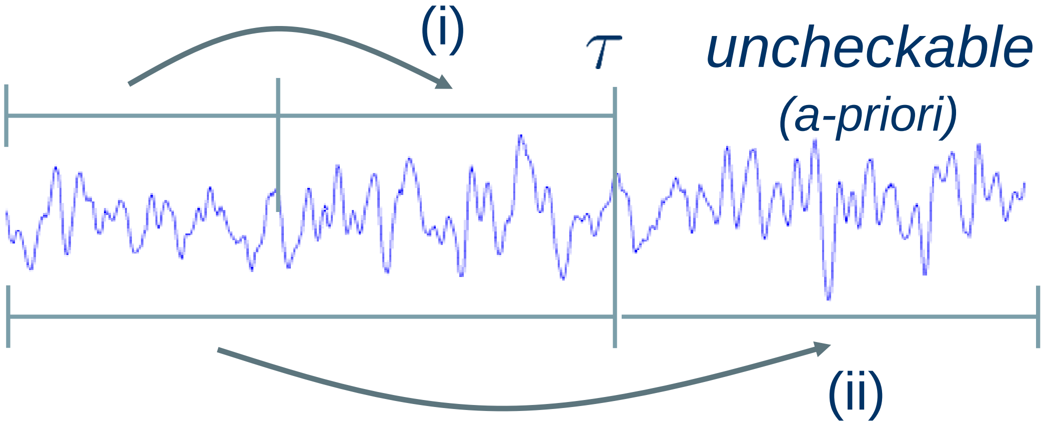
There's nothing one *can* do! … in general.

i.e., generalizability *assumptions* are needed.

Two parts of the assumption: (even in the i.i.d. case!)

(i) the observed data are similar/related - *testable*

(ii) future data are similar/related to observed data



uncheckable … but failure detectable!

(a-priori) *(a-posteriori)*

Time series – generalizability assumptions

Two parts of the assumption:

(i) the observed data are similar/related - *testable*

(ii) future data are similar/related to observed data

Prototypical (but very strong) assumption:

the full series $X = (X_t\, ; t\ \in \mathcal{T})$ is *(strongly) stationary*

Definition: $X = (X_t\, ; t\ \in \mathcal{T})$ is called *(strongly) stationary*

iff the joint law of $X_{S+\delta} = (X_{t+\delta}\, ; t\ \in S)$ where $\delta \in \mathbb{R}, S \subseteq \mathcal{T}$

does not depend on the choice of S, δ

Intuitively: shifting the timestamps by δ does not change statistical properties

Testable by: Unit root tests, spectrum analysis tests

(e.g., Dickey-Fuller) *(e.g., Priestly-Subba Rao)*

Cave: stationarity is strong assumption which is rarely true

but validatory framework has all time series characteristic features

Time series – generalizability assumptions

Two parts of the assumption:

(i) the observed data are similar/related - *testable*

(ii) future data are similar/related to observed data

Prototypical (but very strong) assumption:

the full series $X = (X_t \, ; t \ \in \mathcal{T})$ is *(strongly) stationary*

Testable by: Unit root tests, spectrum analysis tests

(e.g., Dickey-Fuller) *(e.g., Priestly-Subba Rao)*

Cave: stationarity is strong assumption which is rarely true

but validatory framework has all time series characteristic features

Including the „feature“ that validation in time series is difficult:

General tests for the assumption only known in univariate case

Change-point analysis („stationarity breaks“) open research topic

Guarantees and confidence intervals

Recall reason for main guarantees for the i.i.d. case:

(and assume all moments exist)

Central Limit Theorem: Let $X_1, \dots, X_N \underset{\text{i.i.d.}}{\sim} X$

Let $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then $\sqrt{N}\left(\widehat{\mu} - \mathbb{E}[X]\right) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}(X)\right)$ as $N \to \infty$

Central Limit Theorem for stationary time series:

Let $X = (X_t\,; t \in \mathbb{Z})$ be (strongly) stationary (and assume joint mgf is total)

Let $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then $\sqrt{N}\left(\widehat{\mu} - \mathbb{E}[X_{42}]\right) \xrightarrow{d} \mathcal{N}\left(0, \rho \cdot \mathrm{Var}(X_7)\right)$ as $N \to \infty$

where $\rho = \sum_{i=-\infty}^{\infty} \mathrm{Corr}(X_0, X_i) = 1 + 2\sum_{i=1}^{\infty} \mathrm{Corr}(X_0, X_i)$ and where $\mu = \mathbb{E}[X_{42}]$

Proof sketch: Consider joint mgf $M_X(t_1, \dots, t_N)$ of $(X_1, \dots, X_N) - \mu$

and mgf $M_{\widehat{\mu}}(t)$ of $\widehat{\mu} - \mu$

$$M_{\widehat{\mu}}(t) = M_X\left(\frac{t}{N}, \dots, \frac{t}{N}\right) = \frac{t^2}{N^2} \cdot \mathbb{1}^\top \begin{pmatrix} \rho_0 & \rho_1 & \rho_2 & \cdots \\ \rho_{-1} & \rho_0 & \rho_1 & \cdots \\ \rho_2 & \rho_{-1} & \rho_0 & \cdots \\ \vdots & & & \ddots \end{pmatrix} \mathbb{1} + O(N^{-3})$$

„avg of entries“

where $\rho_i = \mathrm{Cov}(X_0, X_i)$

Guarantees and confidence intervals

Central Limit Theorem for stationary time series:

Let $X = (X_t \,; t \in \mathbb{Z})$ be (strongly) stationary (and assume joint mgf is total)

Let $\widehat{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then $\sqrt{N}\,(\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}\,(0, \rho \cdot \mathrm{Var}(X_7))$ as $N \to \infty$

where $\rho = \sum_{i=-\infty}^{\infty} \mathrm{Corr}(X_0, X_i) = 1 + 2\sum_{i=1}^{\infty} \mathrm{Corr}(X_0, X_i)$ and where $\mu = \mathbb{E}[X_{42}]$

Consequence: $\widehat{\mu}$ estimates μ consistently, with normal asymptotic

$\rho \cdot \mathrm{Var}[X_7]$ may be estimated as mean of $\sum_s X_s X_t - \widehat{\mu}^2$

Confidence intervals similar to the i.i.d. setting (additional assumption: decay of autocorrelation)

N/ρ is an „effective sample size“: more correlation, wider CI

CI and tests for location comparisons as in i.i.d. setting

but with effective sample size correction for variance

The Diebold-Mariano test for comparing forecasts

Forecast test set: $X = (X_t \; ; \; t \in \mathcal{T})$ where $\mathcal{T} = \{1, \dots, T\}$

Two predictions to compare: $\widehat{X}^{(1)}$ and $\widehat{X}^{(2)}$, times also $\mathcal{T}$

1. compute loss residuals $L^{(i)} := \left(L(\widehat{X}_t^{(i)}, X_t) \; : \; t \in \mathcal{T} \right)$

2. compute mean estimates $\widehat{\mu}^{(i)} := \sum_{t=1}^{T} L_t^{(i)}$ „performances“

difference = „effect size “

3. compute average autocorrelation $\widehat{v}^{(i)} := \sum_{s=1}^{T} \sum_{t=1}^{T} L_t^{(i)} L_s^{(i)} - \left(\widehat{\mu}^{(i)}\right)^2$

4. α-CI: $\left[\widehat{\mu}^{(i)} + \Phi^{-1}(\alpha/2) \cdot \sqrt{\widehat{v}^{(i)}/T} \, , \widehat{\mu}^{(i)} - \Phi^{-1}(\alpha/2) \cdot \sqrt{\widehat{v}^{(i)}/T} \right]$

5. run two-sided t-test with statistic $\dfrac{\widehat{\mu}^{(1)} - \widehat{\mu}^{(2)}}{\sqrt{\widehat{v}^{(1)} + \widehat{v}^{(2)}}}$ „DM-statistic“

Cave: *compares (sliding) predictions and not strategies!*

(same restrictions as in supervised setting, lecture 2 apply)

Forecasting workflow & validation principles

Checking/testing & keeping track of assumptions is important

For comparison, need to hold only on losses or loss differences

i.i.d.-ness can hold for losses: (WW-)run tests, spectrum analysis tests

Frequent assumptions: (weak/strong) stationarity

ergodicity & mixingness to take care of „coefficient decay“

plus local, periodic and difference versions of the above

Re-sampling schemes need to be temporal! „don‘t predict from future“

Sliding window/shift aggregation compatible with CLTs

necessary for validation, and usually beneficial in tuning

Generally, a broadly open (and difficult) research field…

Full understanding of a model-agnostic validatory workflow is missing

the „right assumptions“ for broad usefulness are unknown/open

should be weaker than stationarity, but stronger than turkey (= none)

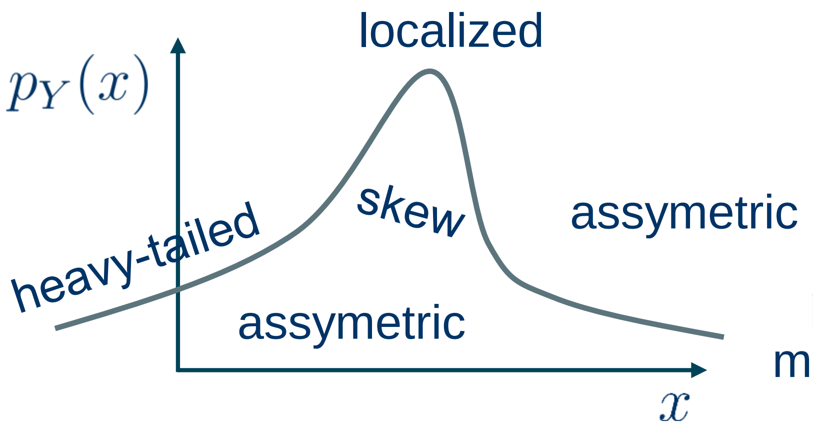
Exhaustive list of ML toolboxes for time series:

(i.e., with a unified sklearn-like interface)

Theory of loss functionals for probabilistic prediction

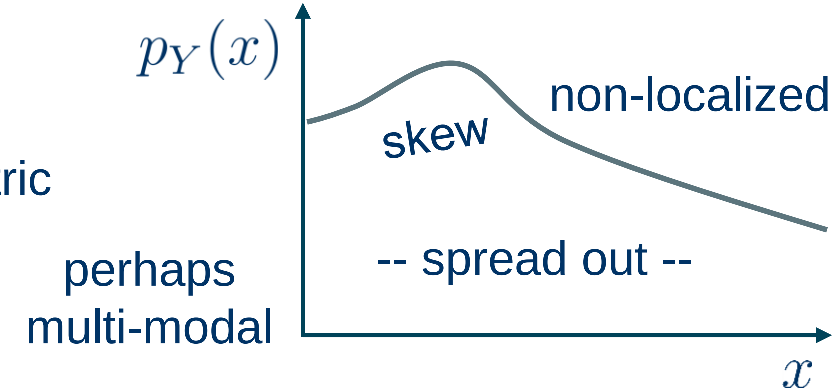
The need for probabilistic prediction

arising from specifics of the data generative process in applications:



Stylized case: „finance“

(e.g., asset price and portfolio modelling)



Stylized case: „medicine“

(e.g., survival and time-to-event modelling)

Both cases: modelling of shape, skewness, and spread *crucial*

Finance: „need to model higher-order moments“ or „tail behaviour“

Medicine: „need to model hazard“ or „survival function“

Mathematical approach: model the *conditional law* of $Y|X = x$

i.d., conditional distribution functional $p_{Y|X}(y|x)$

Point prediction losses & elicitation

Can we use the standard setting for modelling & validation?

Setting: want to predict Y t.v.in $\mathcal{Y}$ prediction $y \in \mathcal{Y}$ (temporarily: $\mathcal{Y} \subseteq \mathbb{R}$)

goodness assessed through convex loss $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

(recall: convex in 1st argument = prediction, by definition)

Lemma: $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is convex iff

$L(\mathbb{E}[Z], y) \leq \mathbb{E}[L(Z, y)]$ for all r.v. Z t.v.in $\mathcal{Y}$ and $y \in \mathcal{Z}$

Proof: Jensen‘s inequality/lemma applied to $[z \mapsto L(z, y)]$

Examples: (of convex losses)

$L : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$ „squared loss“

$L : (\widehat{y}, y) \mapsto |\widehat{y} - y|$ „absolute loss“

Definition: Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a convex loss functional.

(Gneiting, 2000s)

L *elicits* the statistic T : Distr$(\mathcal{Y}) \to \mathbb{R}$

if $T(F_Y) = \underset{y \in \mathcal{Y}}{\arg\min}\, \mathbb{E}[L(y, Y)]$ for all r.v. Y t.v.in $\mathcal{Y}$

(where F_Y denotes cdf of Y)

Point prediction losses & elicitation

Setting: want to predict Y t.v.in $\mathcal{Y}$ prediction $y \in \mathcal{Y}$ (temporarily: $\mathcal{Y} \subseteq \mathbb{R}$)

goodness assessed through convex loss $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Examples: (of convex losses)

$L_{sq} : (\widehat{y}, y) \mapsto (\widehat{y} - y)^2$ „squared loss"

$L_{abs} : (\widehat{y}, y) \mapsto |\widehat{y} - y|$ „absolute loss"

Definition: Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a convex loss functional.

(Gneiting, 2000s)

L *elicits* the statistic $T : \mathrm{Distr}(\mathcal{Y}) \to \mathbb{R}$

if $T(F_Y) = \underset{y \in \mathcal{Y}}{\arg\min}\, \mathbb{E}[L(y, Y)]$ for all r.v. Y t.v.in $\mathcal{Y}$ (well-defined since convex)

Examples: (of elicitation)

squared loss elicits mean: $\mathbb{E}(Y) = \underset{y \in \mathcal{Y}}{\arg\min}\, \mathbb{E}[L_{sq}(y, Y)]$

absolute loss elicits median: $\mathrm{median}(Y) = \underset{y \in \mathcal{Y}}{\arg\min}\, \mathbb{E}[L_{abs}(y, Y)]$

Intuitively: squared and absolute loss

measure how well *location* is predicted!

useless to validate prediction of distributional features! *tails, skew, etc*

Quantile losses for tail predictions

Setting: want to predict Y t.v.in $\mathcal{Y}$ prediction $y \in \mathcal{Y}$ (temporarily: $\mathcal{Y} \subseteq \mathbb{R}$)

goodness assessed through convex loss $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Definition: Let $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a convex loss functional.

(Gneiting, 2000s) L *elicits* the statistic T : Distr$(\mathcal{Y}) \to \mathbb{R}$

if $T(F_Y) = \underset{y \in \mathcal{Y}}{\arg\min}\, \mathbb{E}[L(y, Y)]$ for all r.v. Y t.v.in $\mathcal{Y}$

Definition: $L_\alpha : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}\ ;\ (\widehat{y}, y) \mapsto \alpha \cdot m(y, \widehat{y}) + (1 - \alpha) \cdot m(\widehat{y}, y)$

is called (α-)quantile loss where $m(x, z) = \min(x - z, 0)$

Proposition: L_α elicits the α-quantile

that is, $F_Y^{-1}(\alpha) = \underset{y \in \mathcal{Y}}{\arg\min}\, \mathbb{E}[L_\alpha(y, Y)]$

useful to validate „value at risk“ predictions!

Problems: for „full picture“, multiple predictions/losses needed

quantile measures may have high loss variance (thesis?)

Fully probabilistic predictions

Setting: want to predict Y t.v.in $\mathcal{Y}$

e.g., absolutely continuous (pdf exist)

prediction $p \in \mathcal{P} \subseteq \text{Distr}(\mathcal{Y})$ i.e., p is a distribution in $\mathcal{P}$

goodness assessed through loss $L : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$

Definition: Let $L : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ be a convex loss functional.

(Dawid, 1990s)

L is called *strictly proper* if for all r.v. Y t.v.in $\mathcal{Y}$

$$p_Y = \arg\min_{p \in \mathcal{P}} \mathbb{E}[L(p, Y)] \text{ where } Y \sim p_Y$$

(note: definition is w.r.t. $\mathcal{P}$)

Intuition: best prediction (w.r.t. expected loss) is „true“ distribution

Examples: $L_{log} : (p, y) \mapsto -\log p(y)$

(of strictly proper losses)

„logarithmic loss“ „cross-entropy loss“

Note: defined only for discrete or absolutely continuous distributions

$$L_{sq} : (\widehat{y}, y) \mapsto -2p(y) + \|p\|_2^2$$

„squared integrated loss“ „Brier loss (if discrete)“

where $p(y)$ is shorthand for pdf/pdf at y and $\|p\|_2^2 := \int_{\mathcal{Y}} p(y)^2 \, dy$

Fully probabilistic predictions

Setting: want to predict Y t.v.in $\mathcal{Y}$

e.g., absolutely continuous (pdf exist)

prediction $p \in \mathcal{P} \subseteq \mathsf{Distr}(\mathcal{Y})$ i.e., p is a distribution in $\mathcal{P}$

goodness assessed through loss $L : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$

Examples: (of strictly proper losses)

$$L_{log} : (p, y) \mapsto -\log p(y)$$

„logarithmic loss" „cross-entropy loss"

Note: defined only for discrete or absolutely continuous distributions

$$L_{isq} : (p, y) \mapsto -2p(y) + \|p\|_2^2$$

„squared integrated loss" „Brier loss (if discrete)"

Short proofs of strict properness:

$$\mathbb{E}[L_{log}(p, Y)] - \mathbb{E}[L_{log}(p_Y, Y)] = \int_{\mathcal{Y}} p_Y \cdot \log \frac{p_Y(y)}{p(y)}\, dy = \mathsf{D}_{KL}(p \| p_Y)$$

then use Gibbs' (strict) inequality

„cross-entropy"

„Kullback-Leibler-divergence"

$$\mathbb{E}[L_{isq}(p, Y)] - \mathbb{E}[L_{isq}(p_Y, Y)] = \int_{\mathcal{Y}} (p(y) - p_Y(y))^2 \; dy = \mathsf{D}_{sq}(p, p_Y)$$

then use school math

„integrated squared distance/divergence"

Example: validation of unsupervised methods

Unsupervised setting (full probabilistic model):

Given data $X_1, \dots, X_N \underset{\text{i.i.d.}}{\sim} X \sim p \in \mathcal{P} \subseteq \text{Distr}(\mathcal{X})$

Estimate p via $\widehat{p} \in \mathcal{P}$ (using the data X_i)

(assume discrete, or absolutely continuous)

Goodness assessed through probabilistic loss $L : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$

i.e., $\mathbb{E}[L(\widehat{p}, X)]$ is small.

Validation: use conditionally i.i.d. test sample $L\left(\widehat{p}, X_i^*\right)$

Note: $\frac{1}{M} \sum_{i=1}^{M} L\left(\widehat{p}, X_i^*\right)$ for the log/cross-entropy loss

is the negative out-of-sample likelihood of $\widehat{p}$, up to a constant factor

Cave: this validation strategy applies only to unsupervised methods with a full probabilistic model

WORQ - widely open research questions!

(aka BSc/MSc/PhD topics for theory/practice cross-over inclined)

Completing the model-agnostic validation workflows

What is the right formulation for a fully model-agnostic setting?

Which assumptions are general enough but not too general to be useful?

Right combination of change-point detection and assumption testing?

Basic toolbox design for the time series related tasks

What are the right data containers and data pipelines?

How to encapsulate and interface the different tasks, models, strategies?

Combination of time series, on-line, probabilistic, and heterogeneity?

Meta-learning for time series related tasks

What are good re-sampling schemes, and related learning guarantees?

Tuning, ensembling and composition for time series & probabilistic losses

Composition of prediction and anomaly/change-point detection methods?

Also: systematic study of which methods *really* work

Deep learning for everything? Conditional on complete validation workflow …