Conformal Prediction for detecting hallucinations

Klea Panayidou
Cosmology and Statistics

TITAN ARGOS TOSCA workshop

Paris 2024

Back to the basic principles

All models are wrong, but some are useful

critically (know) ***when*** *the model succeeds/fails and* ***why***

Back to the basic principles

All models are wrong, but some are useful

critically (know) ***when*** *the model succeeds/fails and* ***why***

What about (black box) deep learning?

Back to the basic principles

All models are wrong, but some are useful

critically (know) ***when*** *the model succeeds/fails and* ***why***

What about (black box) deep learning?

Your output should always be delivered with its uncertainty

Quantify uncertainty and pay attention to it

Sir David Cox

Back to the basic principles

All models are wrong, but some are useful

critically (know) ***when*** *the model succeeds/fails and* ***why***

What about (black box) deep learning?

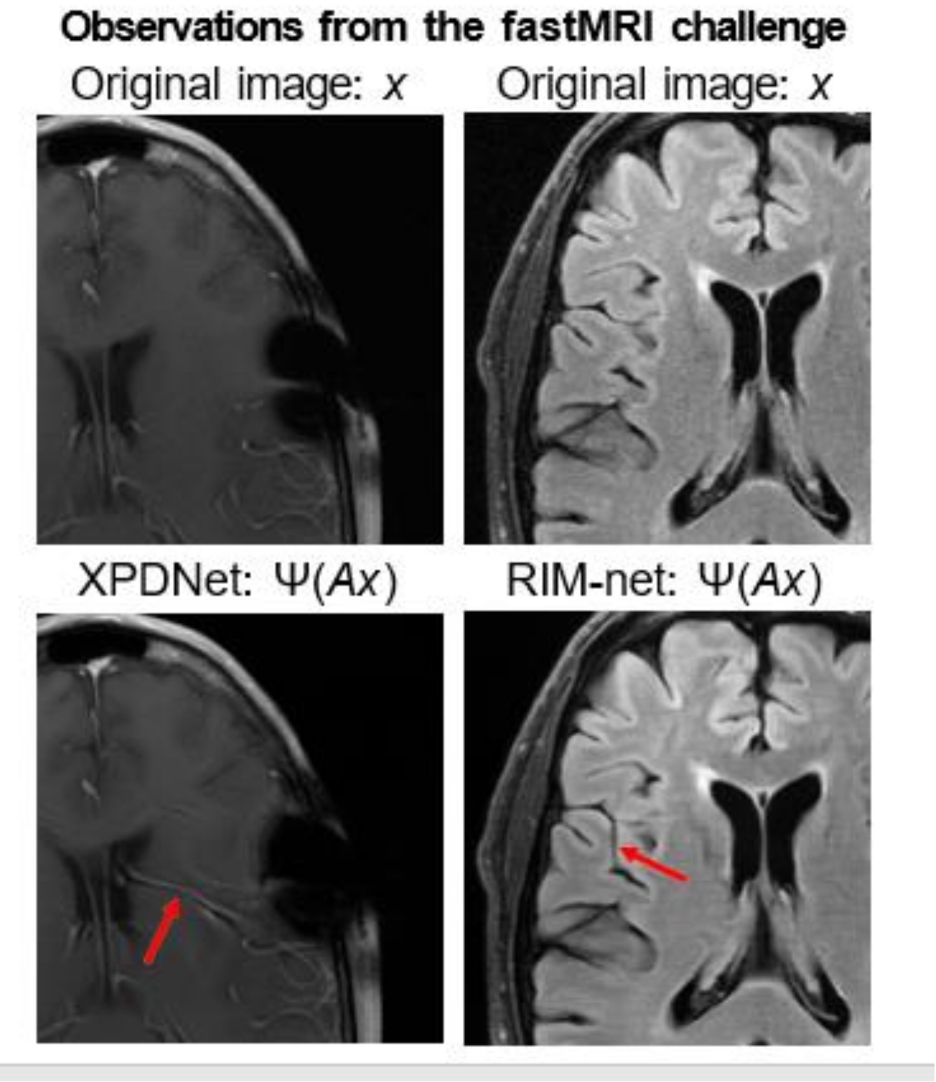
Your output should always be delivered with its uncertainty

Quantify uncertainty and pay attention to it

Can we quantify uncertainty in a meaningful way?

Hallucinations

Concerns In Medical Imaging



Hallucinations

Concerns In Medical Imaging

"The potential lack of generalization of deep learning-based reconstruction methods as well as their innate unstable nature may cause false structures to appear in the reconstructed image that are absent in the object being imaged"
— In "On hallucinations in tomographic image reconstruction", *IEEE T. Med. Imaging* (2021)

In Microscopy

"[...] These hallucinations are deceptive artifacts that appear highly plausible in the absence of contradictory information and can be challenging, if not impossible, to detect."

— In "Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction", *Nature Methods* (2019)

At first, detecting hallucinations felt like

Hallucinations

(Inverse) Problem: Image Reconstruction

Given measurements $y = Ax + e \in \mathbb{C}^m$, of $x \in \mathcal{M}_1$, recover x.

Can we understand and prevent hallucinations?

Hallucination busters

(Inverse) Problem: Image Reconstruction

Given measurements $y = Ax + e$, recover x .

accuracy-stability trade-off for inverse problems

if the accuracy of a method is pushed too far (e.g., by driving the training error to zero), it inevitably becomes unstable.

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

In-distribution hallucinations, yet existence of non-hallucinating algorithm

Theorem 2

Let $A \in \mathbb{C}^{m\times N}$ with $1 \leq \mathrm{rank}(A) < N$, $\mathcal{T} \subset \mathbb{C}^N$ be a non-empty and finite set, $\delta > 0$, $\Psi\colon \mathbb{C}^m \to \mathbb{C}^N$ be a neural network with Lipschitz constant $L > 0$ and $x_{\mathrm{Det}} \in \mathbb{C}^N$ with $\|Ax_{\mathrm{Det}}\| \leq \delta/(4L)$. Suppose that Ψ satisfies

$$\max_{x\in\mathcal{T}} \|\Psi(Ax) - x\| \leq \delta.$$

Then, for any $\epsilon \geq \delta/(2L)$ there is an uncountable family $\mathcal{C}$ of finite or countably infinite sets $\mathcal{M}_1 \subset \mathbb{C}^N$ with $\mathcal{T} \subset \mathcal{M}_1$ and $A\mathcal{M}_1 \subset \mathcal{B}_{\|\cdot\|}(A\mathcal{T}, \epsilon)$, such that for each $\mathcal{M}_1 \in \mathcal{C}$ the following hold simultaneously.

(i) ***(Ψ suffers from in-distribution hallucinations)****. For any probability distribution $\mathcal{D}$ on $\mathcal{M}_1$ with the property that $\mathbb{P}_{X\sim\mathcal{D}}(X \in \mathcal{T}) \leq q$, it holds that*

$$\mathbb{P}_{X\sim\mathcal{D}}\left(\exists\lambda \in \mathbb{C}, |\lambda| = 1 \text{ such that } \|\Psi(AX) - (X + \lambda x_{\mathrm{Det}})\| \leq 2\delta\right) \geq 1 - q.$$

(ii) *(There exists an algorithm that yields non-hallucinating NNs). There exists an algorithm Γ taking inputs in $A(\mathcal{M}_1)$, such that for each $y \in A(\mathcal{M}_1)$, $\Gamma(y) = \Phi_y$ is a NN $\Phi_y\colon \mathbb{C}^m \to \mathbb{C}^N$ that satisfies*

$$\|\Phi_{Ax}(Ax) - x\| \leq \delta, \quad \forall x \in \mathcal{M}_1.$$

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

In-distribution hallucinations

Theorem 2

Let $A \in \mathbb{C}^{m\times N}$ with $1 \leq \mathrm{rank}(A) < N$, $\mathcal{T} \subset \mathbb{C}^N$ be a non-empty and finite set, $\delta > 0$, $\Psi: \mathbb{C}^m \to \mathbb{C}^N$ be a neural network with Lipschitz constant $L > 0$ and $x_{\mathrm{Det}} \in \mathbb{C}^N$ with $\|Ax_{\mathrm{Det}}\| \leq \delta/(4L)$. Suppose that Ψ satisfies

$$\max_{x\in\mathcal{T}} \|\Psi(Ax) - x\| \leq \delta.$$

Then, for any $\epsilon \geq$ … ets $\mathcal{M}_1 \subset \mathbb{C}^N$ with $\mathcal{T} \subset \mathcal{M}_1$ ar… …ultaneously.

(i) **(Ψ suffers** … *on $\mathcal{M}_1$ with the proper…*

… *q.*

(ii) *(There exis…* … *ithm Γ taking inputs in A(…* … *that satisfies*

$$\|\Phi_{Ax}(Ax) - x\| \leq \delta, \quad \forall x \in \mathcal{M}_1.$$

Suppose that the training set T ⊂ C N is a finite set and Ψ is a reconstruction map that achieves small error on M1. Then there are infinitely-many model classes M1 with T ⊂M1 such that Ψ hallucinates on M1 with high probability (regardless of the distribution on M1)

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

In-distribution hallucinations

Theorem 2

Let $A \in \mathbb{C}^{m\times N}$ with $1 \le \mathrm{rank}(A) < N$, $\mathcal{T} \subset \mathbb{C}^N$ be a non-empty and finite set, $\delta > 0$, $\Psi\colon \mathbb{C}^m \to \mathbb{C}^N$ be a neural network with Lipschitz constant $L > 0$ and $x_{\mathrm{Det}} \in \mathbb{C}^N$ with $\|Ax_{\mathrm{Det}}\| \le \delta/(4L)$. Suppose that Ψ satisfies

$$\max_{x\in\mathcal{T}} \|\Psi(Ax) - x\| \le \delta.$$

Then, for any $\epsilon \ge \delta/(2L)$ there is an uncountable family $\mathcal{C}$ of finite or countably infinite sets $\mathcal{M}_1 \subset \mathbb{C}^N$ with $\mathcal{T} \subset \mathcal{M}_1$ and $A\mathcal{M}_1 \subset \mathcal{B}_{\|\cdot\|}(A\mathcal{T}, \epsilon)$, such that for each $\mathcal{M}_1 \in \mathcal{C}$ the following hold simultaneously.

(i) ***(Ψ suffers from in-distribution hallucinations)*** [illegible] *probability distribution $\mathcal{D}$ on $\mathcal{M}_1$ with the property that $\mathbb{P}_{X\sim\mathcal{D}}(X \in \mathcal{T}) \le q$, it holds that*

$$\mathbb{P}_{X\sim\mathcal{D}}\left(\exists \lambda \in \mathbb{C}, |\lambda| = 1 \text{ such that } \|\Psi(AX) - (X + \lambda x_{\mathrm{Det}})\| \le 2\delta\right) \ge 1 - q.$$

(ii) *(There exists an algorithm that yields non-hallucinating NNs). There exists an algorithm Γ taking inputs in $A(\mathcal{M}_1)$, such that for each $y \in A(\mathcal{M}_1)$, $\Gamma(y) = \Phi_y$ is a NN $\Phi_y : \mathbb{C}^m \to \mathbb{C}^N$ that satisfies*

$$\|\Phi_{Ax}(Ax) - x\| \le \delta, \quad \forall x \in \mathcal{M}_1.$$

However, there exists an algorithm for computing NNs that achieve small errors on M1
and therefore do not hallucinate on M1

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

In-distribution hallucinations

Theorem 2

Let $A \in \mathbb{C}^{m \times N}$ with $1 \leq \operatorname{rank}(A) < N$, $\mathcal{T} \subset \mathbb{C}^N$ be a non-empty and finite set, $\delta > 0$, $\Psi: \mathbb{C}^m \to \mathbb{C}^N$ be a neural network with Lipschitz constant $L > 0$ and $x_{\mathrm{Det}} \in \mathbb{C}^N$ with $\|Ax_{\mathrm{Det}}\| \leq \delta/(4L)$. Suppose that Ψ satisfies

$$\max_{x \in \mathcal{T}} \|\Psi(Ax) - x\| \leq \delta.$$

Then, for any $\epsilon \geq \delta/(2L)$ there is an uncountable family $\mathcal{C}$ of finite or countably infinite sets $\mathcal{M}_1 \subset \mathbb{C}^N$ with $\mathcal{T} \subset \mathcal{M}_1$ and $A\mathcal{M}_1 \subset \mathcal{B}_{\|\cdot\|}(A\mathcal{T}, \epsilon)$, such that for each $\mathcal{M}_1 \in \mathcal{C}$ the following hold simultaneously.

(i) ***(Ψ suffers from in-distribution hallucinations)****. For any probability distribution $\mathcal{D}$ on $\mathcal{M}_1$ with the property that $\mathbb{P}_{X \sim \mathcal{D}}(X \in \mathcal{T}) \leq q$, it holds that*

$$\mathbb{P}_{X \sim \mathcal{D}}\left(\exists \lambda \in \mathbb{C}, |\lambda| = 1 \text{ such that } \|\Psi(AX) - (X + \lambda x_{\mathrm{Det}})\| \leq 2\delta\right) \geq 1 - q.$$

(ii) *(There exists an algorithm that yields non-hallucinating NNs). There exists an algorithm Γ taking inputs in $A(\mathcal{M}_1)$, such that for each $y \in A(\mathcal{M}_1)$, $\Gamma(y) = \Phi_y$ is a NN $\Phi_y: \mathbb{C}^m \to \mathbb{C}^N$ that satisfies*

$$\|\Phi_{Ax}(Ax) - x\| \leq \delta, \quad \forall x \in \mathcal{M}_1.$$

In-distribution hallucinations, yet non-hallucinating algorithms exist

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

In distribution hallucinations

(Inverse) Problem: Image Reconstruction

Given measurements $y = Ax + e$, recover x.

Hallucinations arise necessarily as a result of overperformance of a reconstruction map that has no knowledge of the model class M1

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

Hallucinations due to detail transfer

Theorem 1

Let $A \in \mathbb{C}^{m\times N}$, $\delta > 0$ *and* $x, x_{\mathrm{Det}} \in \mathbb{C}^N$ *with* $\|Ax_{\mathrm{Det}}\| \leq \delta$.

(i) *(Ψ hallucinates by transferring details). Let* $\Psi \colon \mathbb{C}^m \to \mathbb{C}^N$ *be Lipschitz continuous with constant at most* $L > 0$ *and suppose that*

$$\|\Psi(A(x + x_{\mathrm{Det}})) - (x + x_{\mathrm{Det}})\| \leq \delta.$$

Then for every $e \in \mathbb{C}^m$, *with* $\|e\| \leq \delta$ *there is a* $z \in \mathbb{C}^N$ *with* $\|z\| \leq (1 + 2L)\delta$, *such that*

$$\Psi(Ax + e) = x + x_{\mathrm{Det}} + z.$$

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

Hallucinations due to detail transfer

Theorem 1
Let $A \in \mathbb{C}^{m\times N}$, $\delta > 0$ and $x, x_{\mathrm{Det}} \in \mathbb{C}^N$ with $\|Ax_{\mathrm{Det}}\| \leq \delta$.

(i) *(Ψ hallucinates by transferring details). Let $\Psi : \mathbb{C}^m \to \mathbb{C}^N$ be Lipschitz continuous with constant at most $L > 0$ and suppose that*

$$\|\Psi(A(x + x_{\mathrm{Det}})) - (x + x_{\mathrm{Det}})\| \leq \delta$$

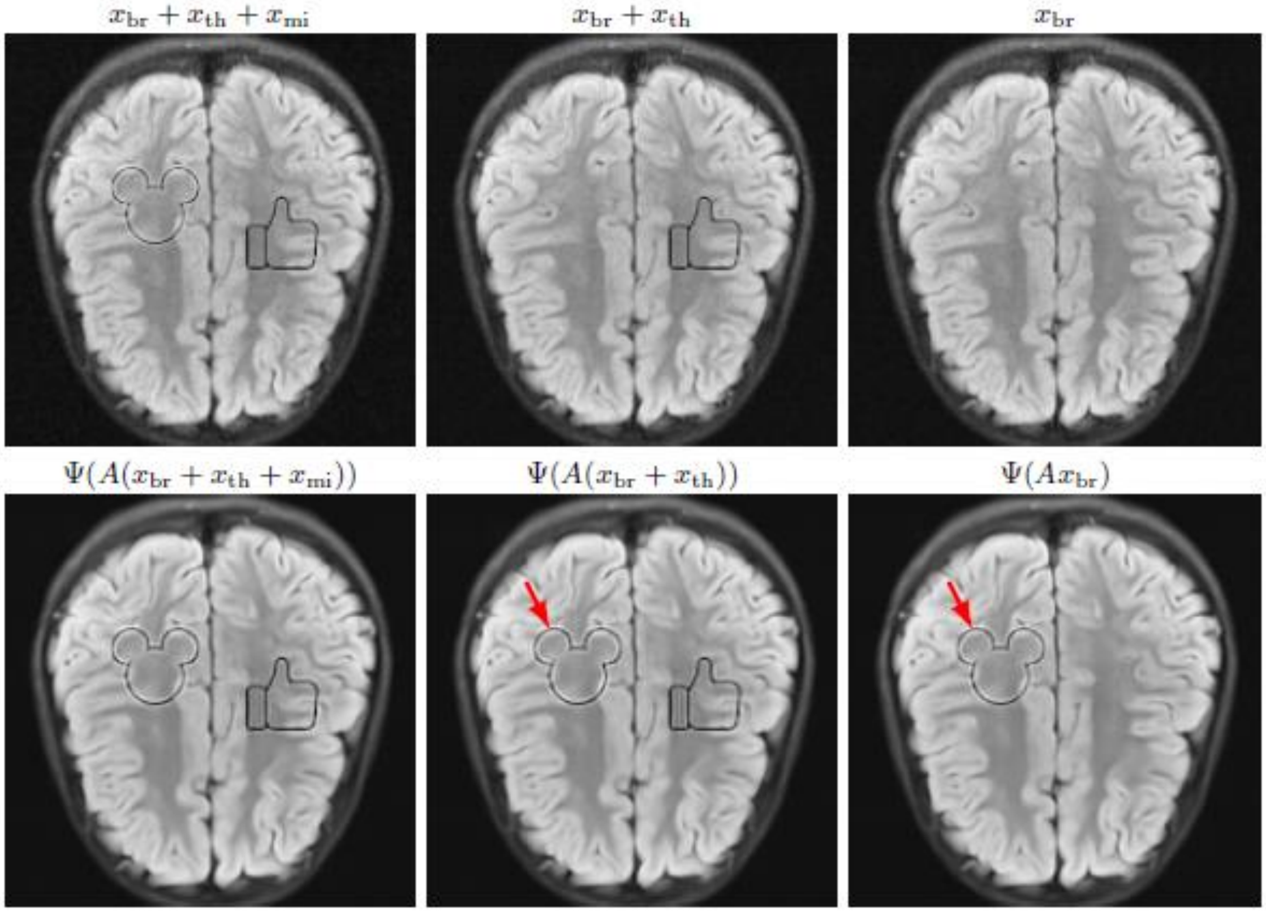
Then for every $e \in \mathbb{C}^m$ with $\|e\| \leq \delta$ there is a $z \in \mathbb{C}^N$ with $\|z\| \leq$ $(1 + 2L)\delta$ such that

$$\Psi(Ax + e) = x + x_{\mathrm{Det}} + z$$

Any map Ψ that recovers the detail image *x+xDet* will hallucinate by incorrectly transferring this detail when reconstructing the detail-free image x, i.e., Ψ(Ax + e) ≈ x + xDet. Thus, a hallucination occurs

See: *N. M. Gottschling, V. Antun, B. Adcock, A. C. Hansen. "The troublesome kernel: – On hallucinations, no free lunches and the accuracy-stability trade-off in inverse problems".*

Hallucinations due to detail transfer



If the
map Ψ performs too well on a certain image x_1 with detail, then it will hallucinate, by incorrectly transferring this detail to another image x_2.

Know thy modelling

Understand (most parts of)

Deep Learning

Understand (most parts of)

Hallucinations

Know thy modelling

Understand (most parts of)

Deep Learning

Understand (most parts of)

Hallucinations

Understand (how to

Quantify) Uncertainty

Conformal Prediction (CP)

CP is a machine learning framework to produce statistically valid regions

- Computes scores on previously trained data
- and using those to create prediction sets on a new test data

Conformal Prediction (CP)

- Provides prediction regions (sets/intervals) that are **guaranteed** to satisfy a required level of confidence
- Prediction regions are **well-calibrated**

Conformal Prediction (CP)

- Provides prediction regions (sets/intervals) that are **guaranteed** to satisfy a required level of confidence
- Prediction regions are well-calibrated
- (only) assumption

data is exchangeable. A set of *N* variables is exchangeable if all the
N! possible orderings of its elements are equally likely

Exchangeable samples should be drawn from the same

distribution but need not be independent (unlike i.i.d.)

Measuring Nonconformity

- For every possible label $Y_j \in \{Y_1, \ldots, Y_c\}$ calculate the *non-conformity scores*

$$\alpha_i^{Y_j} = A(\{z_1, \ldots, z_l, z_{l+1}^{Y_j}\}, z_i), \qquad i = 1, \ldots, l+1$$

  where $z_{l+1}^{Y_j} = (x_{l+1}, Y_j)$.

- Example: Simple regression non-conformity measure:

$$\alpha_i = | y_i - \hat{y}_i |,$$

  where $\hat{y}_i$ is the prediction of the underlying regression technique for x_i.

- Various options

$\alpha = \max o^j - o^u,$

.

$$\alpha_i = \frac{\sum_{j=1}^{k} s_j^i}{\sum_{j=1}^{k} o_j^i}$$

- E.g. *k*-Nearest Neighbours:

Many more

- Multi-label Learning
- Semi-supervised Learning
- Feature selection
- Anomaly detection
- Testing exchangeability / Change Detection in streams
- Active Learning

Conformal Prediction (CP)

- New ways of adapting case for non- exchangeability assumption, distribution shifts

Inverse problems (now/new) attempts

- **Intervals for each pixel by quantile regression** *(Angelopoulos et.al)*
- **Intervals for principal components**

Principal Uncertainty Quantification with Spatial Correlation for Image Restoration Problems

Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, Michael Elad

Conformal Prediction (CP)

- New ways of adapting case for non- exchangeability assumption, distribution shifts

Inverse problems (now/new) attempts

- **Intervals for each pixel by quantile regression** *(Angelopoulos et.al)*
- **Intervals for principal components**

Astrophysics Data

Principal Uncertainty Quantification with Spatial Correlation for Image Restoration Problems

Omer Belhasin, Yaniv Romano, Daniel Freedman, Ehud Rivlin, Michael Elad

Know thy modelling

Understand (most parts of)

Deep Learning

Understand (most parts of)

Hallucinations

Understand (how to

Quantify) Uncertainty

Know thy modelling

Understand (most parts of)

Deep Learning

Understand (most parts of)

Hallucinations

Understand Uncertainty

Hallucinations

Inverse problems