**Predictive Hypothesis Identification**

**Summary**
- If prediction is the goal, but full Bayes is not feasible, one should identify (estimate/test/select) the Hypothesis (parameter/model/interval) that Predicts best (PHI).
- What best is can depend on benchmark (Loss, Loa), distance function (d), how long we use the model (m), compared to how much data we have at hand (n).
- The new principle (PHI) possesses many desirable properties.
- In particular, PHI can properly deal with nested hypotheses, and nicely justifies, reconciles, and blends MAP and ML for m > n, MDL for m ≈ n, and SMF for n > m.

**Background & Idea & Principle**

The Problem - Information Summarization
- Given: Data D ∈ {x1,…,xn} ∈ X^n (any X) sampled from distribution p(D|θ) with unknown θ ∈ Ω.
- Likelihood function p(D|θ) posterior p(θ|D) ≡ p(D,θ|D) contain all statistical information about the sample.
- Information summary or simplification of p(D|θ) is needed: (comprehensibility, communication, storage, computational efficiency, mathematical tractability, etc.).
- Regimes - parameter estimation, - hypothesis testing, - model (complexity) selection.

Ways to Summarize the Posterior
- A single point à = {θ} (ML or MAP or mean or stochastic or...).
- A convex set à ⊆ Ω (e.g. confidence or credible interval).
- A finite set of points θ1,…,θm (mixture models).
- A sample of points (particle filtering).
- The mean and covariance matrix (Gaussian approximation).
- More general density estimation, 
- In a few other ways.

I concentrate on set estimation, which includes (multiple) point estimation and hypothesis testing as special cases.

Call it: Hypothesis Identification.

**Desirable Properties**

Hypothesis identification principle
- leads to good predictions (that’s what models are ultimately for),
- be broadly applicable,
- be analytically and computationally tractable,
- be defined and works also for non-i.i.d. and non-stationary data,
- be reparametrization and representation invariant,
- works for simple and complex hypotheses,
- works for classes containing nested and overlapping hypotheses,
- works in the estimation, testing, and model selection regime,
- reduces in small cases (approximately) to existing other methods.

Here we concentrate on the first item, and will show that the resulting principle nicely satisfies many of the other items.

The Main Idea

Machine learning primarily cares about predictive performance.

We address the problem head on.
- Goal: Predict n future obs. x ≡ {x1,n+1,…,xn+m} ∈ X^n+m well.
- If θ is true parameter, then p(θ|x) is obviously the best prediction.
- If θ unknown, then the Bayesian predictive distribution p(D|θ) = ∫ p(D|θ)p(θ)dθ (ML) is best.
- Approx. full Bayes by predicting with hypothesis H ≡ {θ ∈ Ω, i.e.
- Use (composite) likelihood p(x|θ) = 1/θ p(x|θ)p(θ)dθ for prediction.
- The closer p(θ|x) to p(θ|x) or p(θ|x) the better it’s prediction.
- Measure closeness with some distance function d(·,·)
- Since x and θi are unknown, we must sum or average over them.

**Predictive Hypothesis Identification (PHI)**

Definition 1 (Predictive Loss) The predictive Loss (θ given D) based on distance d be its future observations is

$$L_{P}(\theta, D) = \frac{1}{n} \sum_{i=1}^{n} d(p(x_{i}|\theta), p(x_{i}|D))$$

Definition 2 (PHI) The best (Bayès) predictive hypothesis in hypoth-

Background & Idea & Principle

The Problem - Information Summarization
- Given: Data D ∈ {x1,…,xn} ∈ X^n (any X) sampled from distribution p(D|θ) with unknown θ ∈ Ω.
- Likelihood function p(D|θ) posterior p(θ|D) ≡ p(D,θ|D) contain all statistical information about the sample.
- Information summary or simplification of p(D|θ) is needed: (comprehensibility, communication, storage, computational efficiency, mathematical tractability, etc.).
- Regimes - parameter estimation, - hypothesis testing, - model (complexity) selection.

Ways to Summarize the Posterior
- A single point à = {θ} (ML or MAP or mean or stochastic or...).
- A convex set à ⊆ Ω (e.g. confidence or credible interval).
- A finite set of points θ1,…,θm (mixture models).
- A sample of points (particle filtering).
- The mean and covariance matrix (Gaussian approximation).
- More general density estimation,
- In a few other ways.

I concentrate on set estimation, which includes (multiple) point estimation and hypothesis testing as special cases.

Call it: Hypothesis Identification.

**Desirable Properties**

Hypothesis identification principle
- leads to good predictions (that’s what models are ultimately for),
- be broadly applicable,
- be analytically and computationally tractable,
- be defined and works also for non-i.i.d. and non-stationary data,
- be reparametrization and representation invariant,
- works for simple and complex hypotheses,
- works for classes containing nested and overlapping hypotheses,
- works in the estimation, testing, and model selection regime,
- reduces in small cases (approximately) to existing other methods.

Here we concentrate on the first item, and will show that the resulting principle nicely satisfies many of the other items.

The Main Idea

Machine learning primarily cares about predictive performance.

We address the problem head on.
- Goal: Predict n future obs. x ≡ {x1,n+1,…,xn+m} ∈ X^n+m well.
- If θ is true parameter, then p(θ|x) is obviously the best prediction.
- If θ unknown, then the Bayesian predictive distribution p(D|θ) = ∫ p(D|θ)p(θ)dθ (ML) is best.
- Approx. full Bayes by predicting with hypothesis H ≡ {θ ∈ Ω, i.e.
- Use (composite) likelihood p(x|θ) = 1/θ p(x|θ)p(θ)dθ for prediction.
- The closer p(θ|x) to p(θ|x) or p(θ|x) the better it’s prediction.
- Measure closeness with some distance function d(·,·)
- Since x and θi are unknown, we must sum or average over them.

**Predictive Hypothesis Identification (PHI)**

Definition 1 (Predictive Loss) The predictive Loss (θ given D) based on distance d be its future observations is

$$L_{P}(\theta, D) = \frac{1}{n} \sum_{i=1}^{n} d(p(x_{i}|\theta), p(x_{i}|D))$$

Definition 2 (PHI) The best (Bayès) predictive hypothesis in hypoth-

Background & Idea & Principle

The Problem - Information Summarization
- Given: Data D ∈ {x1,…,xn} ∈ X^n (any X) sampled from distribution p(D|θ) with unknown θ ∈ Ω.
- Likelihood function p(D|θ) posterior p(θ|D) ≡ p(D,θ|D) contain all statistical information about the sample.
- Information summary or simplification of p(D|θ) is needed: (comprehensibility, communication, storage, computational efficiency, mathematical tractability, etc.).
- Regimes - parameter estimation, - hypothesis testing, - model (complexity) selection.

Ways to Summarize the Posterior
- A single point à = {θ} (ML or MAP or mean or stochastic or...).
- A convex set à ⊆ Ω (e.g. confidence or credible interval).
- A finite set of points θ1,…,θm (mixture models).
- A sample of points (particle filtering).
- The mean and covariance matrix (Gaussian approximation).
- More general density estimation,
- In a few other ways.

I concentrate on set estimation, which includes (multiple) point estimation and hypothesis testing as special cases.

Call it: Hypothesis Identification.

**Desirable Properties**

Hypothesis identification principle
- leads to good predictions (that’s what models are ultimately for),
- be broadly applicable,
- be analytically and computationally tractable,
- be defined and works also for non-i.i.d. and non-stationary data,
- be reparametrization and representation invariant,
- works for simple and complex hypotheses,
- works for classes containing nested and overlapping hypotheses,
- works in the estimation, testing, and model selection regime,
- reduces in small cases (approximately) to existing other methods.

Here we concentrate on the first item, and will show that the resulting principle nicely satisfies many of the other items.

The Main Idea

Machine learning primarily cares about predictive performance.

We address the problem head on.
- Goal: Predict n future obs. x ≡ {x1,n+1,…,xn+m} ∈ X^n+m well.
- If θ is true parameter, then p(θ|x) is obviously the best prediction.
- If θ unknown, then the Bayesian predictive distribution p(D|θ) = ∫ p(D|θ)p(θ)dθ (ML) is best.
- Approx. full Bayes by predicting with hypothesis H ≡ {θ ∈ Ω, i.e.
- Use (composite) likelihood p(x|θ) = 1/θ p(x|θ)p(θ)dθ for prediction.
- The closer p(θ|x) to p(θ|x) or p(θ|x) the better it’s prediction.
- Measure closeness with some distance function d(·,·)
- Since x and θi are unknown, we must sum or average over them.