Machine Learning and Interpretability in Biology and Medicine

Neo Christopher Chung
Assistant Professor in Computer Science
Institute of Informatics, University of Warsaw

https://cbml.science

INFORM Seminar 13/01/2023
Advanced Computational Techniques for Oncology: towards Explainable Artificial Intelligence

Back in the days

Prognosis of a heart attack

Cardiovascular diseases the leading cause of death

Fatality rates from heart attacks **were** extremely high.

In 1980s, when a heart attack patient is admitted (University of California, San Diego Medical Center), they would measure **19 variables** within 24 hours:

Blood pressure, age, and 17 clinical variables known to be **highly informative** of the patient's condition.

Additionally, temperature, humidity, upper atmospheric conditions, levels of airborne pollutants, and other meteorological variables.

But they were not being used in clinical practices. How to improve the prognosis?

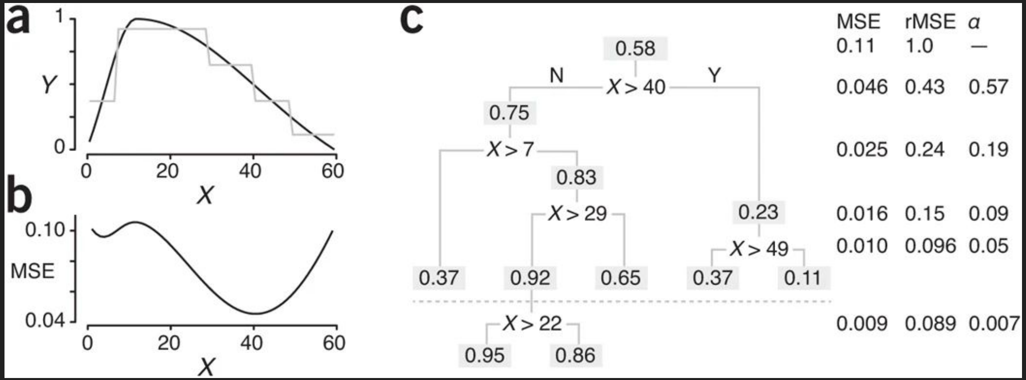
CART™

Breiman, Friedman, Olshen, Stone developed Classification and Regression Trees (CART).

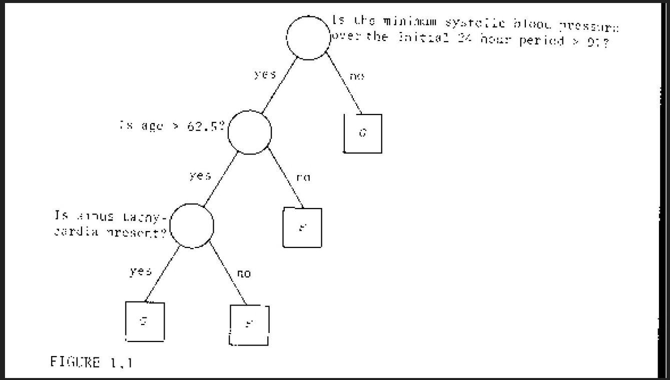
Select a clinical variable, and split (e.g., binary).
No stopping rule, repeat until no more split is possible.
Minimizing a cost function, a greedy algorithm.

Decision trees allowed clinicians to trust the model and apply even without a computer.

Nowadays, more than 90% survival.



Krzywinski and Altman (2017)



Breiman et al (1984)

What is interpretability?

“We define interpretable machine learning as the extraction of relevant **knowledge** from a machine-learning model concerning relationships either contained in data or learned by the model.” – Murdoch et al. (2019)

“Interpretability is the degree to which a human can **understand the cause of a decision**” – Miller (2017)

“The higher the interpretability of a machine learning model, the easier it is for someone to comprehend **why certain decisions or predictions have been made**.” – Molnar (2022)

Why it's so difficult to define interpretability

What is an explanation?

Or a sufficient explanation?

What is understandable to humans?

What if an explanation is understandable to a doctor but not a patient?

How simple should an interpretable model be?

Is a simple model always more interpretable?

How do we compare explanations?

→ **No definition and no quantification**

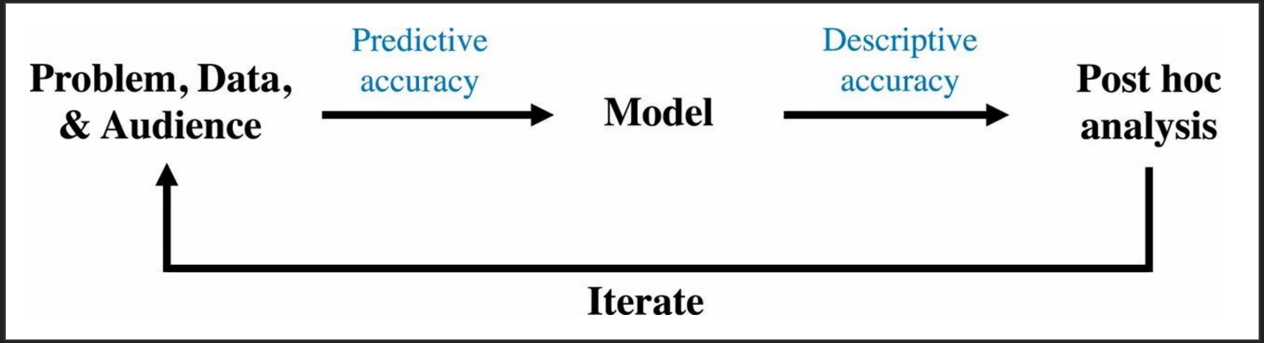
Interpretation in a larger context

Model-based interpretability

Requires modification of existing models
Potentially lower performances
Direct understanding
Simpler models/systems

Post-hoc interpretability

No modification of a model
No change in performance
Potentially ambiguous interpretation



Murdoch et al. (2019)

Trade-off

Predictive accuracy: the performance of the trained ML model

Descriptive accuracy: the accuracy of the post-hoc interpretability

Decrease in Descriptive accuracy

The full model (e.g., coefficients and weights)
Approximation
The top predictors
Examples/prototypes
Linear local approximation
Model compression

Two approaches to interpretability

Interpretability DL can feel enigmatic and ambitious. Where do we start?

Model-centric: explain how the model works in a “simplified” manner while being faithful to the model

Human-centric: explain the model works in a “understandable” manner to humans

→ In the best case scenario, both would converge to the same solution

Models

Statistical and machine learning algorithms (including neural networks):

$$Y = f(X)$$

In a linear model:

$$Y = BX + E$$

In a generalized linear model (e.g., logistic regression),

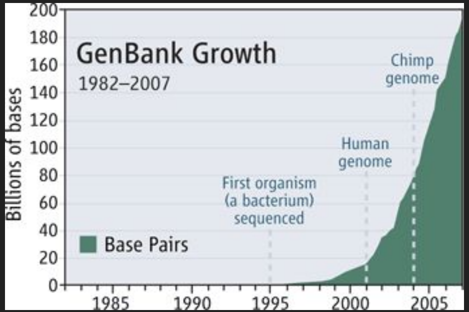
$$g(Y) = BX + E$$

Machine Learning and Interpretability

Increasing data!

Increasing parameter space
e.g., more genes and pixels

Fuller understanding &
Less interpretable

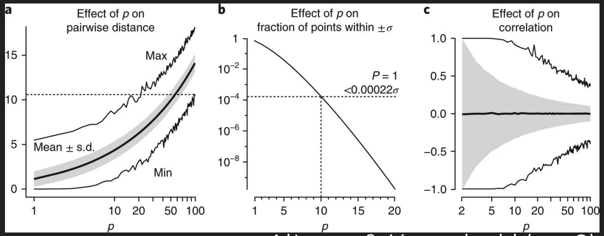


Pennisi 2008

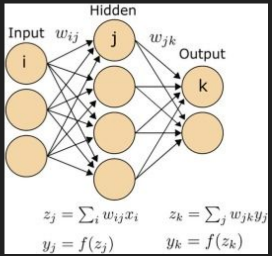
Spatial, contrast, and temporal resolutions of cardiac imaging methods

	Spatial resolution (FWHM), mm	Contrast resolution	Temporal resolution
CT	0.5-0.625	Low to moderate	83-135 ms
MRI	1-2	High	20-50 ms
Catheter angiography	0.16	Moderate	1-10 ms
PET	4-10*	Very high, varies†	5 s to 5 min*
SPECT	4-15*	Very high, varies†	15 min§
Echocardiography	~0.5-2‡	Low to moderate	>200 frames/s (<5 ms)

Lin and Alessio 2009



Altman & Krzywinski (2018)



More complex models, non-linearity
e.g., GAM, DNN

Better performance &
Less interpretable

Machine Learning and Interpretability

Simpler models/algorithms
Easier to interpret

More complex models/algorithms
Harder to interpret

Linear Model

Kernel Methods

Generalized Linear Models

Generalized Additive Models

Decision trees

Rules-based

Bagging, Boosting, Ensemble Models

Perceptron

Sparse DL

Convolutional NN

Machine Learning Models

Statistical and machine learning algorithms (including neural networks):

$$Y = f(X)$$

A linear regression:

$$y = b_0 + b_1 x_1 + \dots + b_n x_n + e$$

y ~ outcomes, disease progression, cancer status, etc

x ~ genes, clinical variables, pixels, features, etc

b ~ coefficients to be **estimated**

Machine Learning Models

Statistical and machine learning algorithms (including neural networks):

$$Y = f(X)$$

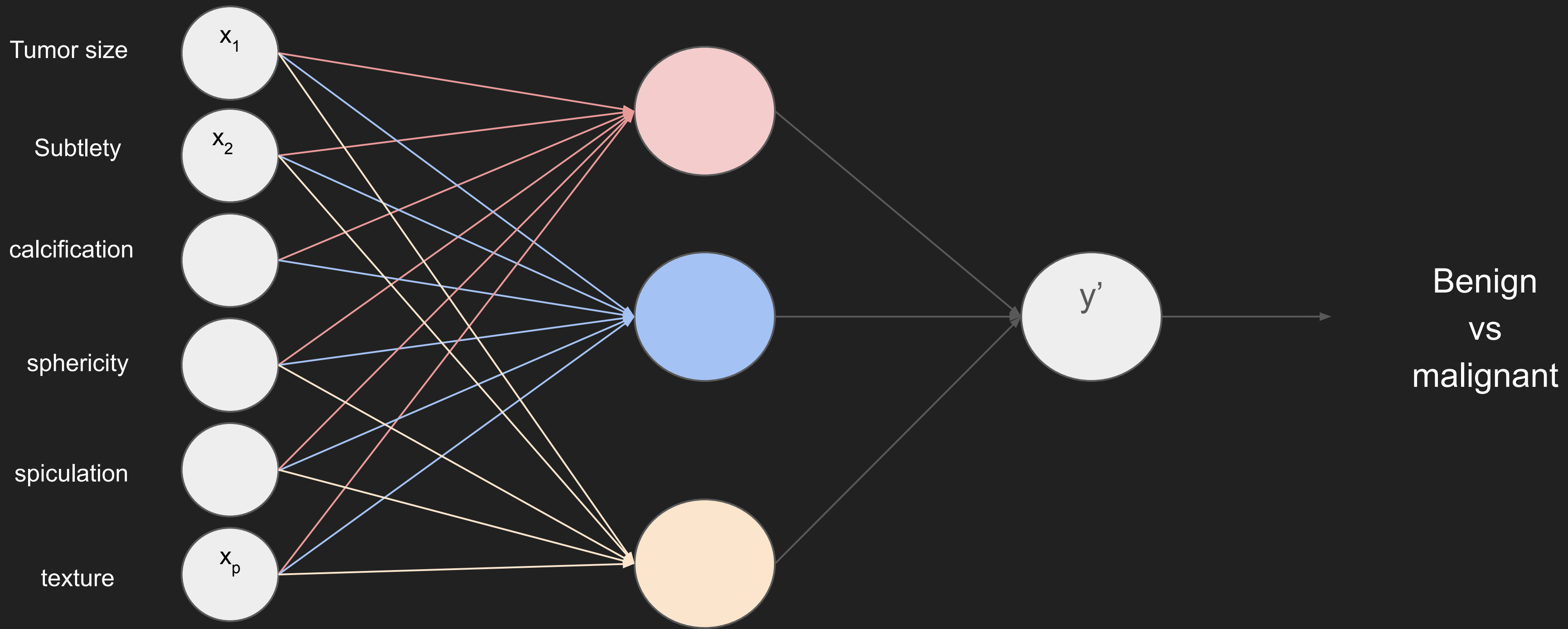
A linear regression:

$$y_1 = b_{1,0} + b_{1,1}x_{1,1} + \dots + b_{1,n}x_{1,n} + e_1$$

$$\dots$$

$$y_m = b_{m,0} + b_{m,1}x_{m,1} + \dots + b_{m,n}x_{m,n} + e_m$$

Neural Networks



Large data means noisy data

$$\mathbf{Y} = \mathbf{BX} + \mathbf{E}$$

Based on observed **X** and **Y**, **B** must be estimated.

In many systems, many variables are expected **not to contribute** to the outcome.
e.g., background pixels in CT/PET images unimportant for tumor/survival/prognosis
e.g., most of genes not related to clinical phenotypes in a genome-wide association study

When data are collected from real world, all of estimated coefficients will be likely non-zero.

Regularization and shrinkage

Lasso adds a L_1 penalty to the least squares:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

B = {B_1, ..., B_p}; t controls the regularization

Ridge (L_2 penalty)

Elastic Net (combining L_1 & L_2 penalties)

Lasso example: prostate cancer (Stamey et al. 1989)

Men who were about to receive radical prostatectomy

Levels of **prostate-specific antigen** and clinical variables:

age, cancer vol, , prostate weight, benign prostatic hyperplasia amount, etc

Fit a linear model,
with a lasso penalty

Shrinkage effect (t) was
selected by cross validation

Three clinical variables are
selected

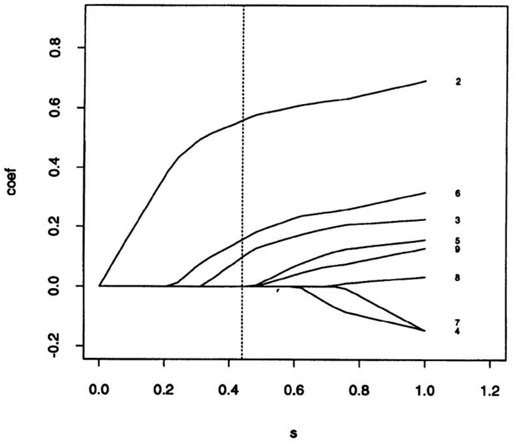


Fig. 5. Lasso shrinkage of coefficients in the prostate cancer example: each curve represents a coefficient (labelled on the right) as a function of the (scaled) lasso parameter $s = t/\Sigma|\beta_j^o|$ (the intercept is not plotted); the broken line represents the model for $\hat{s} = 0.44$, selected by generalized cross-validation

Tibshirani (1996)

Lasso example: bootstrapping

Standard error estimated by bootstrap resampling of residuals

Fit the least squares fit with a fixed lasso penalty from CV

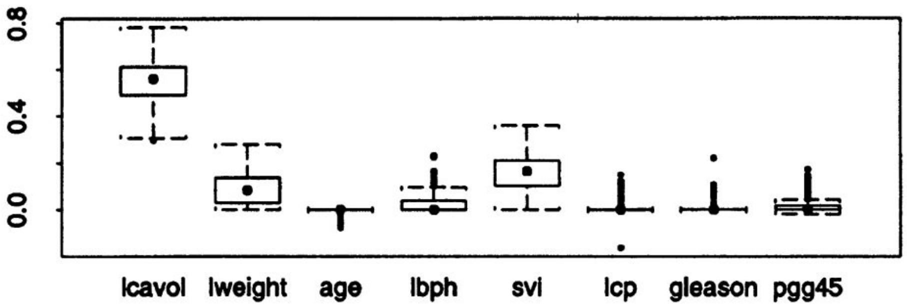


Fig. 6. Box plots of 200 bootstrap values of the lasso coefficient estimates for the eight predictors in the prostate cancer example

Tibshirani (1996)

Shrinkage & variable selection

A large number of predictors → variable/feature selection.

Nowadays, a typical clinical studies would collect > hundreds of variables.

How to intelligently regularize is one of the central goals.

In the feature space of DNN:

Srivastava et al. 2014 Dropout: A Simple Way to Prevent Neural Networks from Overfitting

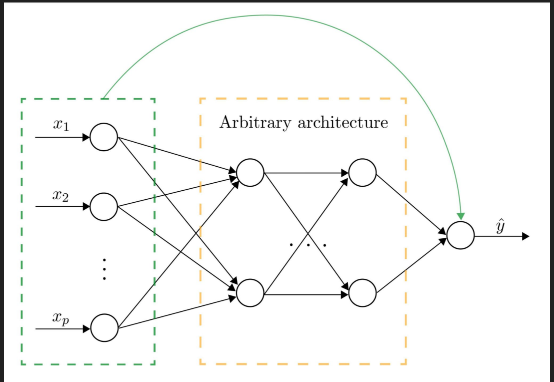
Lemhadri et al 2021 A neural network with feature sparsity

Interpretability methods for neural networks also **implicitly or explicitly**:

Ross et al. 2017 The Neural LASSO: Local Linear Sparsity for Interpretable Explanations

LassoNet for sparse features

The skip-layer (residual) connection allows sparsity in features being used in a classifier. Both linear and non-linear parts are optimized jointly, e.g., L_1 penalty
LassoNet selects the most important features automatically; lead to lower classification errors (sometimes).



$$\underset{\theta, W}{\text{minimize}} \; L(\theta, W) + \lambda \|\theta\|_1$$
$$\text{subject to} \; \left\|W_j^{(1)}\right\|_\infty \leq M|\theta_j|, \; j = 1, \ldots, d.$$

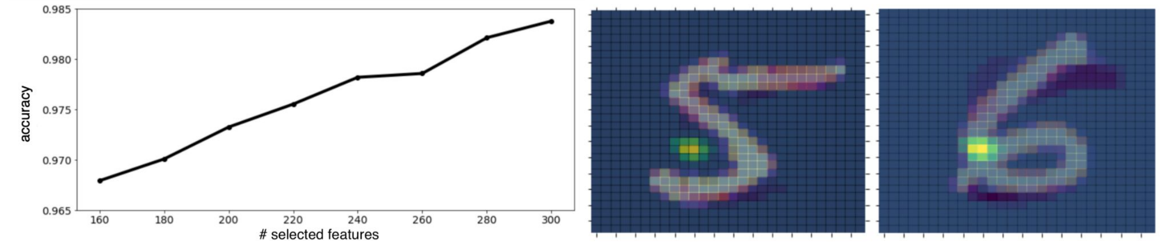
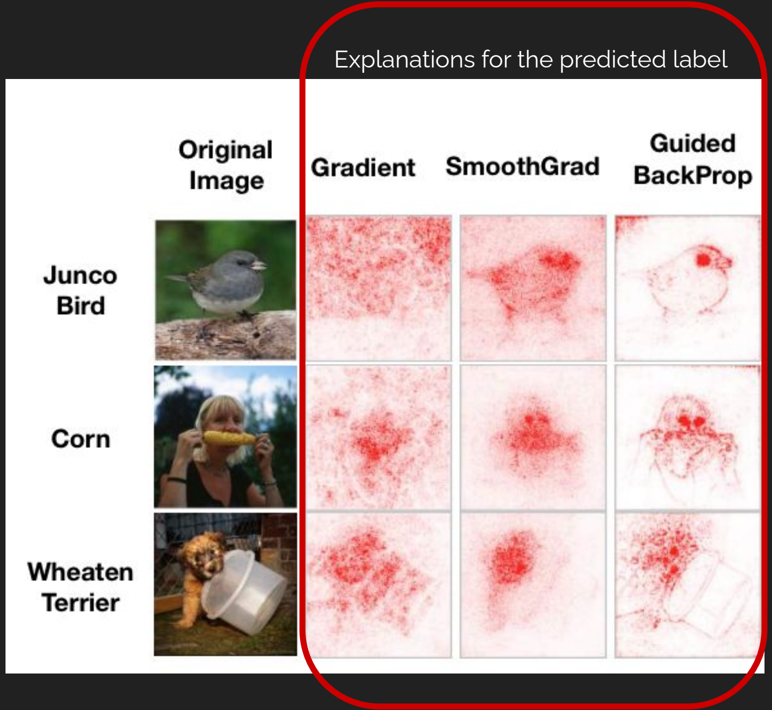


Figure 2. **Demonstrating LassoNet on the MNIST dataset.** Here, we show the results of using LassoNet to simultaneously select informative pixels and classify digits 5 and 6 from the MNIST dataset. **Leftmost graph:** The classification accuracy by number of selected features **Right 2 images:** Individual pixel importance for the model with 220 active features. Here, pixel importance refers to the mean increase in digit 6's predicted probability when setting it to maximum intensity. Lighter colors indicate higher importance, with yellow highest and dark blue lowest. Superimposed are two sample digits. This confirms that the bottom left pixels of digit 6 are the most important.

Lemhadri et al 2021

Implicit sparsity in saliency maps (gradients)



Adebayo et al 2018

But!

Variable selection must reflect the underlying system

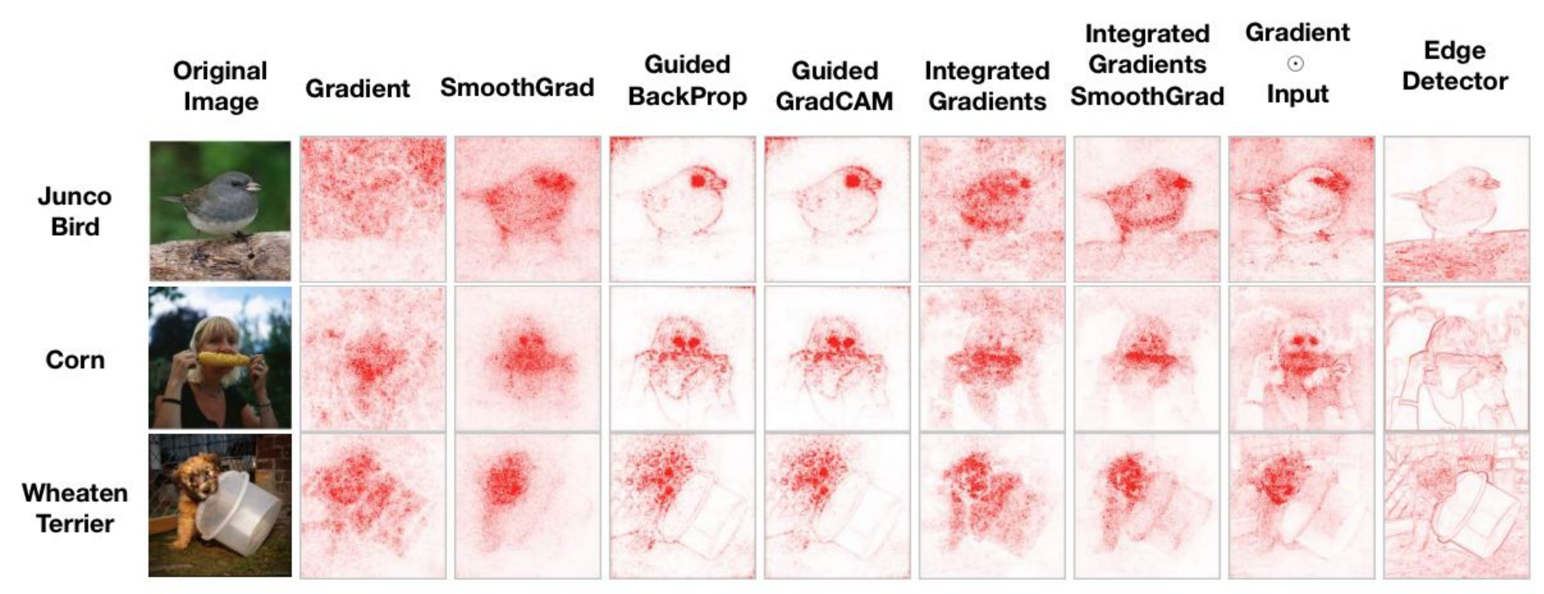
Ideally, sparisty leads to a robust and generalizable model

Quantitative ways (eg, CV) to select shrinkage effects

The trade-off between interpretability and accuracy

Human perception is easily fooled

How do we evaluate?



Adebayo et al 2018

Why seek sparsity?

More interpretable

Better visualization

Build testable hypotheses

Computational efficient

Easier to implement in practice

Why seek sparsity?

More interpretable

Better visualization

Build testable hypotheses

Computational efficient

Easier to implement in practice

Better prediction in high dimensional data → Empirical Bayes

Empirical Bayes & Stein’s Paradox

Stein’s Paradox in Statistics by Efron & Morris (1977)

When multiple variables are estimated simultaneously (in an unbiased and optimal manner), there exists more accurate estimators (lower MSE) on average.

ESTIMATION WITH QUADRATIC LOSS

W. JAMES
FRESNO STATE COLLEGE

AND

CHARLES STEIN
STANFORD UNIVERSITY

Empirical Bayes

In a standard Bayesian, **a prior = fixed** before data

In empirical Bayes, a prior distribution **is estimated from the data**

No need to impose or have a strong prior belief

Bridging two sides (frequentist vs Bayesian) of statistical traditions

Appropriate for modeling repeated ML applications in the system

Least squares with a large m

- Estimate each b_{ij} independently via minimizing the sum of the squares of the residuals
- For each variable $i = 1, \ldots, m$, the errors are uncorrelated, a mean of zero, equal variances (a.k.a. optimal and unbiased)

- However, when we are dealing with a large data -- m variables, measured on a set of n observations -- consider <u>a bias variance tradeoff</u>
- Get “better” estimates by reducing variance and increasing bias
- James–Stein estimator → Empirical Bayes estimator

Shrinkage, a toy example

$$Y = BX + E$$

$$\begin{Bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \\ b_7 \\ b_8 \end{Bmatrix} \xrightarrow{\text{Least Squares estimation}} \begin{Bmatrix} 0.5 \\ 1.1 \\ 0.1 \\ -2.1 \\ 0.2 \\ -1.2 \\ -0.4 \\ -0.2 \end{Bmatrix} \xrightarrow{\text{Shrinking toward the mean by “some method”}} \begin{Bmatrix} 0.4 \\ 0.9 \\ 0.05 \\ -1.9 \\ 0.1 \\ -1.1 \\ -0.3 \\ -.15 \end{Bmatrix}$$

Mean = -.25
Var = 1.01

Mean = -.25
Var = 0.77

A baseball player’s batting average

In sports analytics, we often want to predict a player’s statistics in the future.
E.g., a batting average = # of a baseball player's hits divided by # of at-bats.

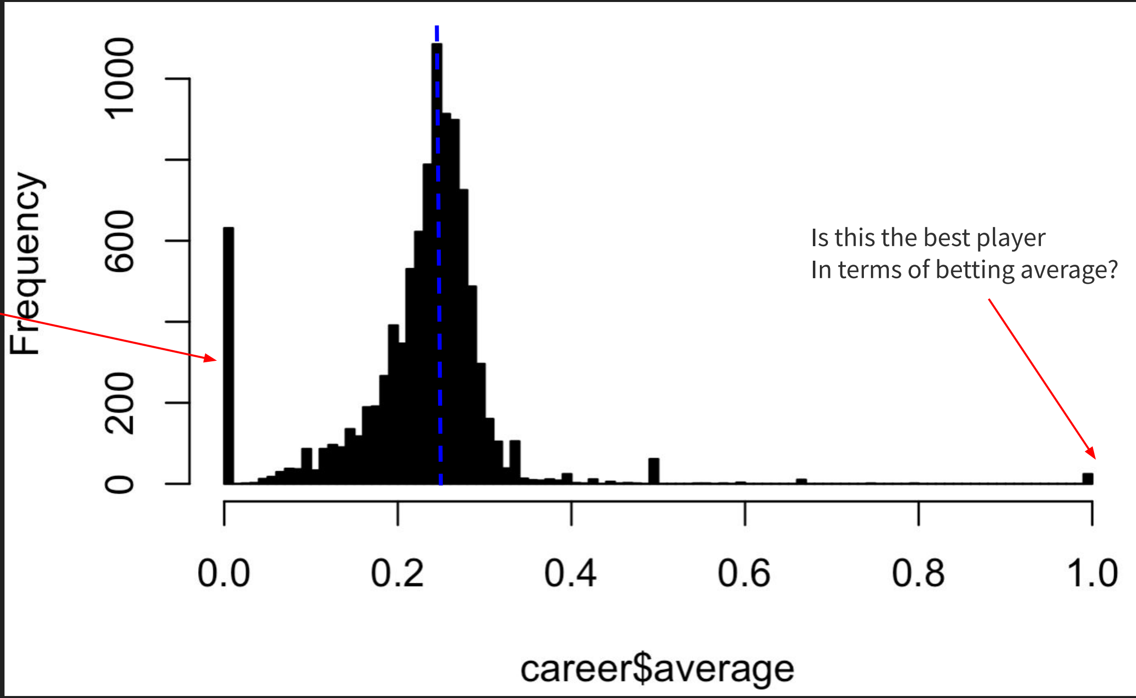
Given the last year’s data on batting averages, you are tasked with predicting players’ batting average next year.

Unbiased LM: the individual player’s batting average from the last year as the predicted average for the next year.

Distribution of batting averages

Betting Averages of 9,256 baseball players in the US major leagues

Will these players have an average of 0.0 again next year?

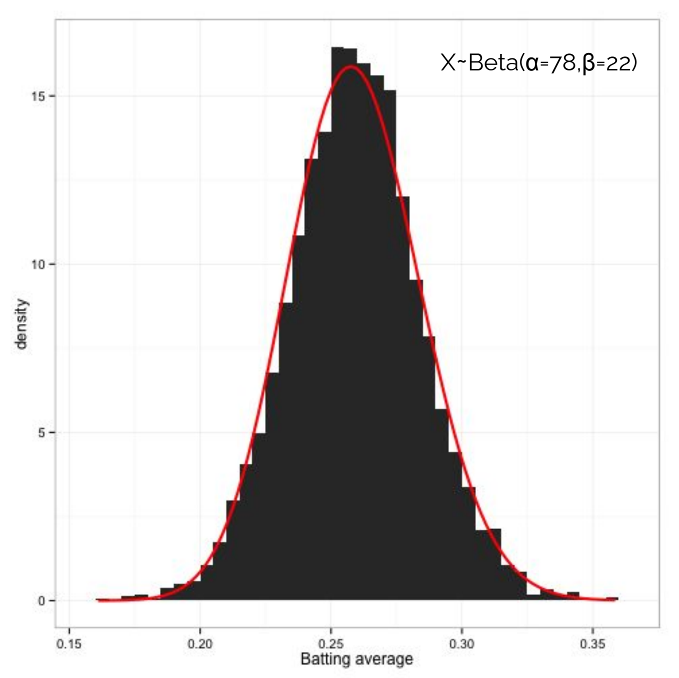


Is this the best player
In terms of betting average?

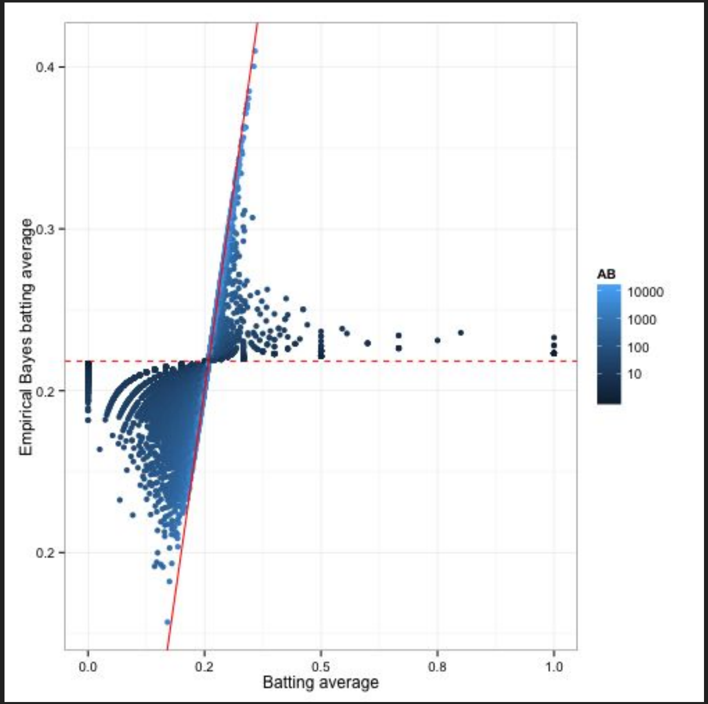
Robinson 2015

Improved prediction of batting averages

Beta distribution fit into the data (after some filtering)



Empirical Bayes Estimates of batting averages



Empirical Bayes: moderated t-statistics

Gene expression of zebrafish with a Swirl mutation in BMP2 gene — Smyth (2004)
BMP2 gene that affects the dorsal/ventral body axis
A microarray experiment gave 8448 probes (~ transcripts)

Goal: identify genes with differential expression in Swirl mutants vs wild-types
We can improve our estimates by borrowing information across all data

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g \sqrt{v_{gj}}}$$

t-test for testing whether means of two groups are equal
s_g = residual sample variance

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}.$$

Moderated t-test
with reduced s_g

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}.$$

The "effective" DOF increases
as extra information borrowed from all data

Swirl gene expression study

Degree of freedom: $d_g = 3$
Estimated prior DoF $d_0 = 4.17$

Sample variance $s^2_g = 0.109$
Estimated prior variance $s^2_0 = 0.0509$

Instead of unbiased t-statistics, we can better identify differentially expressed genes by borrowing information across all genes.

Top 10 genes from the Swirl experiment

ID	Name	M-value	Ord t	Mod t	B
control	BMP2	-2.21	-23.94	-21.1	7.96
control	BMP2	-2.30	-20.20	-20.3	7.78
control	Dlx3	-2.18	-21.03	-20.0	7.71
control	Dlx3	-2.18	-20.09	-19.6	7.62
fb94h06	20-L12	1.27	30.23	14.1	5.78
fb40h07	7-D14	1.35	17.39	13.5	5.54
fc22a09	27-E17	1.27	21.11	13.4	5.48
fb85f09	18-G18	1.28	20.23	13.4	5.48
fc10h09	24-H18	1.20	28.30	13.2	5.40
fb85a01	18-E1	-1.29	-17.39	-13.1	5.32

Smyth (2004)

Concluding remarks

Good performance and understanding are not sufficient for clinical translation

The model and the system guides analytical approaches

Application domains are the most important to consider!

Implicit sparsity-inducing behaviors (in feature space and saliency maps)

Sparse networks, two group models, false discovery rates

Interpretable DL in empirical Bayes frameworks

Cross-disciplinary pollination

References

Breiman, Friedman, Olshen, Stone (1984) Classification and Regression Trees
Krzywinski and Altman (2017) Classification and regression trees. Nature Methods
Murdoch et al. (2019) Definitions, methods, and applications in interpretable machine learning. PNAS
Miller (2017) Explanation in artificial intelligence: Insights from the social sciences. arXiv Preprint arXiv:1706.07269
Molnar (2022) Interpretable Machine Learning. Lulu.com
Lin and Alessio (2009) What are the basic concepts of temporal, contrast, and spatial resolution in cardiac CT? J Cardiovasc Comput Tomogr.
Pennisi (2008). Proposal to 'Wikify' GenBank Meets Stiff Resistance. Science
Donoho (2000) High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. AMS National Meeting
Tibshirani (1996). Regression Shrinkage and Selection via the lasso. JRSS Series B (methodological).
Lemhadri, Ruan, Tibshirani (2021) A neural network with feature sparsity. AISTATS
Altman and Krzywinski (2018) The curse(s) of dimensionality. Nature Methods
Adebayo et al (2018) Sanity Checks for Saliency Maps. NeurIPS
Robinson (2015) Understanding empirical Bayes estimation (using baseball statistics). http://varianceexplained.org/r/empirical_bayes_baseball/
Smyth (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol.