Evaluación de modelos

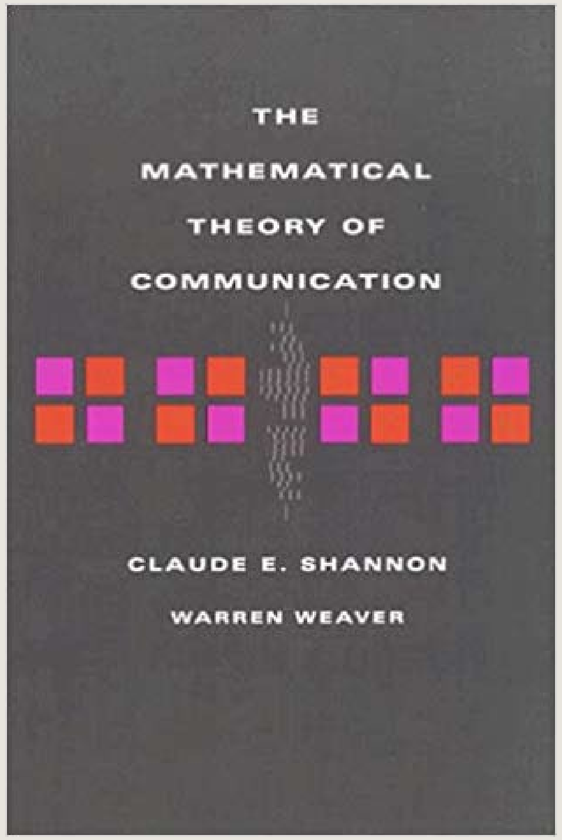
Dr. Héctor Nájera
Dr. Curtis Huffman

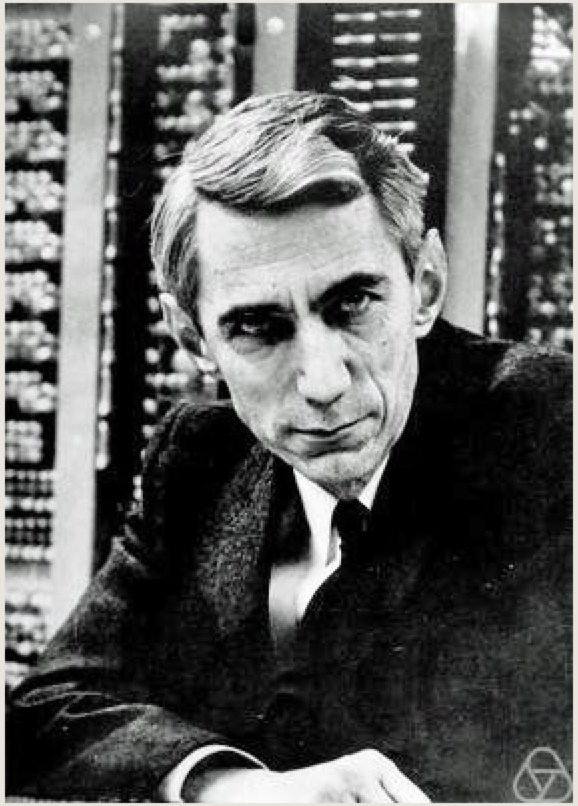
18/04/2024

La sesión anterior

- Exactitud vs. Simplicidad
  - Overfitting: aprender demasiado de la muestra (modelos demasiado flexibles)
  - Underfitting: aprender demasiado poco (modelos no lo suficientemente flexibles)
- Dentro y fuera de la muestra
  - El objetivo nunca es retrodecir la muestra, sino predecir: aprender los aspectos regulares del fenómeno ($p(\tilde{y}_i|y)$; más allá del ECM y el R2)
  - **Simplemente no se puede evaluar el desempeño de los modelos sobre los datos usados para ajustar el modelo (*training* data)**.
  - LOO-CV (computacionalmente intensivo)
- La probabilidad conjunta como (distribución) objetivo: $P(D|\theta)$
- Teoría de la información y divergencia
  - ¿por qué tiene sentido usar información como criterio para evaluar el desempeño de los modelos?

Teoría de la Información





Claude Elwood Shannon
30 de abril de 1916
24 de febrero de 2001

Shannon, C. E. (1948).

Sec. 6 Choice, Uncertainty and Entropy

Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \cdots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say $H(p_1, p_2, \cdots, p_n)$, it is reasonable to require of it the following properties:

1. H should be continuous in the p_i.
2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The

The Bell System Technical Journal

Vol. XXVII | July, 1948 | No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 392-393.

Shannon, C. E. (1948).

- Luego Shannon prueba el teorema

In Appendix II, the following result is established:

Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant.

Shannon buscaba una medida de cuánta "elección" o "incertidumbre" hay en un resultado. Una medida de la información contenida en o asociada a, una distribución de probabilidad.

Definición de entropía

- Let a discrete probability distribution be defined as

$$p_k;\ k = 1, 2, \dots, K$$

$$p_k \ \geq 0\ ,\ \sum_k p_k = 1$$

- The Entropy of the distribution is defined as

$$H = -\sum_k p_k \log\, p_k$$

- Note, by convention $0 \log 0 = 0$.

Medida de información de Shannon

- Interpretaciones
  - La inverosimilitud (o sorpresa) promedio respecto al resultado (todos los posibles) de un experimento
    - Noten que tiene la forma de un promedio: el promedio de la cantidad "$-\log\, p_k$" bajo la distribución de probabilidad "$p_1, \ldots, p_K$".

$$-\sum_k p_k \log\, p_k$$

  - Una medida de incertidumbre
  - Una medida de información

¿De qué va la TI ?

- Piensen que reducir "incertidumbre" (falta de seguridad, de confianza o de certeza sobre algo) es tanto como obtener información
- Estamos interesados en la cantidad de información contenida en un experimento
  - Una manera de medir la cantidad de información con la que se cuenta es por el número de preguntas necesarias para obtener la información requerida.

- Para un experimento con *n* resultados posibles, de probabilidades $p_1, \ldots, p_n$, la cantidad *H* es una medida de cuan “difícil” es descubrir cuál resultado ha tomado lugar.

¿Adivina quién?

Diferentes estrategias

Dumb "Strategy"	Smart "Strategy"
1) Is it Nixon?	1) Is the person a male?
2) Is it Gandhi?	2) Is he alive?
3) Is it me?	3) Is he in politics?
4) Is it Marilyn Monroe?	4) Is he a scientist?
5) Is it you?	5) Is he very well-known?
6) Is it Mozart?	6) Is he Einstein?
7) Is it Niels Bohr?	
8)	

¿con cuál estrategia ganas más información (excluye más posibilidades)?

Diferentes estrategias

Diferentes estrategias

Table 2.2. Diffierent Strategies of Asking Questions

Dumb strategy: Specific questions	Smart strategy: grouping according to some property	Smarter strategy: grouping into two parts nearly half of the range of persons
1. Is it Bush?	1. Does the person have blue eyes?	1. Is the person alive?
2. Is it Gandhi?	2. Is the person living in Paris?	2. Is the person a male?
3. Is it Mozart?	3. Is the person an actor?	3. Does the person live in Europe?
4. Is it Socrates?	4. Does the person work in the field of thermodynamics?	4. Is the person in the sciences?
5. Is it Niels Bohr?	5. Is the person a male?	5. Is the person well known?

Diferentes estrategias

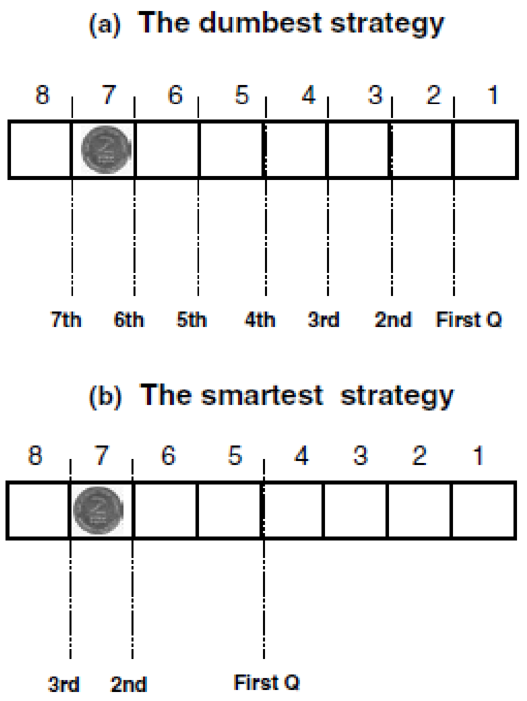


Fig. 2.15 Eight boxes and a coin, the dumbest and the smartest strategies.

Diferentes estrategias

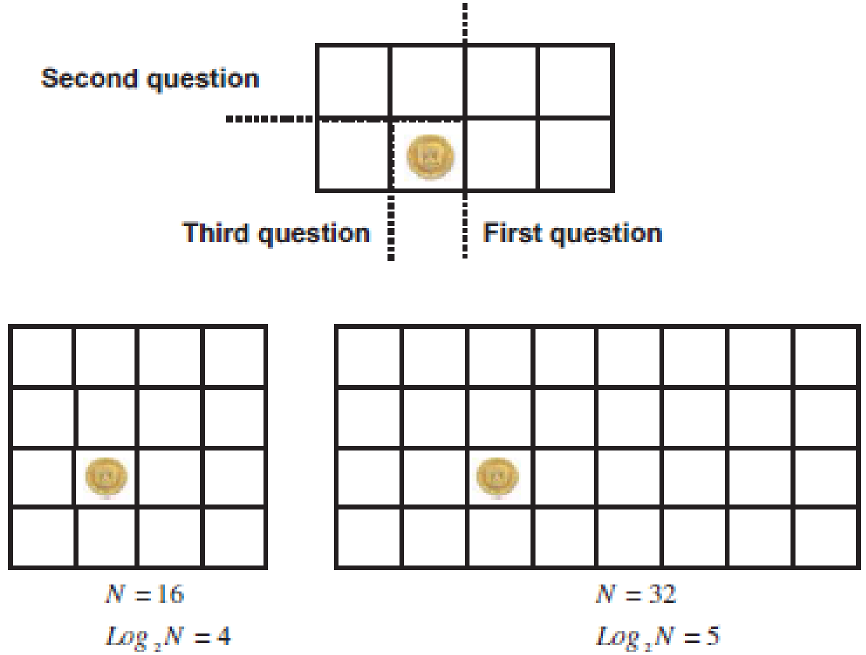


Fig. (2.9)

The Dumbest Strategy	The Smartest Strategy
1) Is the coin in box 1?	1) Is the coin in the right half (of the eight)?
2) Is the coin in box 2?	2) Is the coin in the upper half (of the remaining four)?
3) Is the coin in box 3?	3) Is the coin in the right half (of the remaining two)?
4) Is the coin in box 4?	4) I know the answer!
5)	
⋮ ⋮	

Ben-Naim, A. (2008). *Entropy demystified: The second law reduced to plain common sense*. World Scientific.

Diferentes estrategias

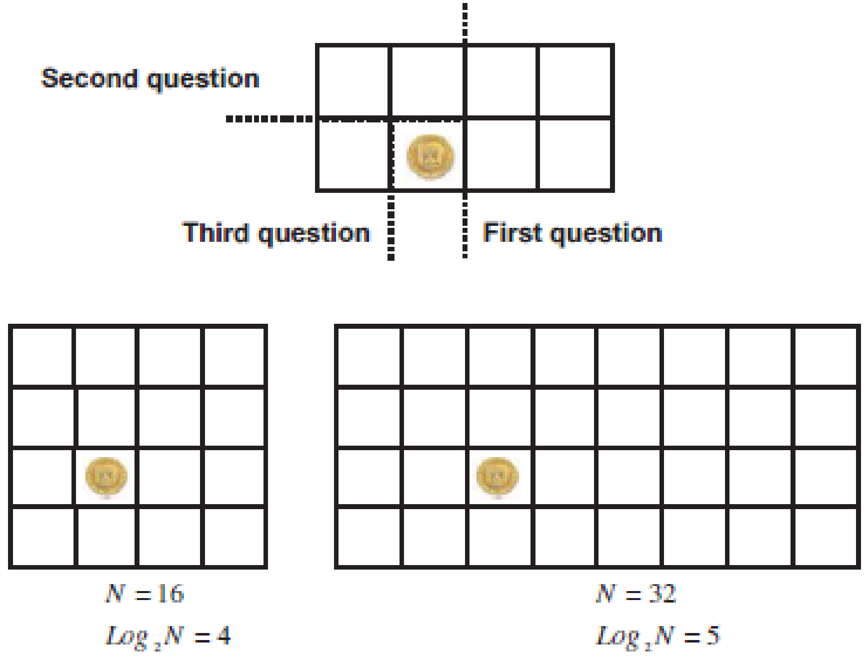


Fig. (2.9)

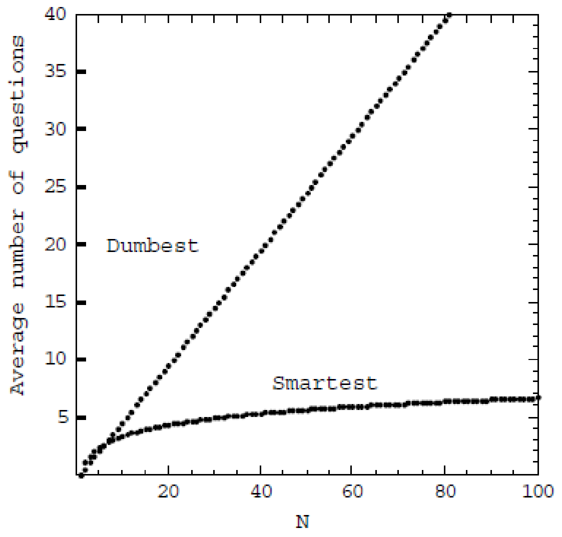


Figure 3.15. The average number of questions as a function N for the two strategies.

Ben-Naim, A. (2008). *Entropy demystified: The second law reduced to plain common sense*. World Scientific.

Diferentes estrategias

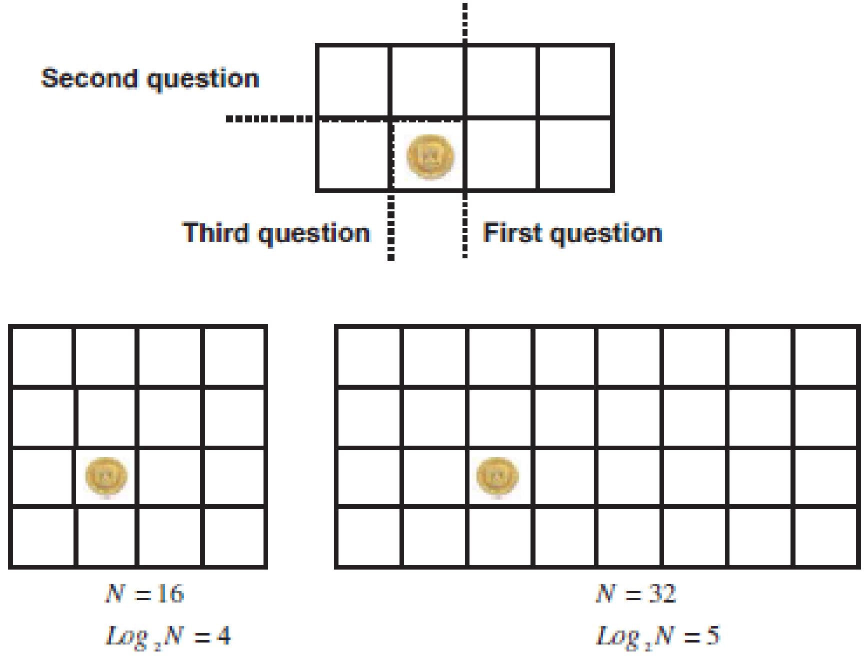


Fig. (2.9)

- La cantidad de información que es necesaria adquirir a base de preguntas está definida en términos de la distribución de probabilidades

$$H = -\sum_k p_k \log\, p_k$$

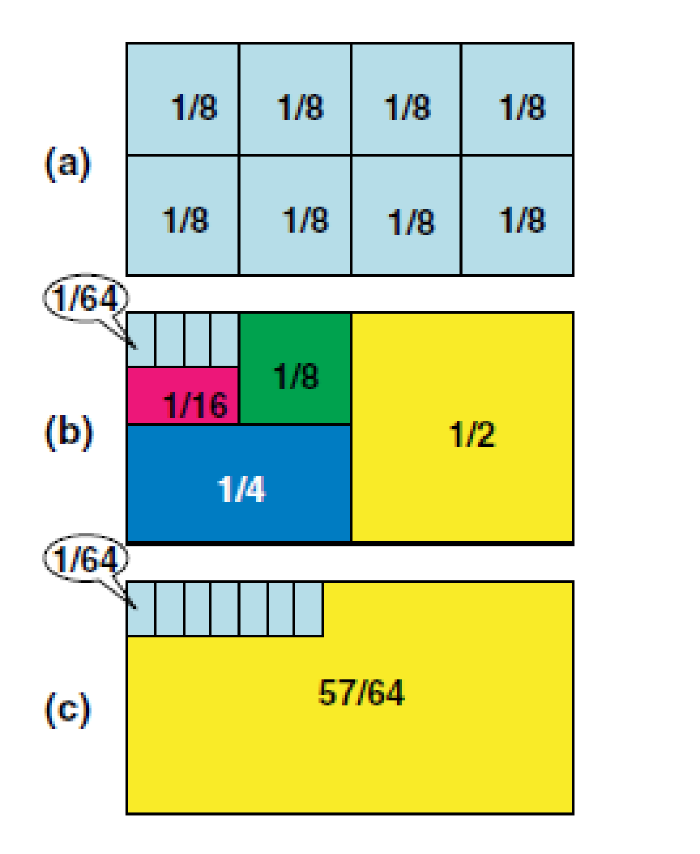
Ben-Naim, A. (2008). *Entropy demystified: The second law reduced to plain common sense*. World Scientific.

- ¿Cuál preferirían jugar?

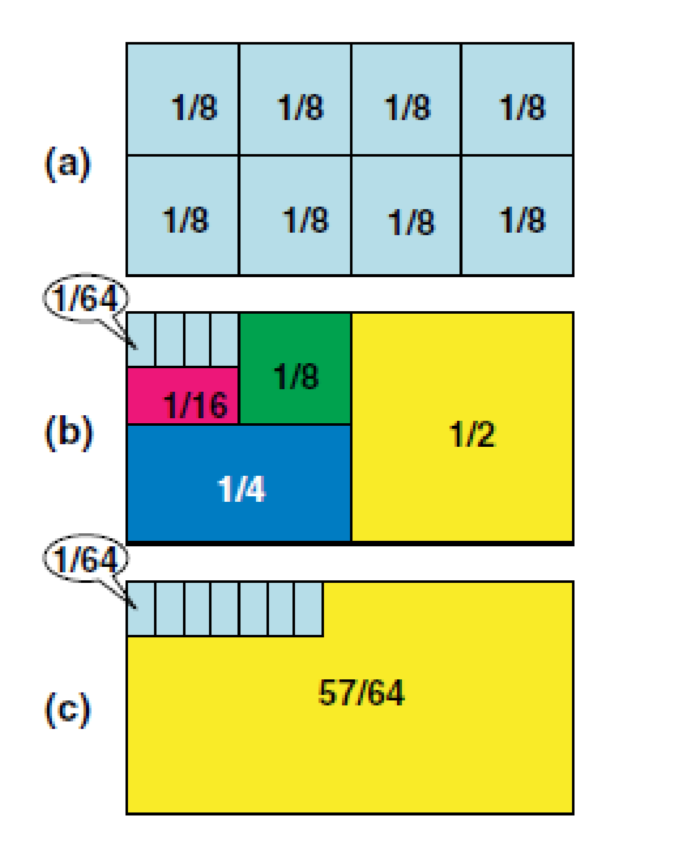
a	b	c	d
1/4	1/4	1/4	1/4

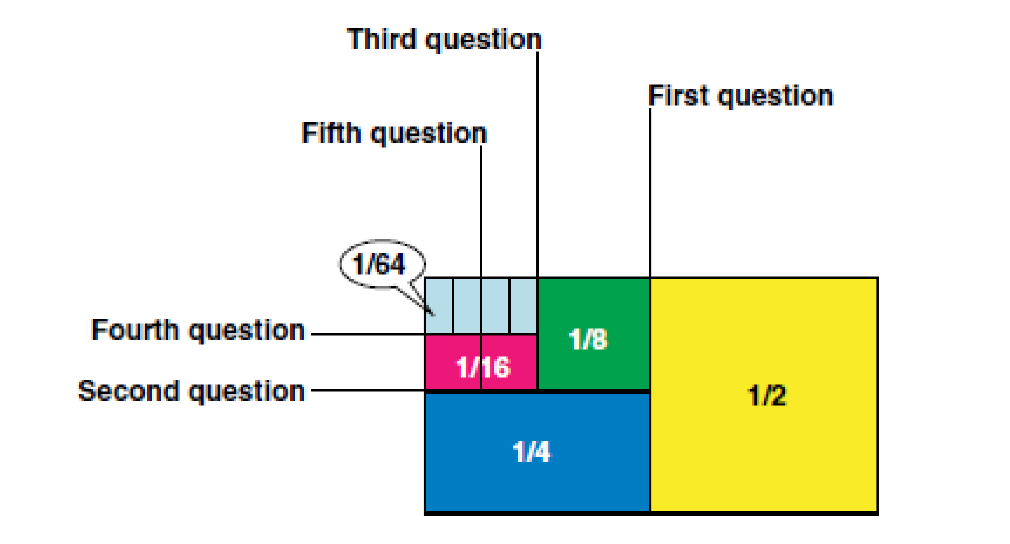
a	b	c
¼	¼	½

- ¿Cuál preferirían jugar?

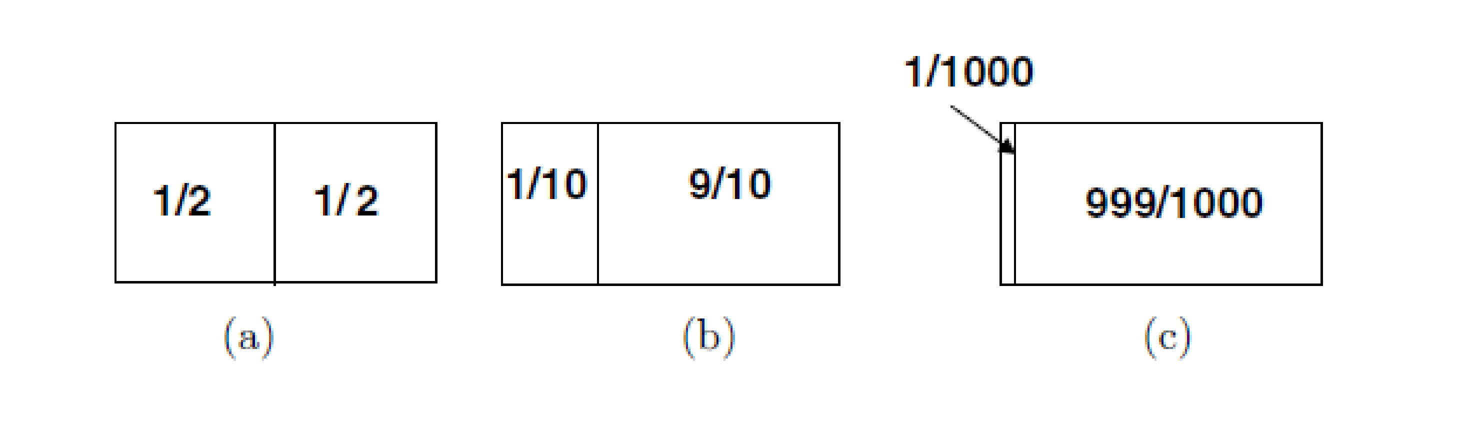


- ¿Cuál preferirían jugar?





- ¿Cuál preferirían jugar?



La entropía de la distribución

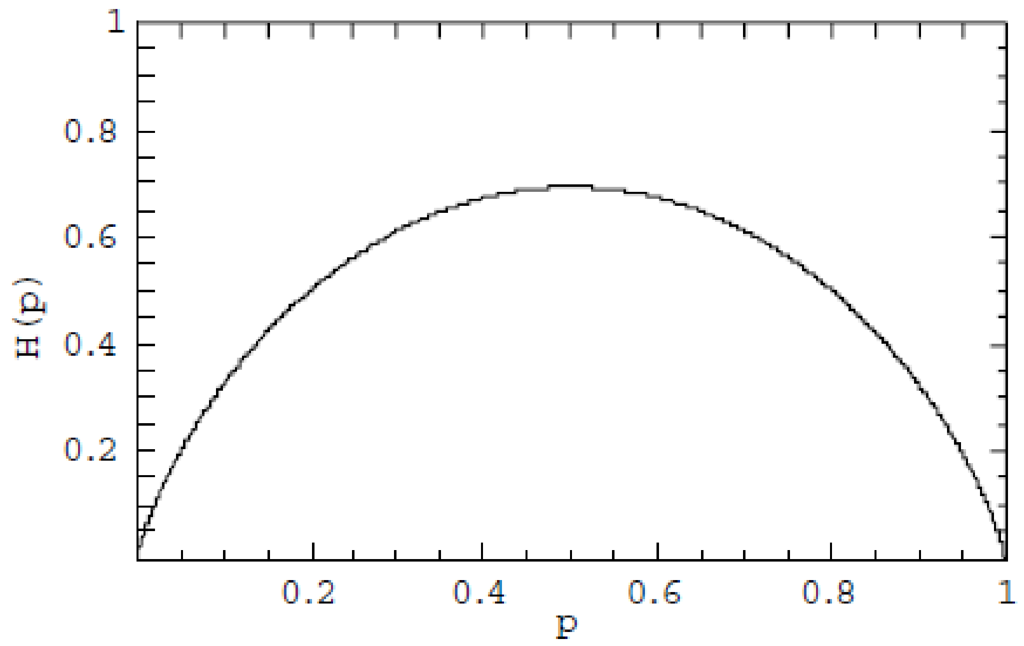


Figure 3.4. The function H for two outcomes; (3.2.2).

$$H = -\sum p_i \log_2 p_i = -p \log_2 p - (1-p)\log_2(1-p).$$

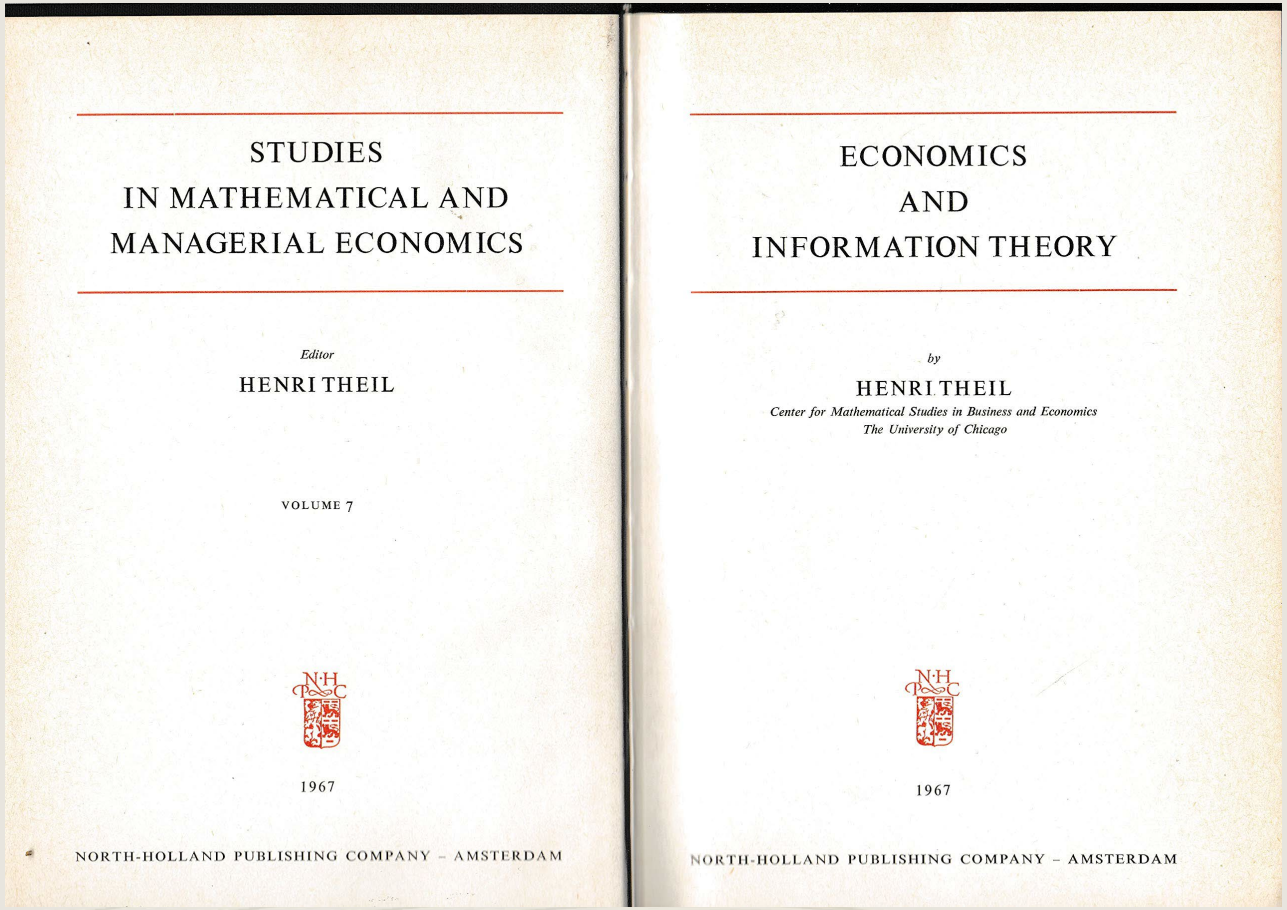
- En un sentido muy intuitivo, si uno de los eventos tuviera probabilidad 1 y el resto probabilidad 0, sabríamos con certeza lo que ocurre y la incertidumbre es baja.
  - Noten que este es el caso en que
    $$H = -\sum_k p_k \log p_k = 0.$$
- De nuevo en un sentido intuitivo, el caso más incierto ocurre cuando todas las probabilidad son iguales $p_k = {}^{1}/_{K}$. Cualquier cosa es igualmente posible.
- Noten que en ese caso la entropía es $\mathrm{H} = -\sum_k p_k \log p_k = \log K$. Ésta crece con K, como más eventos equiprobables haya, mayor será la incertidumbre.

La entropía de la distribución

- *H* (la medida de información de Shannon), provee una medida de información en términos del número **mínimo** de preguntas (binarias) necesarias, en promedio, para conocer el resultado de un experimento, dada la distribución de probabilidad de los posibles resultados.

$$H = -\sum_{k} p_k \log p_k$$

Aplicaciones famosas



STUDIES
IN MATHEMATICAL AND
MANAGERIAL ECONOMICS

Editor
HENRI THEIL

VOLUME 7

1967

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM

ECONOMICS
AND
INFORMATION THEORY

by
HENRI THEIL
Center for Mathematical Studies in Business and Economics
The University of Chicago

1967

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM

Henri (Hans) Theil
October 13, 1924
August 20, 2000

Entropía relativa o divergencia

- El índice de Theil es la redundancia en Teoría de la Información: la entropía máxima posible de los datos menos la entropía observada.
- Piensen en la distribución del ingreso, donde q es el vector de proporciones de ingreso,

$$\mathrm{H} = -\sum_{k} q_k \log\, q_k$$

- Es el menor número más pequeño de preguntas Sí/No necesarias (en promedio) para rastrear el origen de 1 peso extraído aleatoriamente.
- De manera análoga, siendo p el vector de proporciones iguales de ingreso. La entropía de p es el número más pequeño (en promedio) para rastrear el origen de 1 peso extraído aleatoriamente de una distribución uniforme.
- La caída en el número de preguntas necesarias es una medida del grado de desigualdad (distancia en términos de información que media entre dos distribuciones).

Aplicaciones famosas

- **3.5 EXPRESSING THE DEGREE OF UNCERTAINTY: MEAN POSTERIOR PROBABILITIES AND ENTROPY**

$$E = 1 - \frac{\sum_{i=1}^{n} \sum_{c=1}^{C} -p_{ic} \log p_{ic}}{n \log C}, \tag{3.6}$$

• Mean posterior probabilities, the odds of correct classification, and entropy-based measures are tools that can be used to summarize the degree of classification uncertainty in LCA for a particular data set.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons. P. 73-75
https://www.statmodel.com/download/UnivariateEntropy.pdf

Information Entropy

- Information: The reduction in uncertainty when we learn an outcome.
- There is given amount of uncertainty in the true model
- The uncertainty contained (inherent) in a probability distribution is the average log-probability of an event

$$H(p) = -\mathrm{E}\log(p_i) = -\sum_{i=1}^{n} p_i \log(p_i)$$

- How much additional uncertainty is induced by using not the true, but our working model?

Divergence: The additional uncertainty induced by using probabilities from one distribution (working model) to describe another distribution (true model).

$$D_{KL}(p, q) = \sum_i p_i(\log(p_i) - \log(q_i)) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$$

- If we have a pair of candidate distributions, then the candidate that minimizes the divergence will be the closest to the target (true model)

True model?

- KL divergence measures the distance of a working model from our target (true model)
- If we knew the true model, we wouldn't be doing statistical inference!
- We are only interested in in comparing the divergences of different candidates. p just subtracts out leaving the relative distance to the target.
- We can't tell how far any particular archer is from hitting the target, but we can tell which archer gets closer and by how much.

Desviación (deviance)

- As we are only interested in comparing the divergences of different candidates, say q and r, … most of p just subtracts out.
- All we need to know is a model's average log-probability $\mathrm{E}\log(q_i)$ for q and $\mathrm{E}\log(r_i)$ for r.
- Indeed, just summing the log-probabilities of each observed case provides an approximation of $\mathrm{E}\log(q_i)$, just without the final step of dividing by the number of observations.

$$S(q) = \sum_i log(q_i)$$

- To compute the score for a Bayesian model, we have to use the entire posterior distribution. Otherwise, vengeful angels will descend upon you.

Scoring the right data

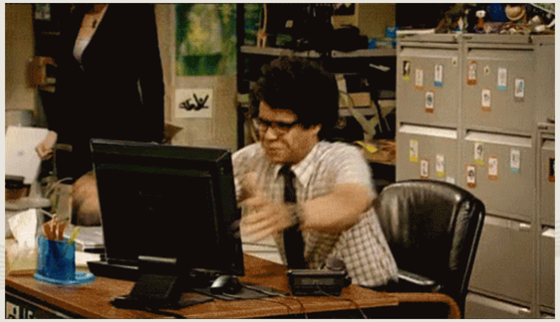
- To compute this score for a Bayesian model, we have to use the entire posterior distribution (log-pointwise-predictive density; lppd).

$$lppd(y, \Theta) = \sum_i log \frac{1}{S} \sum_s p(y_i, \Theta_s)$$

La desviación es $-2lppd$

- We simply cannot score models by their performance on *training* data (more complex models have larger scores!). We need to predict outcomes in a new *test* sample. While deviance on training data always improves with additional predictor variables, deviance on future data may or may not.

LOO-CV does not scale well to large datasets



Validación cruzada y criterios de información

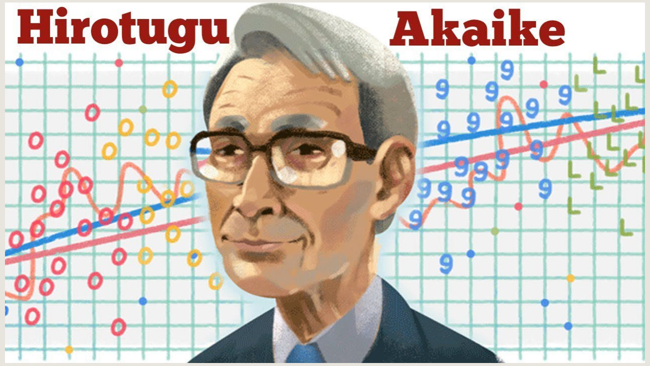
PREDICIENDO EL DESEMPEÑO PREDICTIVO

Information criteria

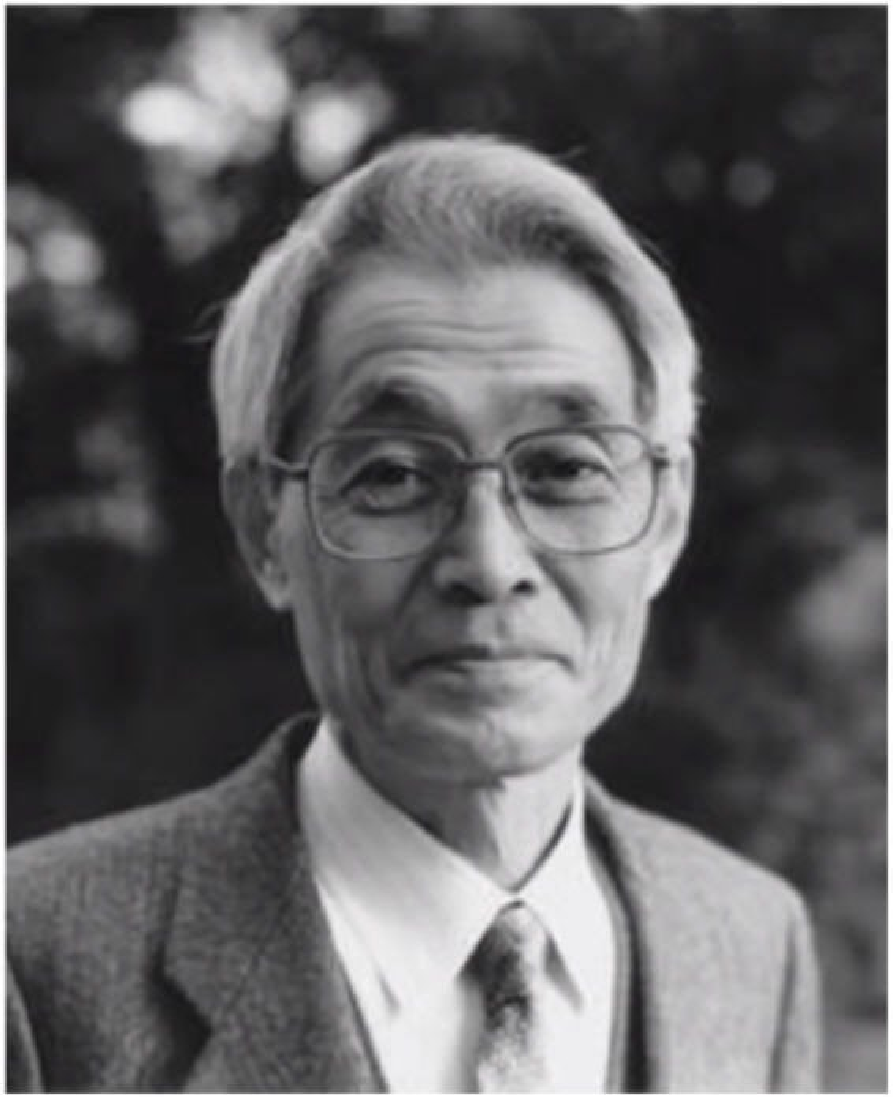
- Constructs a theoretical estimate of the relative out-of-sample KL divergence
  - The difference between training deviance and testing deviance is about twice the number of parameters

$$AIC = D_{train} + 2p = -2lppd + 2p$$

The dimensionality of the posterior distribution is a natural measure of the model's overfitting tendency



"On the morning of March 16, 1971, while taking a seat in a commuter train, I suddenly realized that the parameters of the factor analysis model were estimated by maximizing the likelihood and that the mean value of the logarithmus of the likelihood was connected with the Kullback-Leibler information number."



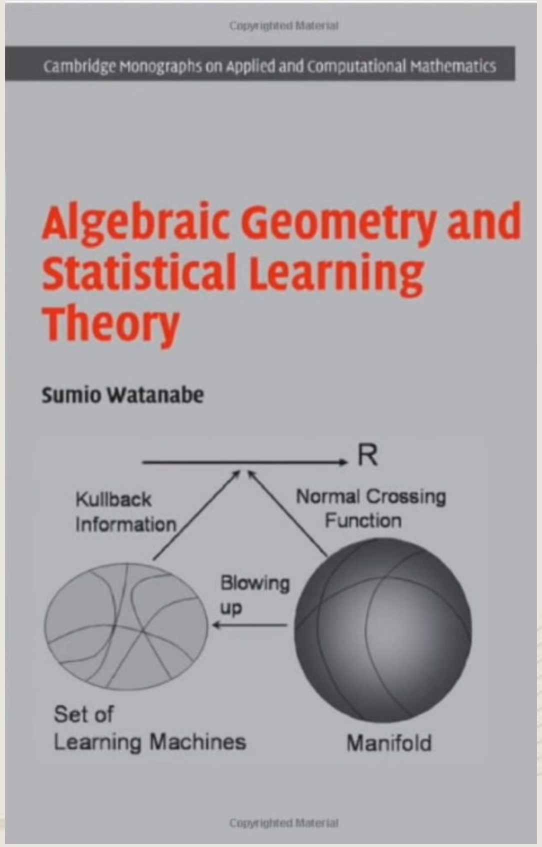
Hirotugu Akaike (1927–2009)

赤池弘次

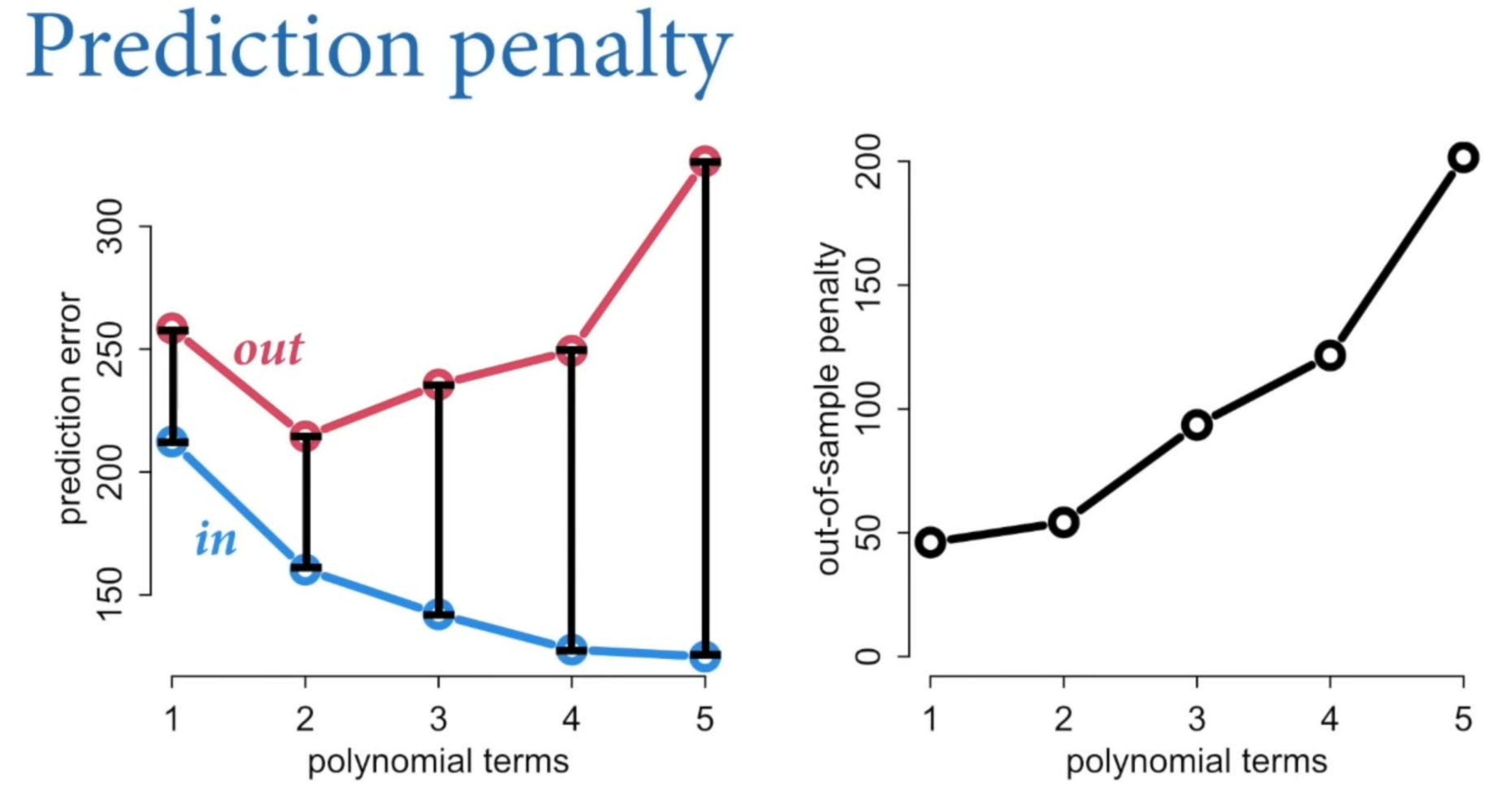
Information criteria

- Newer and more general approximations (of the out-of-sample deviance) exist that dominate AIC in every context: Widely Applicable Information Criterion (WAIC)

$$WAIC(y,\Theta) = -2(lppd - \underbrace{\sum_i \mathrm{var}_\theta \log p(y_i|\theta))}_{\text{Penalty term (effective number of parameters)}}$$

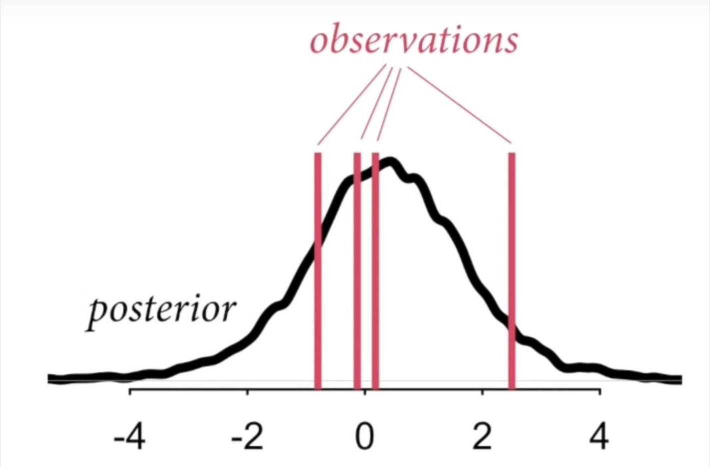


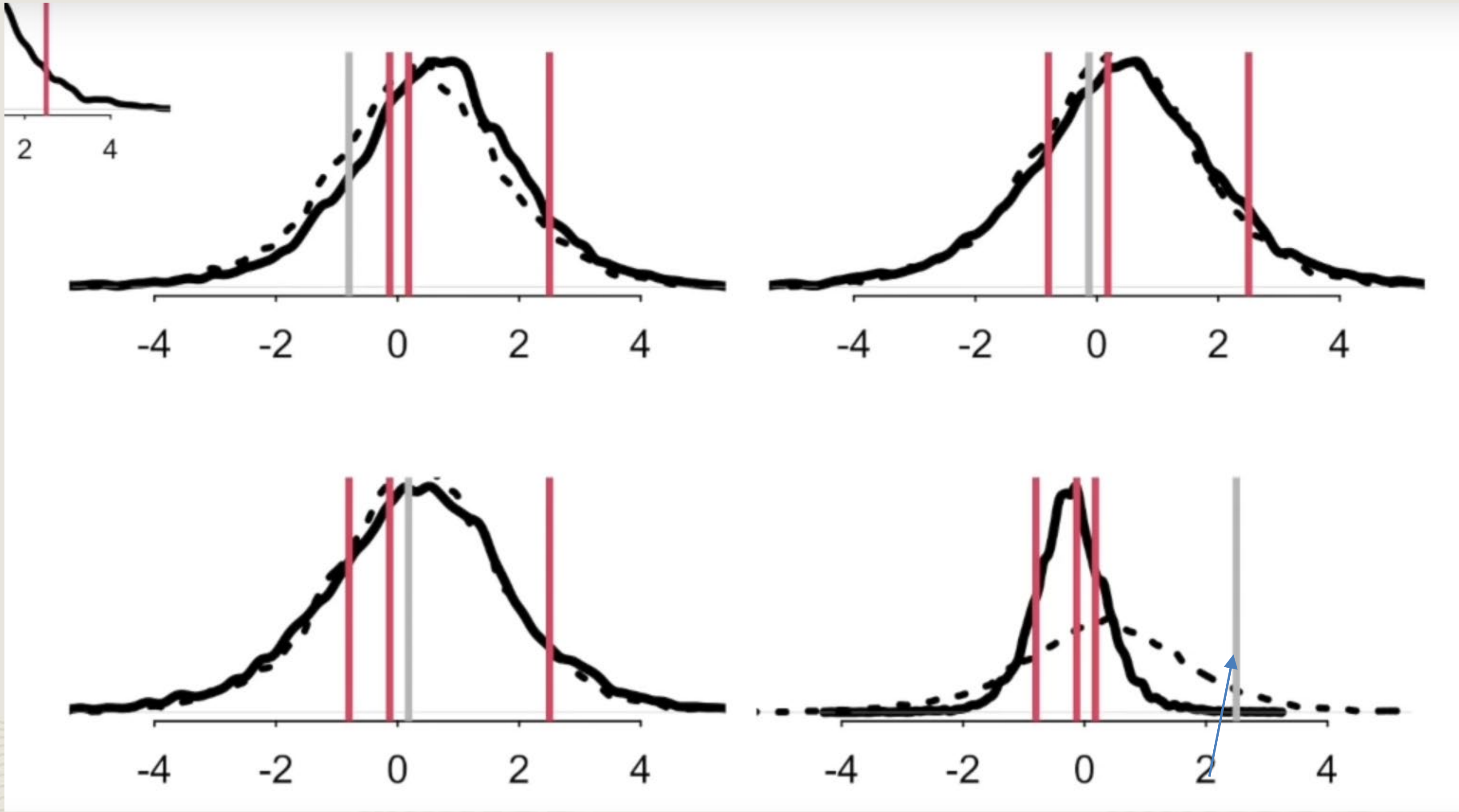
Prediction penality



PSIS $p_M(y_i|y_{-i})$

Pareto smoothed importance sampling for leave-one-out cross-validation (LOO) approximation.





Podemos calcular la probabilidad posterior de dicha observación e inferir la posterior resultante sin la observación

PSIS $p_M(y_i|y_{-i})$

Pareto-smootheed (regularized) importance sampling cross-validation

Estimating out-of-sample pointwise predictive accuracy using posterior simulations

- Importance sampling replaces the computation of *N* posterior distributions by using an estimate of the importance of each i to the posterior distribution.

$$r_i^s = \frac{1}{p(y_i|\theta^s)} \propto \frac{p(\theta^s|y_{-i})}{p(\theta^s|y)}$$

Gelfand, A. E., Dey, D. K., and Chang, H. (1992).

PSIS $p_M(y_i|y_{-i})$

Pareto-smootheed (regularized) importance sampling cross-validation

- Re-weights each sample by the inverse of the probability of the omitted observation
  - performed using existing simulation draws!
  - also obtains approximate standard errors for estimated predictive errors and for comparison of predictive errors between two models.

$$r_i^s = \frac{1}{p(y_i|\theta^s)} \propto \frac{p(\theta^s|y_{-i})}{p(\theta^s|y)}$$

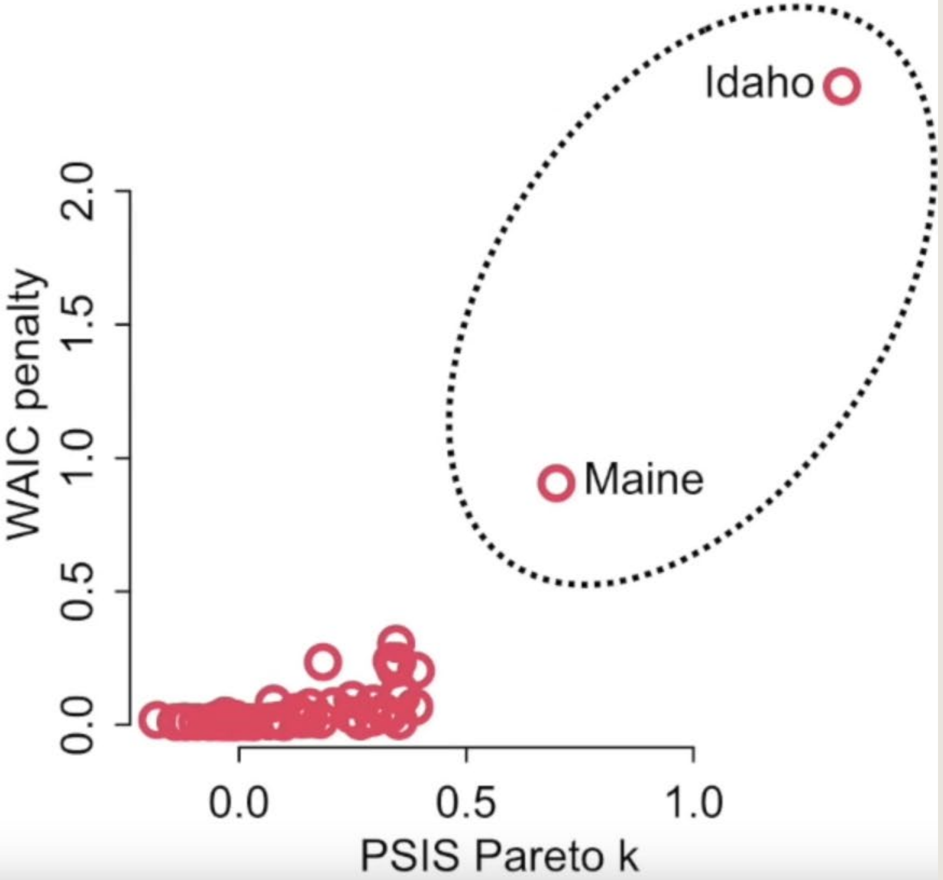
Gelfand, A. E., Dey, D. K., and Chang, H. (1992).

PSIS

Quantify influence:

PSIS *k* statistic

WAIC penalty term ("effective number of parameters")



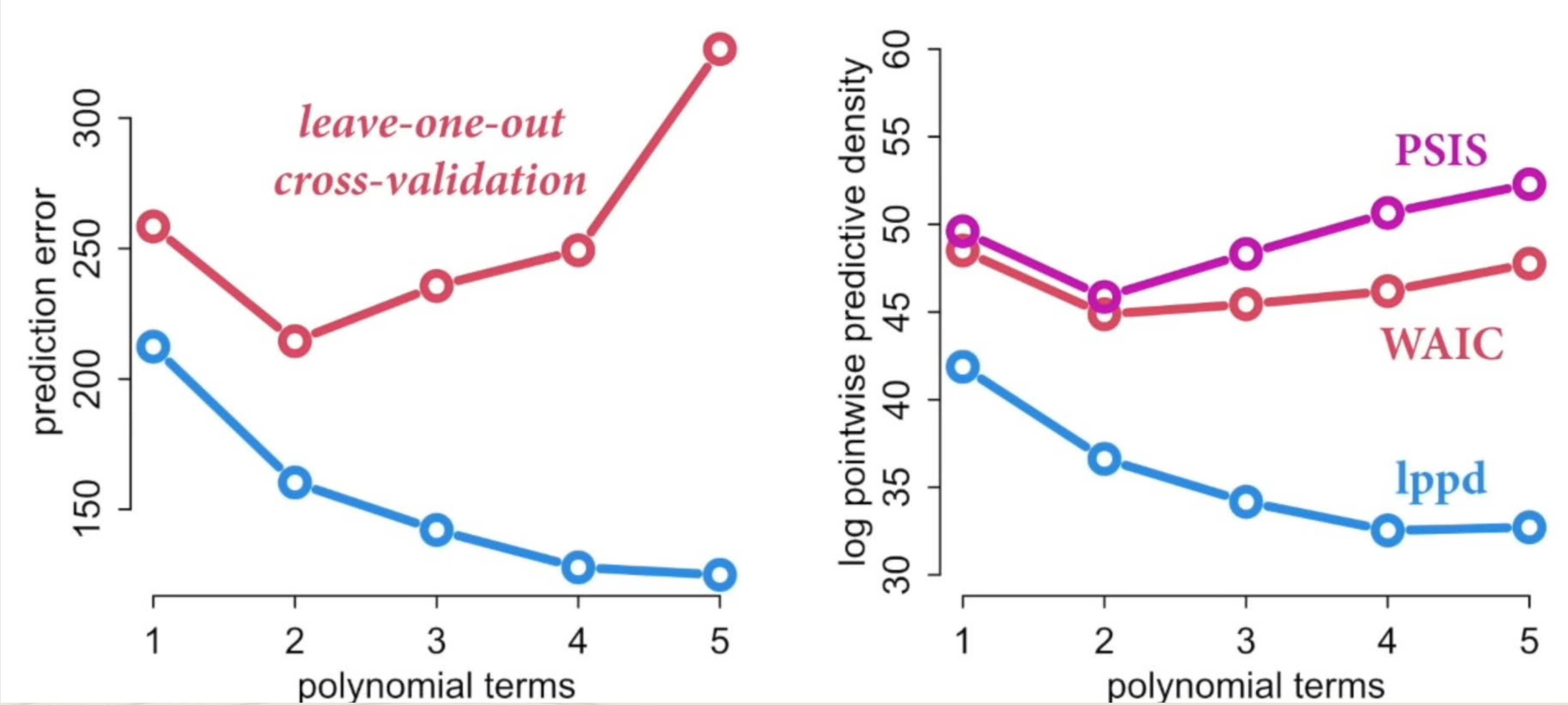
Observaciones en la cola de la distribución predictiva indican un posible exceso de confianza (el modelo no espera suficiente variación). Las predicciones no son confiables

- PSIS: Much more stable and provides diagnostics
- Identifies outliers (indicate something is wrong with the model, dropping them only ignores the problem, predictions will still be bad)
- Estimate of the models performance out of sample

Prof Aki Vehtari (Helsinki), smooth estimator

Vehtari, A., Gelman, A., & Gabry J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. In Statistics and Computing, doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544.

Ejemplo



leave-one-out cross-validation
prediction error
300
250
200
150
1
2
3
4
5
polynomial terms
PSIS
WAIC
lppd
log pointwise predictive density
60
55
50
45
40
35
30
1
2
3
4
5
polynomial terms

Estimación: efficient approximate loo http://mc-stan.org/loo/reference/loo-glossary.html

```
loo1 <- loo(fit1, save_psis = TRUE)
print(loo1)
```

```
Computed from 4000 by 262 log-likelihood matrix

         Estimate     SE
elpd_loo  -6247.8  728.0
p_loo       292.4   73.3
looic     12495.5 1455.9
------
Monte Carlo SE of elpd_loo is NA.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     239   91.2%   200
 (0.5, 0.7]   (ok)         6    2.3%   56
   (0.7, 1]   (bad)        8    3.1%   25
   (1, Inf)   (very bad)   9    3.4%   1
See help('pareto-k-diagnostic') for details.
```

Cuando K pareto es grande!

$P\_loo > p$: Mala especificación

It describes how much more difficult it is to predict future data than the observed data!

Quisiéramos que todas fueran "buenas" <.7

Malas predicciones

Lecciones

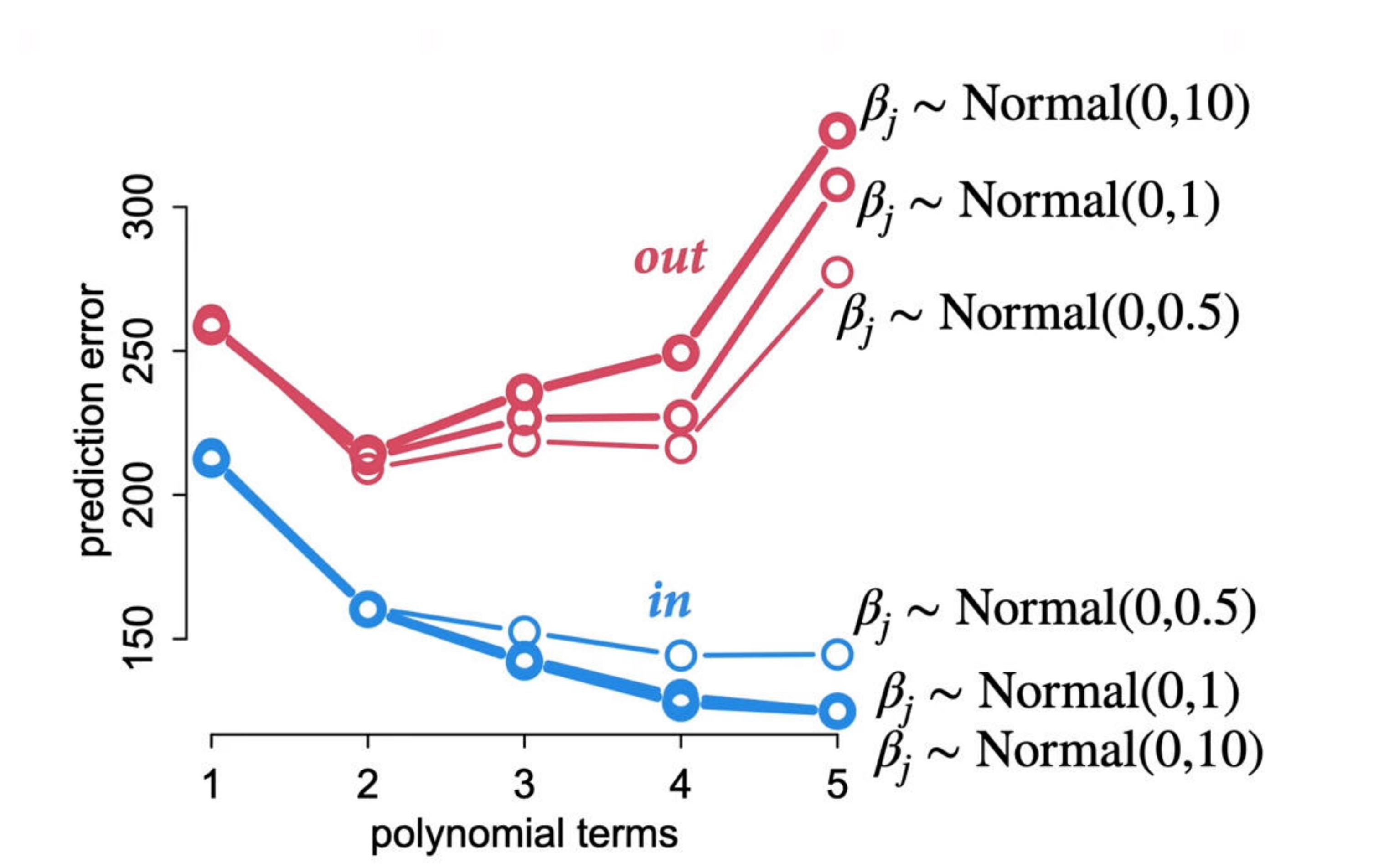
- NO USEN WAIC or PSIS para elegir modelos explicativos –causales-
- El modelo incorrecto puede ser mejor para predecir!

¿Y los priors?

- Cross-validation measures predictive accuracy, but does nothing about it.
- For pure prediction….Tune the prior using cross-validation (regularize, skeptical models tend to do better as they are less excitable)

Lecciones

- Flat priors are the worse
- You can make priors too tight/skeptical and learn nothing from the sample
- For causal inference… use science!
- Muchas tareas son una mezcla de inferencia y predicción



Referencias

Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle", in Petrov, B. N.; Csáki, F. (eds.), 2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971, Budapest: Akadémiai Kiadó, pp. 267–281. Republished in Kotz, S.; Johnson, N. L., eds. (1992), Breakthroughs in Statistics, vol. I, Springer-Verlag, pp. 610–624

Akaike, H. (1985), "Prediction and entropy", in Atkinson, A. C.; Fienberg, S. E. (eds.), A Celebration of Statistics, Springer, pp. 1–24.

Ben-Naim, A. (2017). Entropy, Shannon's measure of information and Boltzmann's H-theorem. *Entropy*, *19*(2), 48.

Ben-Naim, A. (2008). *A Farewell to Entropy: Statistical Thermodynamics Based on Information: S.* World Scientific.

Ben-Naim, A. (2008). *Entropy demystified: The second law reduced to plain common sense.* World Scientific.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification*, *13*(2), 195-212.

Collins, L. M., & Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics* 4, ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 147{167. Oxford University Press.

Magnusson, M., Andersen, M. Jonasson, J., and Vehtari, A. (2019). Bayesian leave-one-outcross-validation for large data. *Proceedings of the 36th International Conference on Machine Learning*, 97:4244–4253.

Rioul, O., & LTCI, T. P. (2018, November). This is IT: A Primer on Shannon's Entropy and Information. L'Information, S´eminaire Poincar´e XXIII. http://www.bourbaphy.fr/rioul.pdf

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27, July and October, 379-423 and 623-656.

Theil, H. (1967). *Economics and Information Theory*. Chicago: Rand McNally and Company.

Vehtari, A., Gelman, A., & Gabry J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. In Statistics and Computing, doi:10.1007/s11222-016-9696-4. arXiv preprint arXiv:1507.04544.

https://www.youtube.com/watch?v=odGAAJDIgp8