LECTURE 10: BEST PRACTICES OF STATISTICAL ANALYSIS

- How to handle missing data and selection effects
- Model checking, evaluating: how to check whether the models we developed fit the data: Bayesian goodness of fit test
- Model comparison: which model fits the data better and how to answer this when number of parameters differ
- Dirty error estimations: bootstrap, jackknife, cross-validation tests
- Bad practices (p-hacking), good practices (blind analysis)
- Decision theory: how to make the final decision

How to handle missing data and selection effects

- We have already seen this in HW 1 (MacKay 3.1): decay events can only be observed between 1cm and 20cm away from the source.
- We can restrict the data into an interval that cover the interval $1\text{cm} < x < 20\text{cm}$, such that the probability distribution over the interval integrates to 1
- Once we have a proper PDF we can turn it into a likelihood, which multiplied with the prior gives the posterior
- These effects need to be modeled, even more so if we have noise

Example: Malmquist Bias

- Flux limited surveys of galaxies have a luminosity limit that changes with distance and the measured sample is not representative of the whole sample

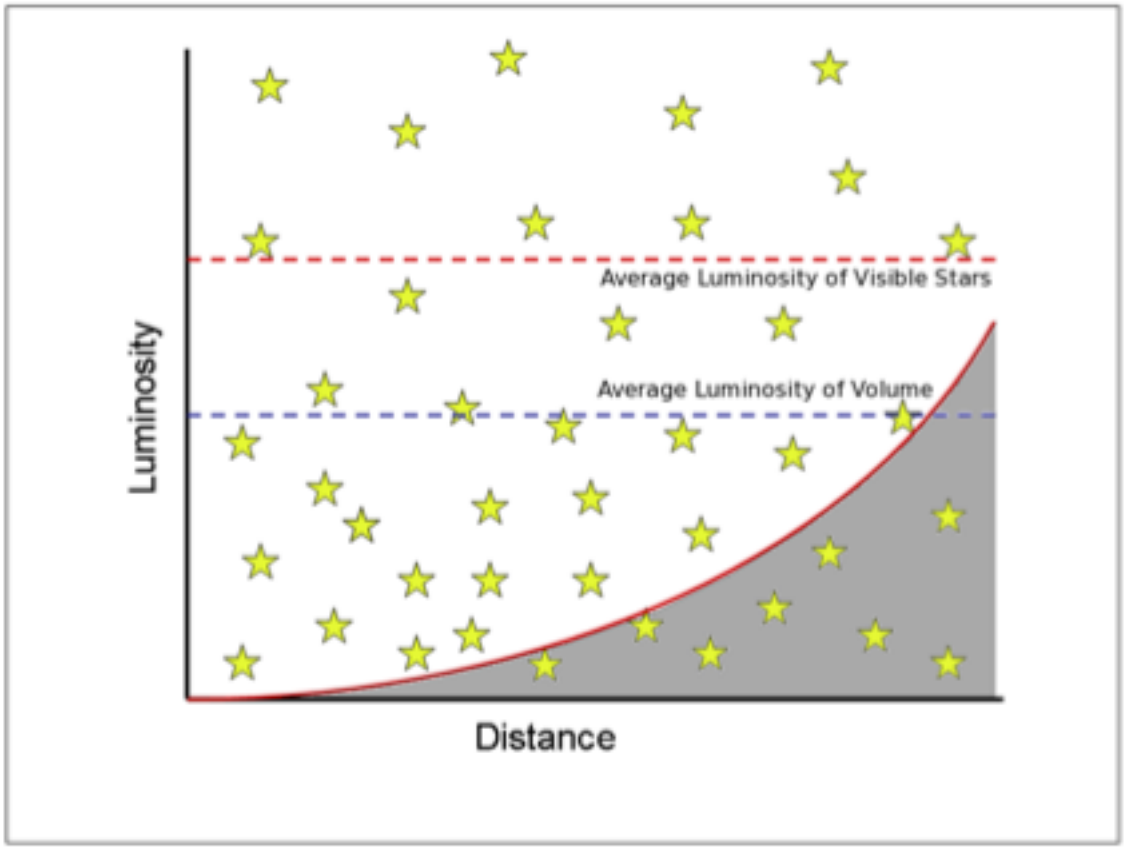
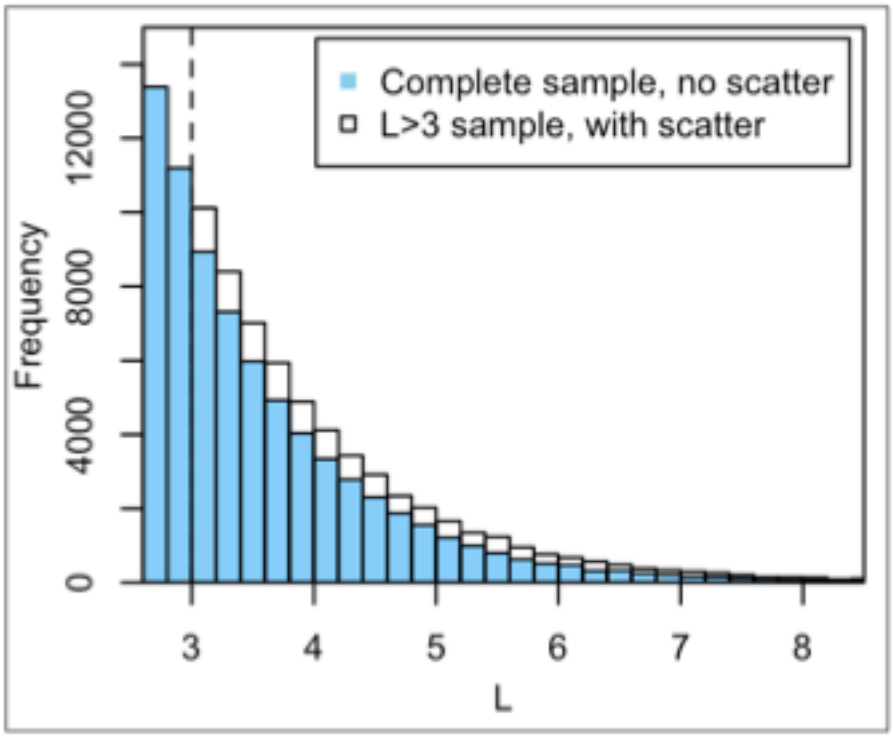


Image credit: Wikimedia Commons user Galaxy1F10 (public domain)

Example: Eddington Bias

- We have a population of sources with different luminosities, which we can observe only down to a certain value (due to the flux limit). There are more faint sources than bright ones. The data are noisy and we impose a flux limit.

  Because of this, there will be a scatter of faint sources into the sample and a scatter of bright sources out of the sample, but since there are more faint than bright sources the net effect is to have more faint sources in the sample than in the absence of noise.



A bit more formally

- Stability assumption: a measurement does not affect the value of the data (obviously we are not in quantum realm here)
- This also means we can define inclusion vector I, which tells us which of the data y are included (value 1) and which excluded (value 0, with no overlap between the two)
- $P(y,I|\theta,\phi) = p(y|\theta)p(I|y,\phi)$ where ϕ are parameters defining the inclusion vector I, y is all the data, which can be decomposed into y_{obs} and y_{mis}
- We need to marginalize over missing data y_{mis}
- $P(y_{obs},I|\theta,\phi) = \int P(y,I|\theta,\phi)dy_{mis}$
- This is the observed data likelihood

Ignorability

- Just because the data are missing does not mean we have to worry about its selection: e.g. if the data are a random subsample of the original full data
- When can we ignore the selection? Obviously when the posterior $P(\theta|y_{obs},I)= P(\theta|y_{obs})$. We need:
- Missing at random: $P(I \mid y, \phi)= P(I \mid y_{obs}, \phi)$ and distinct parameters ϕ (independent of θ in prior): $P(\phi|\theta)=P(\phi)$
- If this is not satisfied we need to model missing data
- We do this with simulations: 1) imputation of missing data from their posterior predictive distribution given observed data
- 2) draw from this posteriors of θ

Model Checking and Sensitivity Analysis

- After we have done our Bayesian analysis we want to make sure the model is good
- We also want to look at how sensitive our conclusions are to a change of the model (e.g. change of prior, likelihood etc)
- there is one proper Bayesian way to do this and many non-Bayesian ways that are often simpler
- In theory, Bayesian analysis is supposed to perform the analysis over all the viable models. If so, and if the likelihood is correct, then the posterior correctly summarizes the information gained relative to the prior
- In practice this may be impossible to do: we never have access to all parameters and their predictions
- Likelihood is also an approximation that needs to be tested

Sensitivity Analysis

- For example, we try to determine x marginalized over nuisance parameters y. If we worry about sensitivity to nuisance parameters y we can enlarge their parameter space and look at marginalized posterior $p(x)$.
- For example, suppose we know our model is not very good over some range of data space, but is very good elsewhere. We enlarge nuisance parameter space such that its spans all possible outcomes of data in that part of space, even if the nuisance parameters are artificial.
- If the posterior $p(x)$ does not change as we add more nuisance parameters this suggests results are not sensitive to priors and conclusions are robust.
- Alternatively, we can also increase the errors in the data space that is poorly modeled, but that is harder to do.

Agreement between Datasets

- Suppose we want to test agreement between two data sets that are suppose to measure the same parameter x. For example, we have noticed some discrepancy between the two, but we are not sure if it is statistically significant.
- Best approach is to introduce a new physical parameter that encapsulates the differences between what the two experiments are measuring: perhaps there is an enlarged physical model that allows the two experiments to probe another parameter via the difference between the two experiments
- In this case the discrepancy is leading to a new scientific discovery

Agreement between Datasets

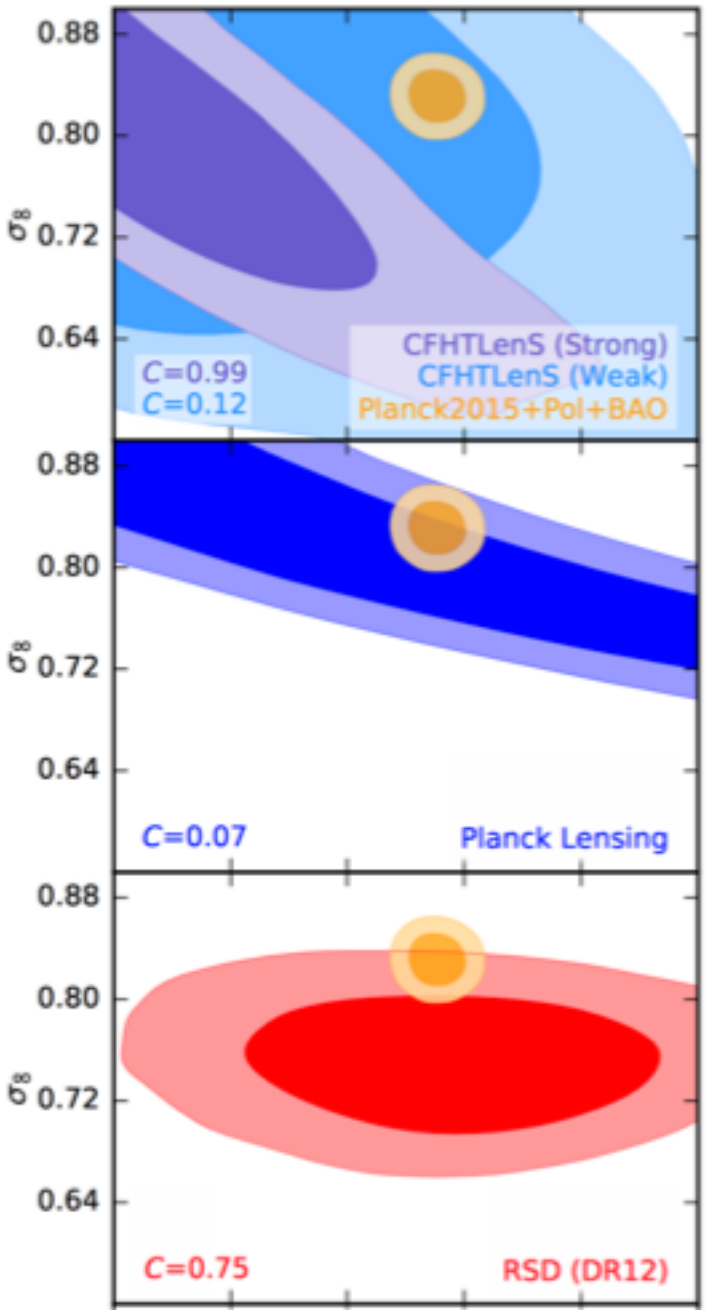
- Less well physically motivated, but more non-parametric, is to simply define $x_1 = x$ and $x_2 = x+\delta x$ where experiment 1 measures x_1 and experiment 2 measures x_2 and compute the posterior of δx:

  $$P(\delta x) = \int P_1(x_1) P_2(x_2 = x_1 + \delta x) dx_1$$

- If posterior of δx has a strong support away from $\delta x = 0$ this suggests the two experiments are not agreeing on x
- Note that δx can be multidimensional, in which case one is looking at integrals over constant isocontours of $P(\delta x)$ integrated over $P(\delta x) > P(0)$. $C = \int P(\delta x) d\delta x$
- If C is close to 1 it means most of the posterior volume excludes $\delta x = 0$

Example

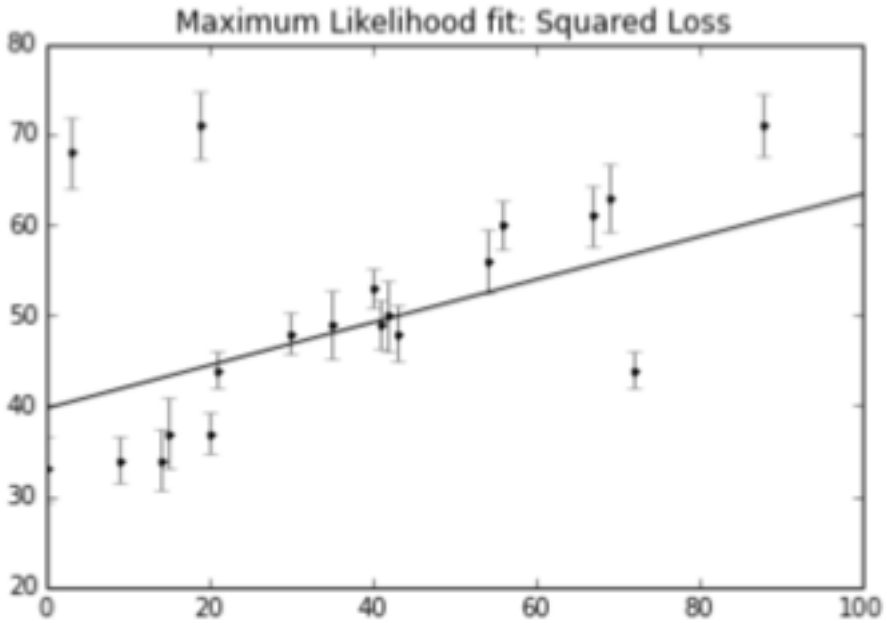
- Here Planck data are compared to other data on 2 parameters
- There is discrepancy against CFHTLens (Strong) with $C = 0.99$
- $C = 0.99$ corresponds to 2.5 sigma



Credit: arXiv:1703.05959

Beyond Prior-to-Posterior Analysis

- Often it is not practical to look at all possible model variations, so we want to test the model in a non-parametric way
- First thing to do is visual inspection: we have seen this in the outlier example
- Clearly the model is inadequate: can we quantify it?



Model Accuracy and Efficiency

- Accuracy: an accurate model is one that can generate data like observed data: it fits the data
- We asses this visually, with test statistics and with adding additional parameters
- An efficient model is one that achieves this with the least number of parameters (Occam's razor)
- We asses this with Bayesian evidence, information criteria or cross-validation

Goodness of Fit Tests

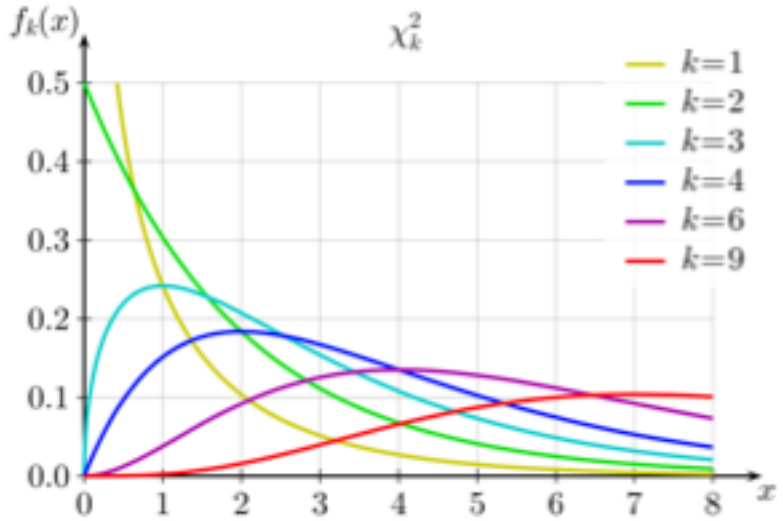
- A typical frequentist method is to compute χ^2 evaluated at the MAP/MLE

$$\chi^2(\mathbf{a}) = \sum_{i=0}^{N-1} \left[\frac{y_i - y(x_i|\mathbf{a})}{\sigma_i} \right]^2$$

- For Gaussian errors its distribution follows $\chi^2(N_{dof})$ distribution: this can be used as a summary measure of discrepancy using a point estimator (MLE or null hypothesis)
- Here $N_{dof} = N - M$ (N data, M model parameters)

Goodness of Fit Tests

- By central limit theorem as we increase N_{dof} the distribution is closer to a Gaussian



- The distribution peaks at N_{dof} and its variance is $(2N_{dof})^{1/2}$ (each measurement is independent and the variance of each term is 2)
- Sometimes one uses reduced χ: χ^2 / N_{dof}, which is peaked at 1, with variance $(2/N_{dof})^{1/2}$
- We can convert deviation of χ^2(MLE) from N into a p-value: this fails if PDF is not Gaussian. We will deal with this (slide 17,18).

Example

- Compute reduced χ^2, 1 sigma χ^2 range, number of sigma deviation, approximate p-value for the following:

1. $\chi^2(\text{MLE}) = 10$, $N_{dof} = 8$
2. $\chi^2(\text{MLE}) = 100$, $N_{dof} = 80$
3. $\chi^2(\text{MLE}) = 1000$, $N_{dof} = 800$

1. Reduced $\chi^2 = 1.25$, $d\sigma = 2/16^{1/2} = 0.5$, $p = 0.6$
2. Reduced $\chi^2 = 1.25$, $\sigma = 20/160^{1/2} = 1.6$, $p = 0.1$
3. Reduced $\chi^2 = 1.25$, $\sigma = 200/40 = 5$, $p = 3\text{x}10^{-7}$

- Note: we need to decide if we want one tailed (from +/- infinity to the value) or two tailed p-values

Discrepancy Measures *T*

- *T* can be χ^2 (MLE) or some other statistic that describes the deviation of the model from data
- If the likelihood is not Gaussian we can use $T = -\log L$.
- Define replicated data y^{rep} using predictive distribution of similar data, which are drawn from a fixed θ (simulations)

- Classical p-value for a test statistic *T* $p_C = \Pr(T(y^{\text{rep}}) \geq T(y)|\theta)$

- Here parameters θ are fixed (at MLE or null hypothesis): the test says that if the data are similar to the replicas p_c will not be an outlier

A more Bayesian Version: Posterior Predictive p-value

- Bayesian version: we marginalize over θ: we draw θ from posterior and then draw y^{rep} given θ, then compute $T(y^{rep}, \theta)$

$$p(y^{\rm rep}|y) = \int p(y^{\rm rep}|\theta)p(\theta|y)d\theta.$$

- If needed we also need to compute $T(y, \theta)$ at θ. Then we ask for each draw of θ and y^{rep} whether $T(y^{rep}, \theta) > T(y, \theta)$. If yes we assign 1 otherwise 0 (Indicator function I), and average over y^{rep}, θ
- This works for any probability distribution, not just gaussian χ^2 : no need to compute it analytically, use simulations
- PPP: posterior predictive p-value p_B

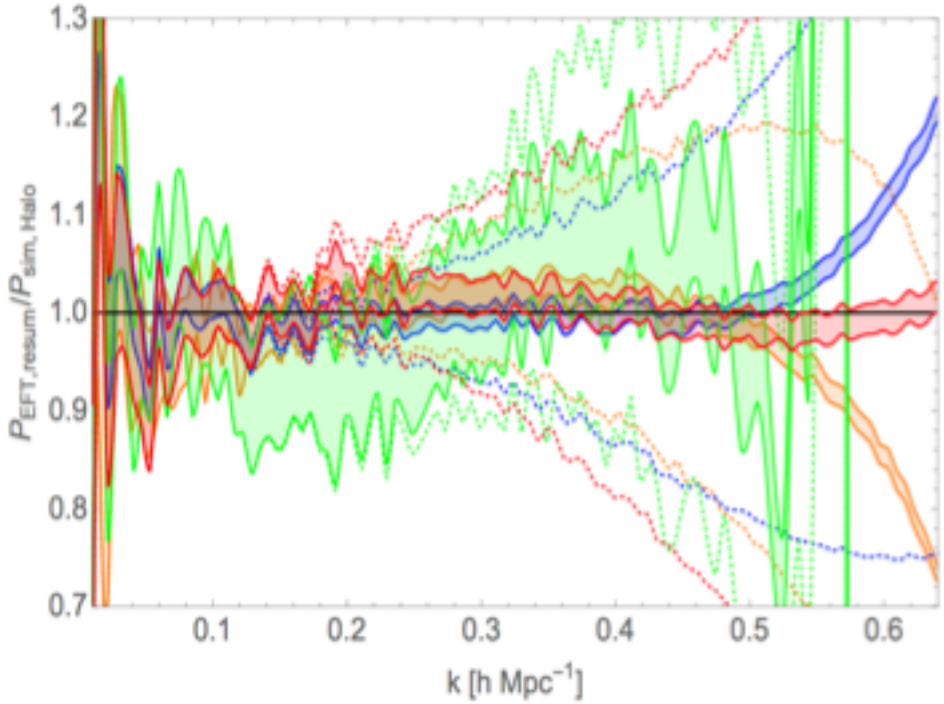
$$p_B = \Pr(T(y^{\rm rep}, \theta) \geq T(y, \theta)|y)$$

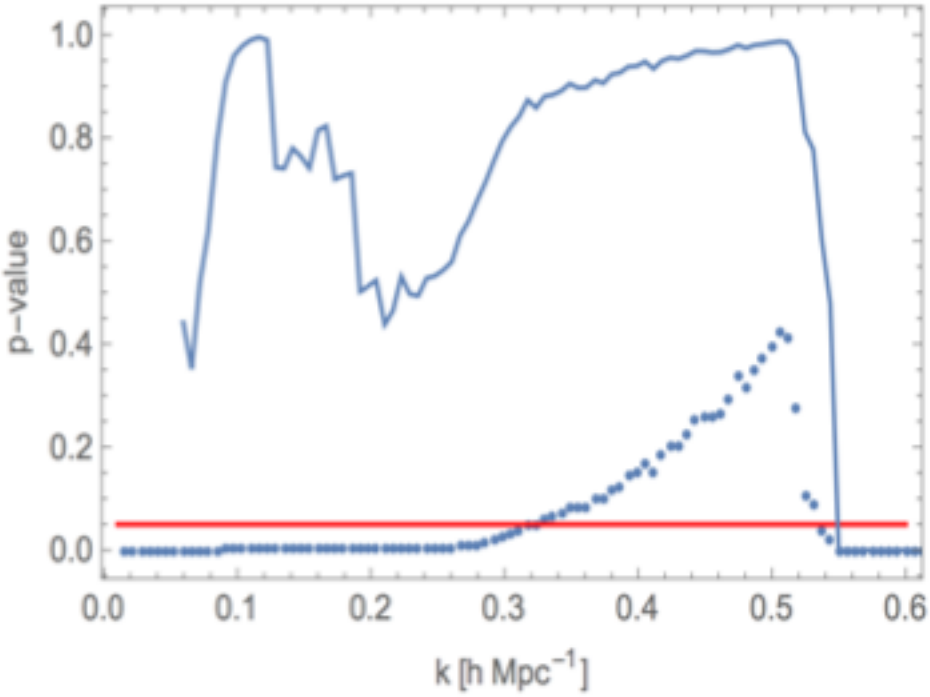
$$p_B = \int\int I_{T(y^{\rm rep},\theta)\geq T(y,\theta)} p(y^{\rm rep}|\theta)p(\theta|y)dy^{\rm rep}\, d\theta,$$

- PPP is still not fully Bayesian: we had to make up T. No guarantee we have chosen it well. Many possible discrepancy measures T, not just χ^2

PPP/Goodness of Fit is not sufficient

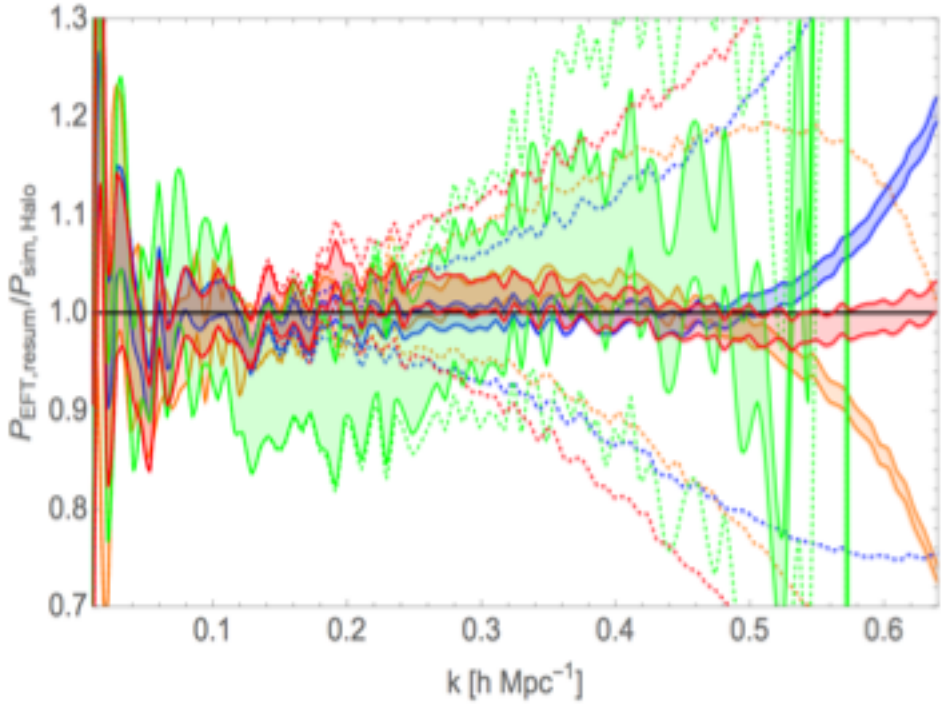
- Example: goodness of fit can be good, but the model can still be bad
- For example, with N_{dof} = 800 one is allowed to have 1 sigma deviation of χ^2 of $(2*800)^{1/2} = 40$ and still have a good goodness of fit

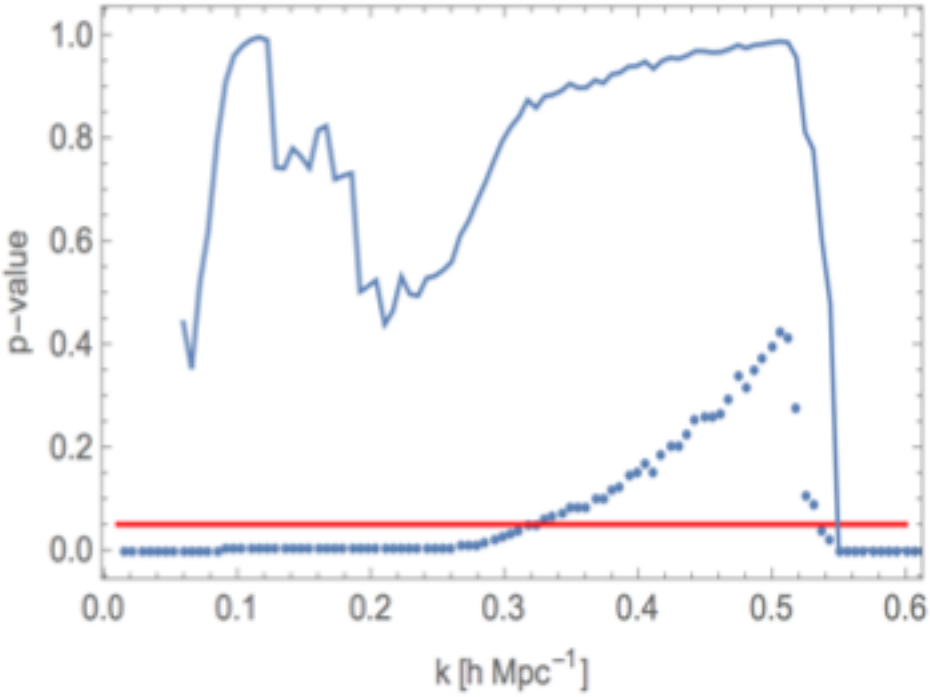




PPP/Goodness of Fit is not sufficient

- We can introduce a new parameter that improves the fit by $\Delta\chi^2 = 40$ and the fit is now better, the new model is preferred by more than 6 sigma
- The strongest model checking is always based on adding additional parameters
- Things to look for in visual inspections: correlated model deviations from data





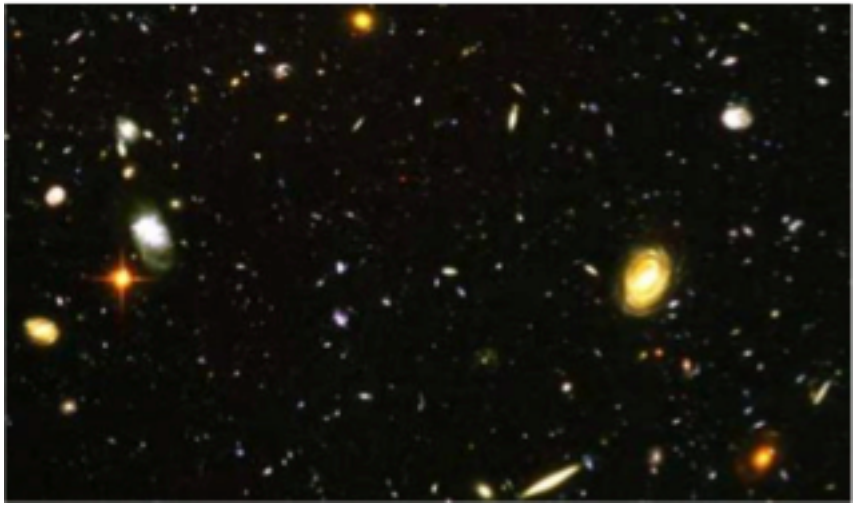
Bayesian Model Comparison

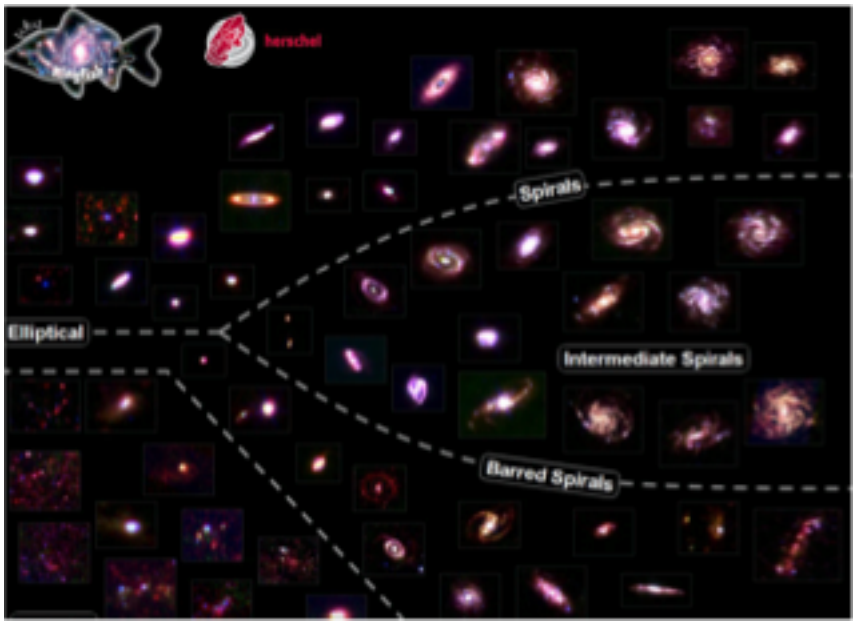
- We discussed hypothesis testing in the beginning of the class, where we computed marginal likelihood or evidence (Bayes factor): $p(d|H)=\int p(d|\theta,H)p(\theta|H)d\theta$
- Evidence is the mean likelihood averaged over parameters θ of H
- Computing evidence requires an integral over all prior space of θ: this includes regions where likelihood $p(d|\theta,H)$ is very small. This makes evidence very susceptible to the prior. This is qualitatively different from the methods used so far, and is the reason why model comparison is sometimes viewed with suspicion: this is a valid criticism if priors are not well justified
- Computing evidence is expensive and requires specialized methods (e.g. nested sampling etc.) because one may have large volumes with low (but non-zero) likelihood so one needs to evaluate likelihood everywhere
- To compare two models, assuming their priors are equal, compute the ratio of their evidence $p(d|H_1)/p(d|H_2)$

Bayesian Model Comparison for Classification: example

- Suppose you want to decide if an object is an elliptical (E) or spiral (S) galaxy.
- There is no single description of E and S galaxy: they come in different sizes, shapes, profiles, ellipticities etc
- We can parametrize the hypothesis classes H_E and H_S in terms of these parameters. Each class has its own prior distribution $p(\theta|H)$.
- For example, ellipticals have very compact radial distribution relative to spirals, so $p(\theta|H_E)$ will be high for compact radial profile, relative to $p(\theta|H_S)$
- If it is true H the data likelihood will give high weight to high prior region

$$p(d|H) = \int p(d|\theta,H)p(\theta|H)d\theta$$





Information Criteria

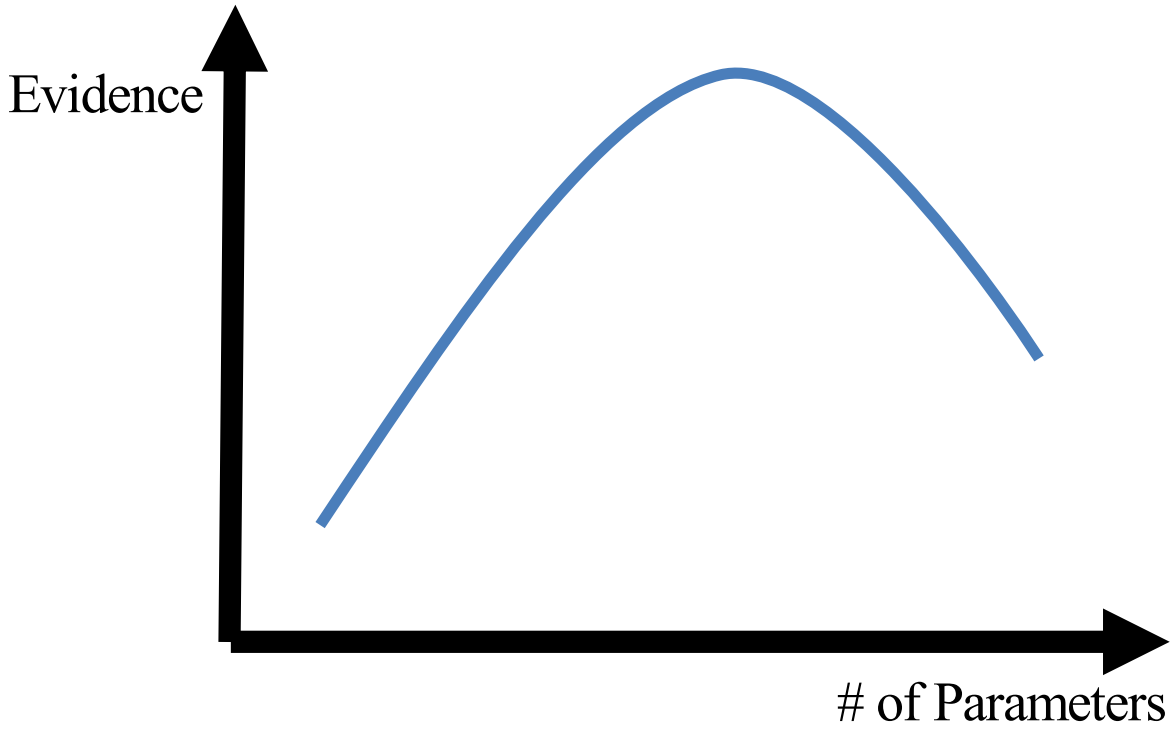
- These have been developed as simpler (and sometimes less controversial) alternatives to Bayesian evidence
- To simply: these methods use some form of $-2\log L$, which for gaussian errors is $\chi^2(N_{dof})$, where $N_{dof} = N - M$, with M number of parameters: more parameters, better χ^2, but at a cost of higher complexity (lower efficiency). Estimating number of parameters is hard, specially with regularization
- To look at efficiency one thus adds some version of effective number of parameters to the information criterion (e.g. AIC adds 2M: derived using KL divergence information theory)
- Names such as AIC (Akiake information criterion), WAIC (Watanabe-Akaike), BIC (Bayesian, similar to evidence, but cheaper), DIC (deviance)…
- Cross-validation another option, discussed in slide 30

Information Criteria

		No pooling ($\tau = \infty$)	Complete pooling ($\tau = 0$)	Hierarchical model (τ estimated)
AIC	$-2\,\text{lpd} = -2\log p(y\|\hat{\theta}_{\text{mle}})$	54.6	59.4	
	k	8.0	1.0	
	$\text{AIC} = -2\widehat{\text{elpd}}_{\text{AIC}}$	70.6	61.4	
DIC	$-2\,\text{lpd} = -2\log p(y\|\hat{\theta}_{\text{Bayes}})$	54.6	59.4	57.4
	p_{DIC}	8.0	1.0	2.8
	$\text{DIC} = -2\widehat{\text{elpd}}_{\text{DIC}}$	70.6	61.4	63.0
WAIC	$-2\,\text{lppd} = -2\sum_i \log p_{\text{post}}(y_i)$	60.2	59.8	59.2
	$p_{\text{WAIC 1}}$	2.5	0.6	1.0
	$p_{\text{WAIC 2}}$	4.0	0.7	1.3
	$\text{WAIC} = -2\widehat{\text{elppd}}_{\text{WAIC 2}}$	68.2	61.2	61.8
LOO-CV	$-2\,\text{lppd}$		59.8	59.2
	$p_{\text{loo-cv}}$		0.5	1.8
	$-2\,\text{lppd}_{\text{loo-cv}}$		60.8	62.8

Evidence/IC as a Function of # of Parameters

- $p(d|H) = \int p(d|\theta,H)p(\theta|H)d\theta$
- Without regularization we expect it to peak somewhere (with regularization not): initially we fit the data better with more parameters. With large number of parameters we start getting more an d more regions of low likelihood within the prior

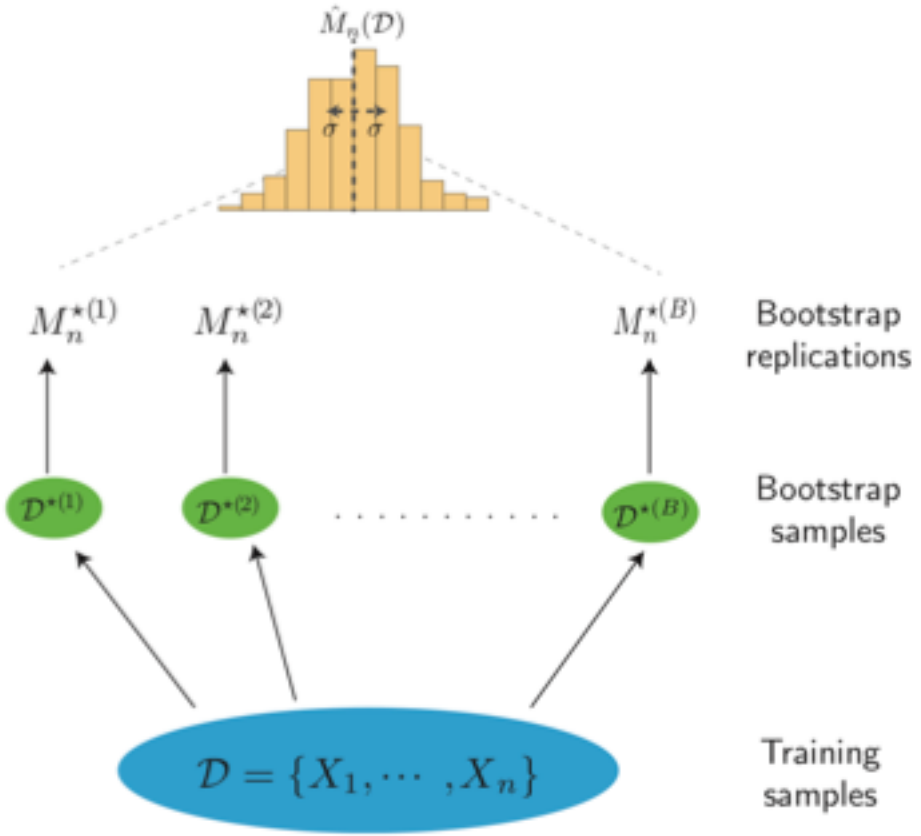


Resampling Methods: Bootstrap, Jackknife, …

- These are classical (frequentist) methods, but c an be useful also for Bayesian analysis
- Main idea: assuming exchangeability of data (a bit more general than iid-identical, independent distributed), we can divide the data into random or systematic subsamples and use the resulting distribution to determine the errors
- This is a classical analysis since it tries to mimic random data realizations (remember that in Bayesian analysis we only view model parameters as random and data is fixed).
- How can we use this in Bayesian analysis? Suppose we are working with summary statistics for which we do not have a good way to evaluate the likelihood. We can compute their covariance matrix (or even full posterior) using resampling methods, and then use this in the subsequent Bayesian posterior analysis of parameters that influence summary statistics

(Nonparametric) Bootstrap

- Nonparametric: drawn from the data
- Parametric: drawn from simulations
- We take a sample of *N* data measurements then draw from it a random sample of *N* data, with replacement: we simply draw at random *N* numbers from 1 to *N*: on average 1/e=37% of data are replaced with a duplicate of another data.
- We perform data analysis as usual, estimating a summary point statistic M for parameters θ (which does not have to be MAP/MLE). We obtain a PDF which mimics well the posterior of the parameter.



(Nonparametric) Bootstrap

- This fails if we fail to have exchangeability or iid. For example, if the data are a realization of a variable, but the variable is stochastic, then we will get an underestimate of variance. This often happens with Fourier analysis where we try to estimate the variance on the largest scale of the problem, but we only have one realization of the variable.
- Parametric bootstrap (simulations) is a solution to this problem, as long as the generative model is valid: we call this Monte Carlo simulations and is the most common method to determine the (co)variance of the parameters.

Jackknife Resampling

- This is a linear approximation of bootstrap
- Here we leave one measurement out, do the analysis on remaining $n - 1$

$$\bar{x}_i = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} x_j$$

- The mean is the same as if we averaged original n samples

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} \bar{x}_i$$

- The variance is $\mathrm{Var}(\bar{x}) = \frac{n-1}{n} \sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2$

Cross-Validation Methods

- The closest analog to jackknife/bootstrap etc. in Bayesian analysis is to

  1) remove part of the data

  2) fit the model to the smaller data set

  3) sample from the predicted model posterior

  4) evaluate the log posterior of the left out data for each of the model samples

  5) average log posterior over the model samples

  6) repeat by removing some other data; average

- Options: leave-one-out (LOO-CV, similar to jackknife), expensive since one needs to repeat *N* times
- Leave-out p (expensive if all possible permutations!), or remove 1/10 and repeat 10 times, or ½, 1/100…
- Bootstrap

Information Criteria

		No pooling ($\tau = \infty$)	Complete pooling ($\tau = 0$)	Hierarchical model (τ estimated)
AIC	$-2\,\mathrm{lpd} = -2 \log p(y\|\hat{\theta}_{\mathrm{mle}})$	54.6	59.4	
	k	8.0	1.0	
	$\mathrm{AIC} = -2\,\widehat{\mathrm{elpd}}_{\mathrm{AIC}}$	70.6	61.4	
DIC	$-2\,\mathrm{lpd} = -2 \log p(y\|\hat{\theta}_{\mathrm{Bayes}})$	54.6	59.4	57.4
	p_{DIC}	8.0	1.0	2.8
	$\mathrm{DIC} = -2\,\widehat{\mathrm{elpd}}_{\mathrm{DIC}}$	70.6	61.4	63.0
WAIC	$-2\,\mathrm{lppd} = -2 \sum_i \log p_{\mathrm{post}}(y_i)$	60.2	59.8	59.2
	$p_{\mathrm{WAIC\,1}}$	2.5	0.6	1.0
	$p_{\mathrm{WAIC\,2}}$	4.0	0.7	1.3
	$\mathrm{WAIC} = -2\,\widehat{\mathrm{elppd}}_{\mathrm{WAIC\,2}}$	68.2	61.2	61.8
LOO-CV	$-2\,\mathrm{lppd}$		59.8	59.2
	$p_{\mathrm{loo-cv}}$		0.5	1.8
	$-2\,\mathrm{lppd}_{\mathrm{loo-cv}}$		60.8	62.8

No pooling prediction is not possible for LOO-CV since we cannot determine the parameter given that the other 7 data points do not inform it

Complete pooling: the fit on “out of sample” is worse by 1: complete pooling model is ”good”

LOO-CV is in a reasonable agreement with other information criteria

Machine Learning

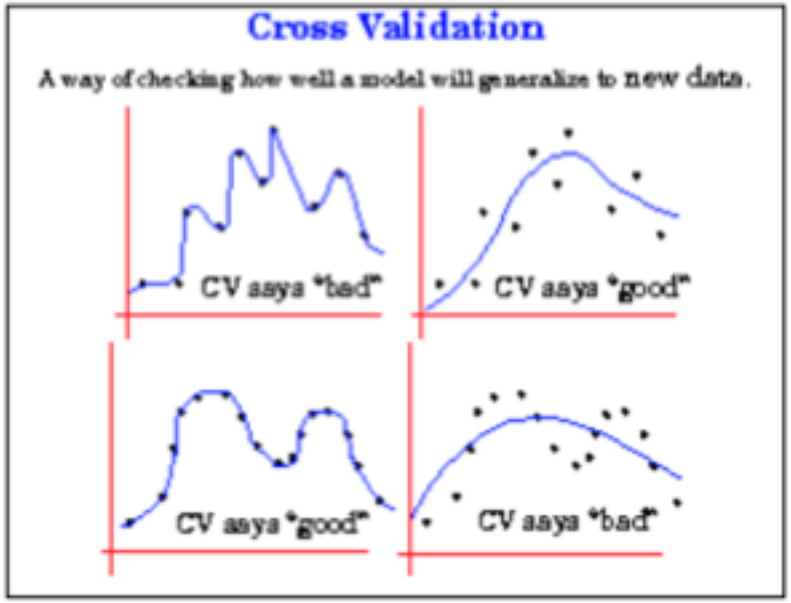
- Machine learning tries to predict true target variables with to be predicted output variables
- input variables (data): input vector, attribute vector, feature vector…
- Typical tasks are regression (output variable is a real number) and classification (output variable is a class: concept learning predicts whether input vector is a member of a class)
- It attempts to do this without a generative model, instead training on the input-output variable pairs (supervised learning)
- Unsupervised learning usually means without pre-classification of variables
- Machine learning is often data analysis without statistics
- We cannot reliably use the concept of posterior or evidence

Cross-validation in ML: Holdout Method

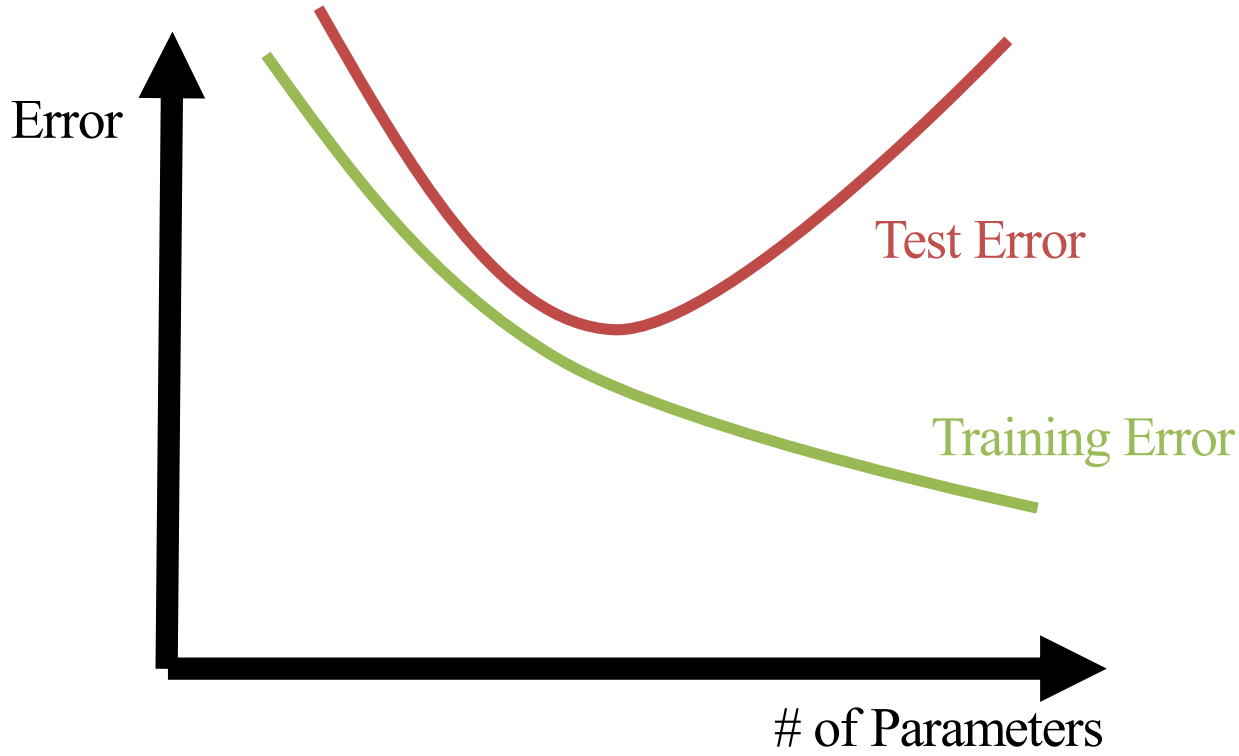
- Supervised learning: we have a set of input/output variables called training set (seen cases)
- We also have another set of input/output variables called testing set (unseen cases)
- We perform the training on the training set and use testing set to determine the error rate (for classification) or error for regression. This is also called off-training error (out of sample) as opposed to training error (in sample).
- Training and testing data must be iid (testing set cannot be correlated with training set)
- Holdout set: a small set used to tune the parameters after training but before testing

Cross-validation in ML: K-fold, leave p out

- Holdout method suffers from randomness of splitting the data into training and testing
- To reduce that we can split the data into K-folds instead of just 1
- If we take K to N we get leave-one-out method
- Or we can leave p out



Cross-validation as a Function of # of Parameters



Bagging or Bootstrap Aggregating

- Instead of using resampling only for cross-validation we may also use it for improving the fits and reduce overfitting
- We can generate bootstrap resamples of the data, proceed with the ML classification or regression for each, then average over them
- Some ML methods such as decision trees (CART: classification and regression trees) have high variance and benefit from bagging
- Same is true for random forest (to be discussed later)

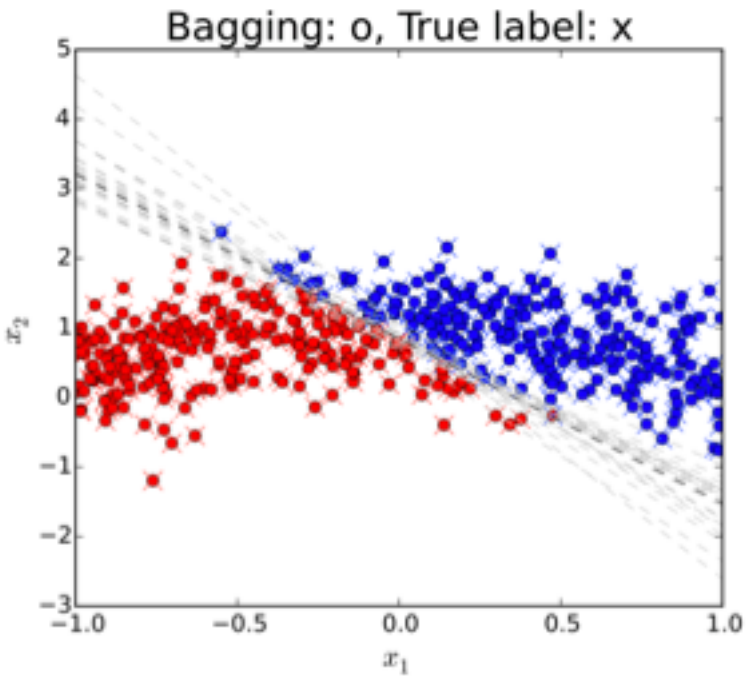


FIG. 29 **Bagging applied to the perceptron learning algorithm (PLA)**. Training data size $n = 500$, number of bootstrap datasets $B = 25$, each contains 50 points. Colors corresponds to different classes while the marker indicates the how these points are labelled: cross for true label and circle for that obtained by bagging. Each gray dashed line indicates the prediction made, based on every bootstrap set while the dark dashed black line is the average of these.

Bad Practices: p-hacking …

- Analyzing data always involves numerous decisions how to cut and slice the data, and what to look for
- For example, we may do many analyses until we find something with $p<0.05$, without devising an underlying causality model. This is p-hacking, also called data fishing or dredging, a posteriori statistics or look-elsewhere effect.
- Or maybe we have a hypothesis we want to prove and we decide on the cuts of the data: if these decisions are correlated with the hypothesis we wish to test this is also an example of p-hacking (which does not have to be conscious)
- p-hacking is a serious problem in many fields, specially soft science where $p=0.05$ is used (less of a problem if we use 5 sigma, i.e. $p=10^{-7}$)
- Publishing only results with $p<0.05$ is called publication bias
- It can go both ways: if we do not want to find an effect we call it confirmation bias: suppose we find an interesting and unexpected effect. We start looking at it and change the selection cuts until it disappears. We are also doing bad science

Good Practices: Blind Analysis

- Best practice is to do blind analysis
- For example, do not look at the real data until you have done all of your selection cuts on simulated data
- In practice this is not always possible: you may discover real data have a systematic that is not present in simulated data
- Alternatives: working with null tests on real data without doing the full analysis. Null tests are tests that the data must pass if you understand their generative model (even if just partial). For example, if we think we understand noise then we can devise data analysis tests that are testing this only. However, what do you do if you fail a null test? You need to go back and understand its origin.
- scramble the parameter outputs in your code in a random fashion so you do not have confirmation bias (or publication bias). Don't be tempted to find out the random seed.

Decision Theory

- How to decide what to do? Optimization over decisions and averaging over uncertainties. We have many different states x, which can be influenced by our action a. We have a payoff $U(x,a)$ when we are in state x and decide action a. We want to maximize it. We need to

  1) enumerate all decisions a and outcomes x (e.g. a are decisions what to purchase, outcome is price)

  2) determine $p(x|a)$

  3) define utility function $U(x,a)$ that maps outcomes into a real number. This could be maximizing net return

  4) compute expected utility $E(U(x)|a) = \int dx U(x,a) p(x|a)$ and maximize it with respect to outcomes a

Decision Theory in in ML: reinforcement learning

- In practice this is a very complicated problem: how to choose $U(x,a)$ and how to update it based on previous actions
- Machines have made a lot of progress in board games and routinely beat people in chess, go …
- This also has a connection to control theory: how to control a robot/car/spacecraft to move along some trajectory given current position
- For more details look MacKay or Gelman

Literature

- *Bayesian Data Analysis,* Gelman et al. , Chapter 6-9
- D. Mackay, *Information Theory, Inference, and Learning Algorithms* (See course website), Chapter 36