Bayesian model selection and MCMC

DAWGI guest lecture 29/01/2021

Peter Scicluna

Why do we need model selection?

Yesterday: Given a model and some data, what part of its parameter space is most likely?

Why do we need model selection?

Yesterday: Given a model and some data, what part of its parameter space is most likely?

But the other aspect of inference is identifying which model describes the observations the best.

Why do we need model selection?

Yesterday: Given a model and some data, what part of its parameter space is most likely?

But the other aspect of inference is identifying which model describes the observations the best.

This is the *model selection* problem. This is the key to deciding between different theoretical frameworks given uncertain observations.

Hypothesis testing: Frequentist approaches and issues

Various hypothesis testing approaches:

Z-test and *t*-test - usually the first tests we learn. Simple ways to test if means of two samples are different, or if mean of sample is different from a model. Z-test requires that the variance of the population is known, t-test is more general. But samples must be drawn from a Gaussian population. Set a p-value threshold to reject null hypothesis.

Hypothesis testing: Frequentist approaches and issues

Various hypothesis testing approaches:

Z-test and *t*-test - usually the first tests we learn. Simple ways to test if means of two samples are different, or if mean of sample is different from a model. Z-test requires that the variance of the population is known, t-test is more general. But samples must be drawn from a Gaussian population. Set a p-value threshold to reject null hypothesis.

Likelihood-ratio test - more general, should we reject the null model given the maxima of the likelihoods of the two hypotheses? Under certain assumptions, is the test with the highest *statistical power*. Compute ratio of the maximum likelihoods, if it is less than a threshold, reject the null hypothesis.

$$\log R_L = 2\left(\log \hat{L}_0 - \log \hat{L}_1\right)$$

Hypothesis testing: Frequentist approaches and issues

Various hypothesis testing approaches:

Z-test and *t*-test - usually the first tests we learn. Simple ways to test if means of two samples are different, or if mean of sample is different from a model. Z-test requires that the variance of the population is known, t-test is more general. But samples must be drawn from a Gaussian population. Set a p-value threshold to reject null hypothesis.

Likelihood-ratio test - more general, should we reject the null model given the maxima of the likelihoods of the two hypotheses? Under certain assumptions, is the test with the highest *statistical power*. Compute ratio of the maximum likelihoods, if it is less than a threshold, reject the null hypothesis.

Information criteria (AIC, BIC, WAIC …) - information-theoretic or entropy-based approaches instead. Use maximum likelihood and some information about the model to calculate statistic for each model (lower is better). Difference in statistic used to estimate relative probability of information loss, while penalising more complex models.

$$AIC = 2k - 2\log \hat{L} \qquad BIC = k\log n - 2\log \hat{L}$$

Hypothesis testing: Frequentist approaches and issues

Various hypothesis testing approaches:

Z-test and *t*-test - usually the first tests we learn. Simple ways to test if means of two samples are different, or if mean of sample is different from a model. Z-test requires that the variance of the population is known, t-test is more general. But samples must be drawn from a Gaussian population. Set a p-value threshold to reject null hypothesis.

Likelihood-ratio test - more general, should we reject the null model given the maxima of the likelihoods of the two hypotheses? Under certain assumptions, is the test with the highest *statistical power*. Compute ratio of the maximum likelihoods, if it is less than a threshold, reject the null hypothesis.

Information criteria (AIC, BIC, WAIC …) - information-theoretic or entropy-based approaches instead. Use maximum likelihood and some information about the model to calculate statistic for each model (lower is better). Difference in statistic used to estimate relative probability of information loss, while penalising more complex models.

And many more …

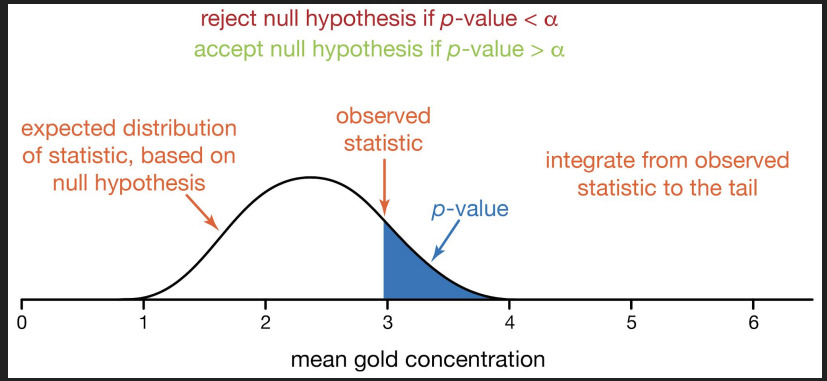
Hypothesis testing: some issues

- *Nested models* often required
  - is constrained version of model superior? (e.g. is $\Omega_\Lambda > 0.7$?)
  - What if we want to compare ΛCDM with MOND?

Hypothesis testing: some issues

- *Nested models* often required
  - is constrained version of model superior? (e.g. is $\Omega_\Lambda > 0.7$?)
  - What if we want to compare ΛCDM with MOND?
- Often misinterpreted
  - p-value is *not* the probability of the Null, it is the probability of getting the same result by chance (i.e. false-positive probability)!
  - Statistic is *uncertain*. This should be propagated into the test.

Figure by Steve Holland

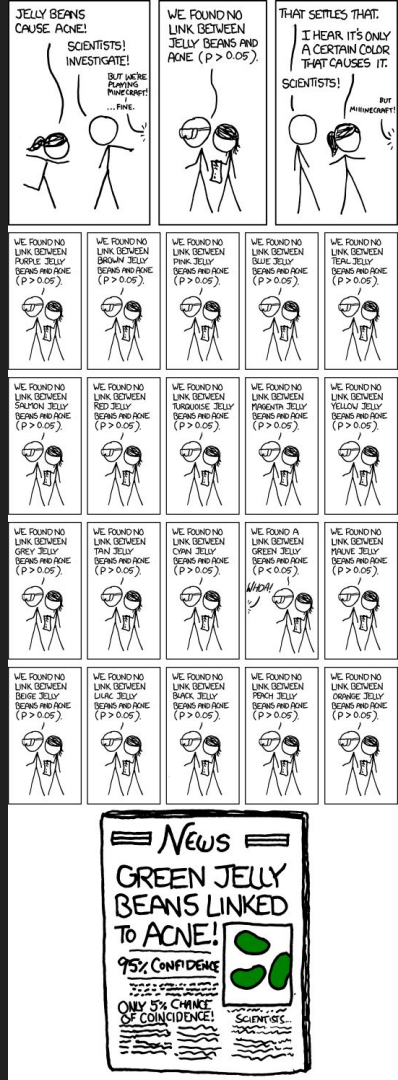


Hypothesis testing: some issues

- *Nested models* often required
  - is constrained version of model superior? (e.g. is $\Omega_\Lambda > 0.7$?)
  - What if we want to compare ΛCDM with MOND?
- Often misinterpreted
  - p-value is *not* the probability of the Null, it is the probability of getting the same result by chance (i.e. false-positive probability)!
  - Statistic is *uncertain*. This should be propagated into the test.
- Null hypothesis is somehow special
  - It must be rejected

Hypothesis testing: some issues

- *Nested models* often required
  - is constrained version of model superior? (e.g. is $\Omega_\Lambda > 0.7$?)
  - But what if we want to compare ΛCDM with MOND?
- Often misinterpreted
  - p-value is *not* the probability of the Null, it is the probability of getting the same result by chance (i.e. false-positive probability)!
  - Statistic is *uncertain*. This should be propagated into the test.
- Null hypothesis is somehow special
  - It must be rejected
- *Look-elsewhere effect*
  - If we set a 5% FPP and do 20 tests, we expect 1 false positive on average.



Looking for a Needle in a Haystack? Look Elsewhere!

A statistical comparison of approximate global p-values.

Sara Algeri [a,1,2], David A. van Dyk [b,1], Jan Conrad [c,2,3,4,1], Brandon Anderson [d,2,3]

[1]Department of Mathematics, Imperial College London, South Kensington Campus, London SW7 2AZ, United Kingdom
[2]Department of Physics, Stockholm University, AlbaNova, SE-106 91 Stockholm, Sweden
[3]The Oskar Klein Centre for Cosmoparticle Physics, AlbaNova, SE-106 91 Stockholm, Sweden
[4]Wallenberg Academy Fellow

Hypothesis testing: some issues

- *Nested models* often required
  - is constrained version of model superior? (e.g. is $\Omega_\Lambda > 0.7$?)
  - What if we want to compare ΛCDM with MOND?
- Often misinterpreted
  - p-value is *not* the probability of the Null, it is the probability of getting the same result by chance (i.e. false-positive probability)!
  - Statistic is *uncertain*. This should be propagated into the test.
- Null hypothesis is somehow special
  - It must be rejected
- *Look-elsewhere effect*
  - If we set a 5% FPP and do 20 tests, we expect 1 false positive on average.
- Penalties for complexity must be included by hand
  - Occam’s razor is intuitive, but only IC statistics include a penalty and those depend on specific assumptions

Hypothesis testing: some issues

- *Nested models* often required
  - is constrained version of model superior? (e.g. is $\Omega_\Lambda > 0.7$?)
  - What if we want to compare ΛCDM with MOND?
- Often misinterpreted
  - p-value is *not* the probability of the Null, it is the probability of getting the same result by chance (i.e. false-positive probability)!
  - Statistic is *uncertain*. This should be propagated into the test.
- Null hypothesis is somehow special
  - It must be rejected
- *Look-elsewhere effect*
  - If we set a 5% FPP and do 20 tests, we expect 1 false positive on average.
- Penalties for complexity must be included by hand
  - Occam's razor is intuitive, but only IC statistics include a penalty and those depend on specific assumptions
- No penalties for fine turing
  - Only look at maximum likelihood, but intuitively models which only work well in a small region of parameter space are bad

Hypothesis testing: a Bayesian perspective

Key question: Which of a set of models is best able to describe our data?

Requirements:

- Robust against multiple comparisons
- Provide probability (or *odds*) of each model compared to each other
    - Note: symmetrical i.e. does not ‘privilege’ one model as null!
- Suitable for nested models but also more general
- Naturally penalises complexity and fine-tuning

Information-theoretic approaches do deal with many of these, but there is a better approach …

The Bayes factor

Described by Kass & Raftery (1995, see right) based on Jeffreys (1931, 1961).

N.B.: $P\left(D|H_k\right) = \int P\left(D|\theta_k, H_k\right) P\left(\theta_k|H_k\right) d\theta_k$

So the Bayes factor is the *ratio of the **integral** of the posteriors!* This is also the *ratio of the **evidences!***

Prior odds: ratio of probability of the two models. Hence, if the two models have equal probability (prior odds = 1) then the Bayes factor = posterior odds.

When prior is a delta function, the Bayes factor reduces to the likelihood ratio

3.1 Definition

We begin with data $\mathbf{D}$, assumed to have arisen under one of the two hypotheses H_1 and H_2 according to a probability density $\mathrm{pr}(\mathbf{D}|H_1)$ or $\mathrm{pr}(\mathbf{D}|H_2)$. Given a priori probabilities $\mathrm{pr}(H_1)$ and $\mathrm{pr}(H_2) = 1 - \mathrm{pr}(H_1)$, the data produce a posteriori probabilities $\mathrm{pr}(H_1|\mathbf{D})$ and $\mathrm{pr}(H_2|\mathbf{D}) = 1 - \mathrm{pr}(H_1|\mathbf{D})$. Because any prior opinion gets transformed to a posterior opinion through consideration of the data, the transformation itself represents the evidence provided by the data. In fact, the same transformation is used to obtain the posterior probability, regardless of the prior probability. Once we convert to the odds scale (odds = probability/(1 − probability)), the transformation takes a simple form. From Bayes's theorem, we obtain

$$\mathrm{pr}(H_k|\mathbf{D}) = \frac{\mathrm{pr}(\mathbf{D}|H_k)\mathrm{pr}(H_k)}{\mathrm{pr}(\mathbf{D}|H_1)\mathrm{pr}(H_1) + \mathrm{pr}(\mathbf{D}|H_2)\mathrm{pr}(H_2)} \qquad (k = 1, 2),$$

so that

$$\frac{\mathrm{pr}(H_1|\mathbf{D})}{\mathrm{pr}(H_2|\mathbf{D})} = \frac{\mathrm{pr}(\mathbf{D}|H_1)}{\mathrm{pr}(\mathbf{D}|H_2)} \frac{\mathrm{pr}(H_1)}{\mathrm{pr}(H_2)},$$

and the transformation is simply multiplication by

$$B_{12} = \frac{\mathrm{pr}(\mathbf{D}|H_1)}{\mathrm{pr}(\mathbf{D}|H_2)}, \qquad (1)$$

which is the *Bayes factor*. Thus, in words,

posterior odds = Bayes factor × prior odds,

and the Bayes factor is the ratio of the posterior odds of H_1 to its prior odds, regardless of the value of the prior odds.

The Bayes' factor

Advantages:

- Makes assumptions (like prior odds) explicit
- Integration over more parameters expands prior volume
  - Natural penalty for more parameters
  - Models which only work well in a small region will also have lower evidence
- Works fine when $\theta_1 \subset \theta_2$ i.e. nested models but also when they're not!
- Can make any number of comparisons between different models after calculating their evidences, as long as we handle uncertainty
- **Can express support for either hypothesis!**

The Bayes' factor

Issues:

- Requires integration over *entire* prior volume for robust estimate of evidence
  - This is both difficult and time consuming
- Choice of *prior odds* allows choice of winning model
- Also sensitive to the priors themselves
  - Adding new regions to search that don't contain any probability dilutes the evidence

Computing Bayes factors

Key problem: integration is hard!

If we have samples from the posterior, we could integrate over them to get evidence

- Most MCMC implementations aren’t very good at exploring the full prior volume

Need a method that gives shape of posterior and is guaranteed to explore whole volume

- Nested Sampling (Skilling 2004) is designed to do exactly this.
  - Samples parameter vectors from the prior and evaluates the likelihood
  - Gradually shrinks volume searched, by creating nested iso-likelihood contours
  - Evidence estimate is a natural outcome.

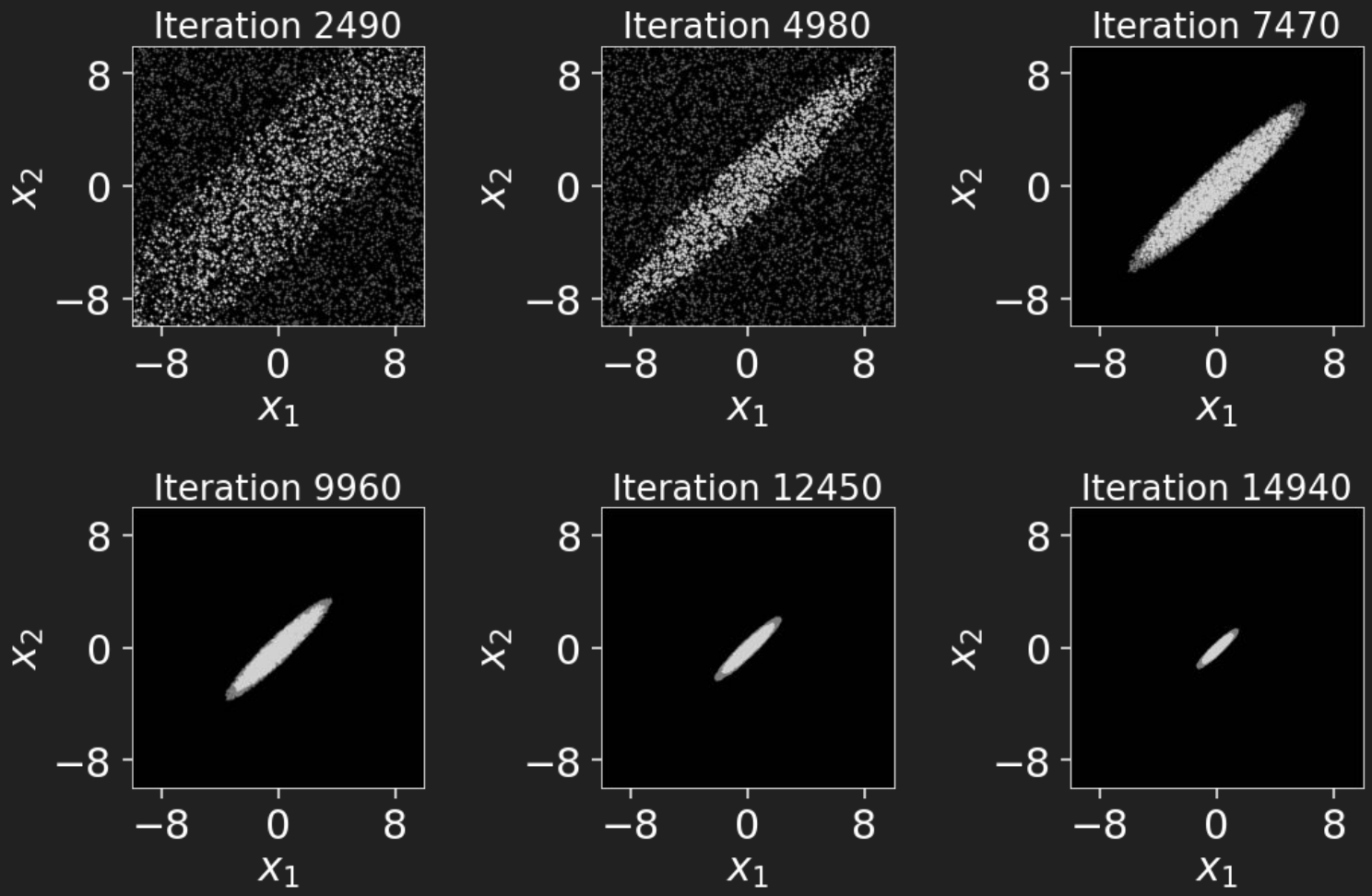


Image by Josh Speagle

Computing evidence with nested sampling

Important aspect of NS is ensuring that the prior is proper.

Typically achieved through a function that transforms from the unit hypercube to the prior, known as *prior transform*.

Simple example from dynesty docs →

The transformed points are then fed to a likelihood function (similar to emcee) so the likelihood can be evaluated.

Example: Uniform Priors

Suppose we want our prior to be Uniform from [-10, 10) for all variables:

$$p(x) \propto \begin{cases} 1 & -10 \leq x < 10 \\ 0 & \text{otherwise} \end{cases}$$

The prior transform for this distribution would be:

```
def prior_transform(u):
    """Transforms the uniform random variable `u ~ Unif[0., 1.)`
    to the parameter of interest `x ~ Unif[-10., 10.)`."""

    x = 2. * u - 1.  # scale and shift to [-1., 1.)
    x *= 10.  # scale to [-10., 10.)

    return x
```

Computing evidence with nested sampling

A number of other important choices to make

- How to sample from the prior
- How to bound the prior volume
- How many *live points* to use

As results can be very sensitive to these.

Nested sampling also has a few handy byproducts, like samples from the posterior, so we can also do all the same inference as with MCMC

However, it is typically slower because it explores entire prior volume.

Decisions with Bayes Factors

Once you have computed the evidence for your models, you can compute the Bayes factor and the posterior odds.

For Bayes factor $B_{12} = \frac{P(D|H_1)}{P(D|H_2)}$ we must decide what constitutes a significant result. It is common to use tables similar to that from Jeffreys (1961, see right). Remember, it can express support for either hypothesis!

Assuming all models are equally likely, any number of models can be compared and the one with the highest evidence is preferred. This also indicates which models are effectively indistinguishable.

APPENDIX B

TABLES OF K

We have defined $$K = \frac{P(q \mid \theta H)}{P(q' \mid \theta H)},$$

where q is the null hypothesis, q' the alternative, H the previous information, and θ the observational evidence. We take the standard case where q and q' are equally probable given H. In most of our problems we have asymptotic approximations to K when the number of observations is large. We do not need K with much accuracy. Its importance is that if $K > 1$ the null hypothesis is supported by the evidence; if K is much less than 1 the null hypothesis may be rejected. But K is not a physical magnitude. Its function is to grade the decisiveness of the evidence. It makes little difference to the null hypothesis whether the odds are 10 to 1 or 100 to 1 against it, and in practice no difference at all whether they are 10^4 or 10^{10} to 1 against it. In any case whatever alternative is most strongly supported will be set up as the hypothesis for use until further notice. The tables give values of χ^2, t, or z for $K = 1, 10^{-1/2}, 10^{-1}, 10^{-3/2}, 10^{-2}$. The last will be regarded as a limit for unconditional rejection of the null hypothesis. $K = 10^{-1/2}$ represents only about 3 to 1 odds, and would be hardly worth mentioning in support of a new discovery. It is at $K = 10^{-1}$ and less that we can have strong confidence that a result will survive future investigation. We may group the values into grades, as follows.

Grade 0.	$K > 1$.	Null hypothesis supported.
Grade 1.	$1 > K > 10^{-1/2}$.	Evidence against q, but not worth more than a bare mention.
Grade 2.	$10^{-1/2} > K > 10^{-1}$.	Evidence against q substantial.
Grade 3.	$10^{-1} > K > 10^{-3/2}$.	Evidence against q strong.
Grade 4.	$10^{-3/2} > K > 10^{-2}$.	Evidence against q very strong.
Grade 5.	$10^{-2} > K$.	Evidence against q decisive.

Bayes factors in action

Table 5. Model selection results for 4U 1820-30.

Model	log B	ΔAIC	ΔBIC
ism	0.0	0	0
ism + photo	+1.3	+19	+40
ism + ism	+25.5	+156	+204
photo	+77.3	+351	+363
photo + photo	+90.1	+389	+421

Notes. The last three columns show the model comparison based on log-evidence, AIC and BIC. The log-evidence is normalised to the maximum value found, whereas both AIC and BIC are normalised to the minimum value which indicates the preferred model. Models with log $B > 1.5$ or $\Delta AIC(\Delta BIC) > 10$ can be ruled out as a plausible model that generates the data (Jeffreys 1961; Burnham & Anderson 2002).

Rogantini et al (2021)

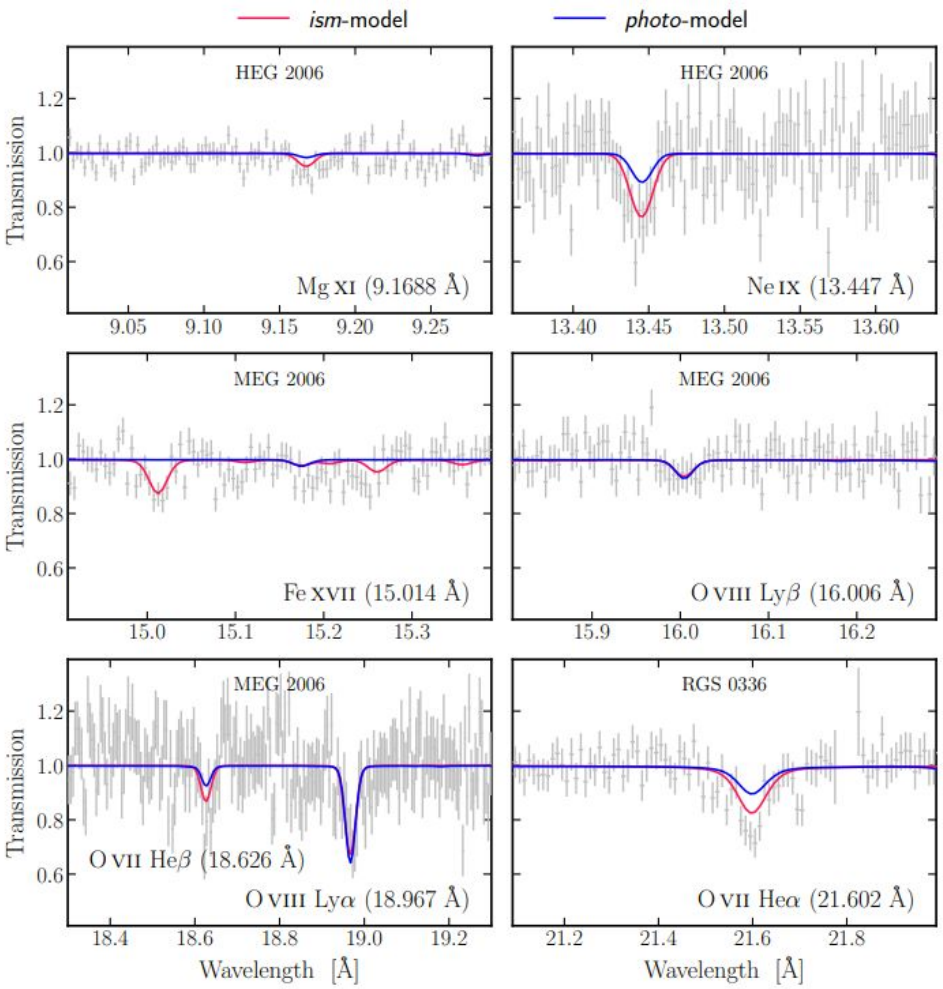


Fig. 2. High ionisation lines detected in the spectra of 4U 1820-30. Here, the spectrum is shown in unit of transmittance: the observed counts are divided by the underlying continuum together with the cold and warm absorption. For clarity, we do not display all the datasets. We superimpose the *ism* and *photo* models (in red and blue, respectively) obtained with the Bayesian parameter inference.

Mixture models

Important note: Mixture models are ***not*** a model-selection method, they do ***classification!***

BUT this is often what astronomers mean when they say they want to choose between two models.

Basic idea: Source could be one of two classes, can we tell which from the data?

Try modelling data as *superposition* of models for both classes $P\left(D_i|\theta,\pi_k\right) = \sum_{k=1}^{2} P\left(D_i|z_i=k,\theta\right) P\left(z_i=k|\pi_k\right)$

z_i is the *latent variable* indicating which class it belongs to (which model describes the data the best)

This generalises trivially to many classes

Functionally equivalent to clustering for unsupervised classification

Mixture models

- Each model in the mixture is known as a *component*
  - The mixture is said to have *K* components
- Each data point has a *latent variable z* associated with it indicating which component it belongs to.
- Each component is associated with a *mixture weight* π_k
  - $\sum_{k=1}^{K} \pi_k = 1$
  - Basically prior on *z*
- Prior on mixture weights: Dirichlet distribution
  - Weights lie on a *simplex*
- Each component still has their respective parameters
  - Total number of parameters = sum of parameters for each component + *N* latent variables + *K* mixture weights
- With enough components can reproduce arbitrary distributions
- Can also be thought of as *compositional models*

Mixture models

How does this work?

- If we have *K* components, as stated the priors on the components are π_k, with $\sum_{k=1}^{K} \pi_k = 1$
- It therefore follows that the prior probability that an observation is from component *k* is
  - $P(z_i = k|\theta, \pi_k) = P(z_i = k|\pi_k) = \pi_k$
- While the likelihood of an observation conditioned on being drawn from component *k* is
  - $P(D_i|z_i = k, \theta)$
- And marginalising over the classes, this becomes
  - $P(D_i|\theta, \pi_k) = \sum_{k=1}^{K} P(D_i|z_i = k, \theta) P(z_i = k|\pi_k)$
- Given this likelihood, one could in principle infer the distribution of z_i to classify this observation
  - $P(z_i, \theta|D_i, \pi_k) \propto P(\theta, \pi_k) \sum_{k=1}^{K} P(D_i|z_i = k, \theta) P(z_i = k|\pi_k)$
- However, we would need to know π_k very well, not necessarily informative
- Real power comes with *hierarchical inference* (see next lecture by B. Johnson) to learn about a sample
  - Infer π_k, z_i and θ simultaneously

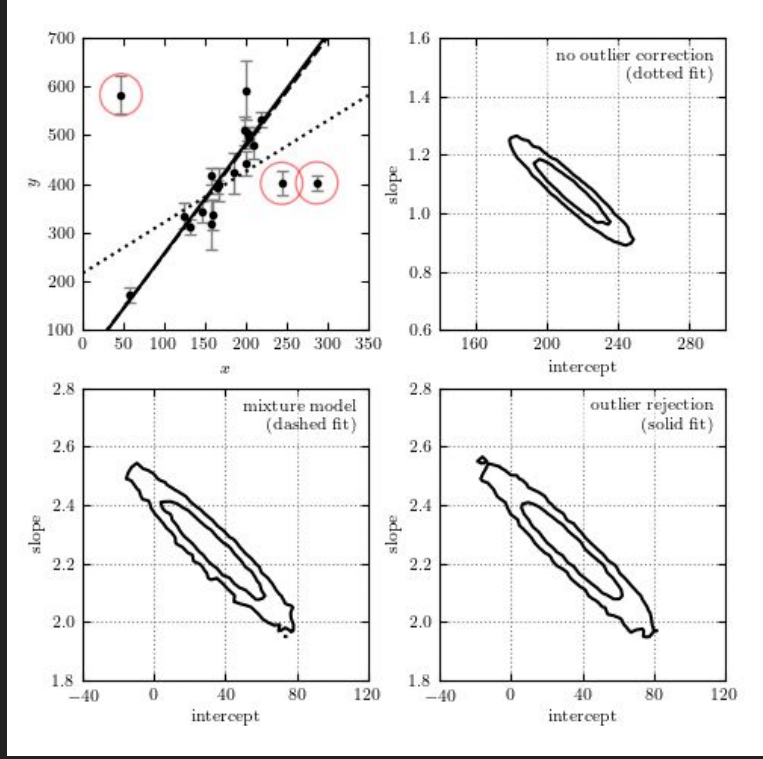
Mixture models - a classic example

Given a dataset, what is the probability that some points are outliers?

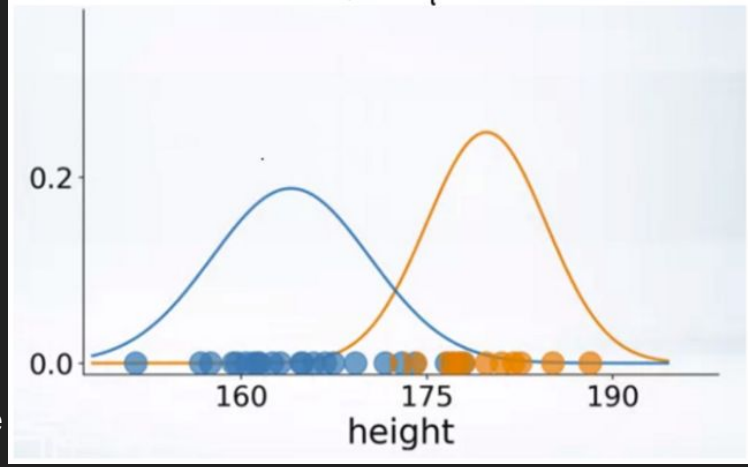
Two classes: good points and bad points

$$p(\{y_i\}|\{x_i\},\{\sigma_i\},\{g_i\},\theta_0,\theta_1,\mu_b,V_b) \propto \prod_{i=1}^{N}\left[\frac{g_i}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{(y_i-\theta_1 x_i-\theta_0)^2}{2\sigma_i^2}\right)\right.$$
$$\left.+\frac{1-g_i}{\sqrt{2\pi(V_b+\sigma_i^2)}}\exp\left(-\frac{(y_i-\mu_b)^2}{2(V_b+\sigma_i^2)}\right)\right]. \quad (8.68)$$

Covered in detail by AstroML and Hogg+ 2010.



Gaussian Mixture models



What if we want to classify sources in colour-colour space, but don't know how many classes to expect or what they represent?

- We expect similar sources to be grouped together i.e. clustering
  - Simple assumption - classes might be described by multivariate Gaussians in colour-colour space
- So we have a model of K 2D Gaussians
  - Mixture weights, means and variances are our free parameters for these
- Plus each object has its latent variable for classification
- So total number of variables = N (objects) + K (mixture weights) + 4K (means and variances)
- Optimise all of this simultaneously

$$P\left(z,\theta|D,\pi_k\right) \propto P\left(\theta,\pi_k\right) \prod_{i=1}^{N} \sum_{k=1}^{K} P\left(D_i|z_i=k,\theta\right) P\left(z_i=k|\pi_k\right)$$

- Outcome: probability of each class for each object, parameters of Gaussians, mixture weights

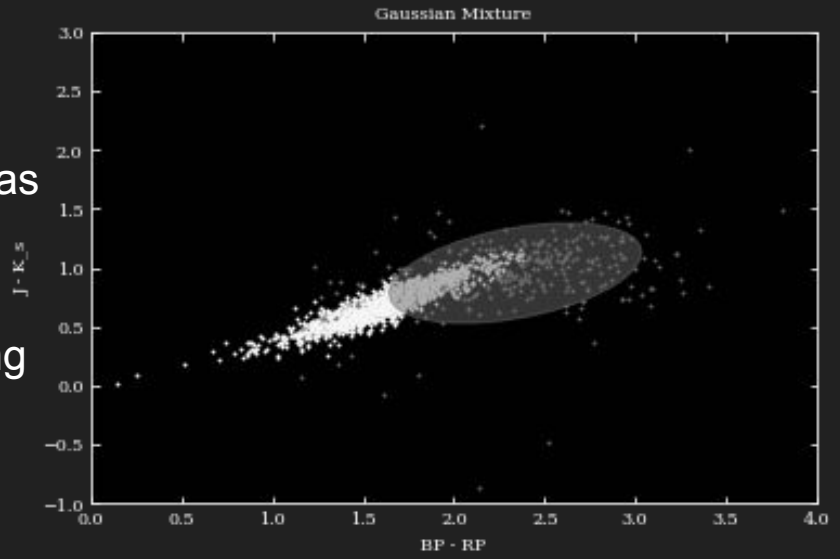
Gaussian Mixture models

Lots of python implementations exist

- sklearn has a well-featured starting point, but has limitations
- AutoGMM (part of Graspy)
- pyGMMis (handles noisy data properly, including correlated noise)
- …

And it’s not too hard to put it together yourself.

Important question - how many components do you need?



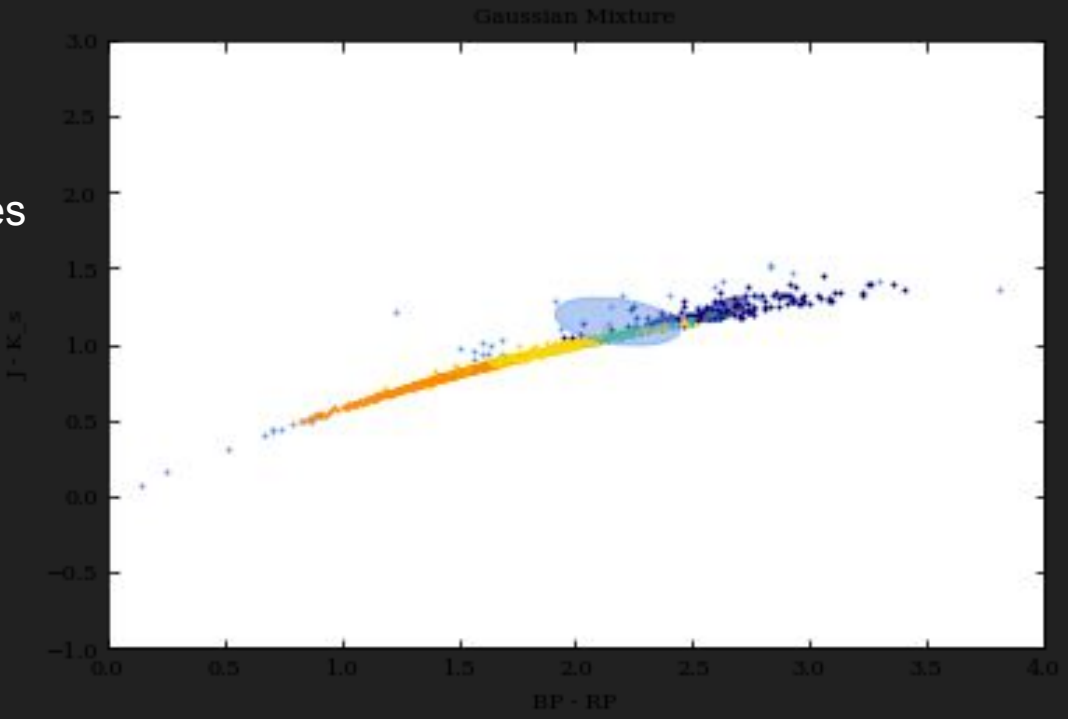
Gaussian Mixture models - how many components?

This is a model-selection problem!

Obvious strategy - keep adding more components until it no longer improves the fit

How do you measure that? *Bayes factors!*

However, these can be very complex models, so AIC or BIC might be good enough for many cases.



GMMs for population analysis

Rojas-Arriagada
2016

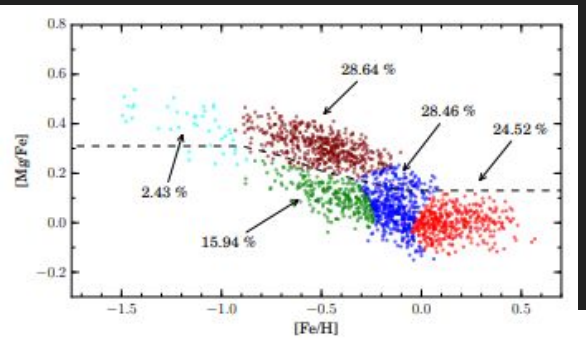


Fig. 2. GMM best model for the GES data distribution in the [Mg/Fe] vs. [Fe/H] plane. The colors highlight the different data families corresponding to individual modes of the Gaussian mixture. Stars are classified according to the highest component-associated responsibilities. Three black dashed line segments displays the same division into thin and thick disk sequences as in Fig. 1. We quote the percentage of the sample that each component encompasses.

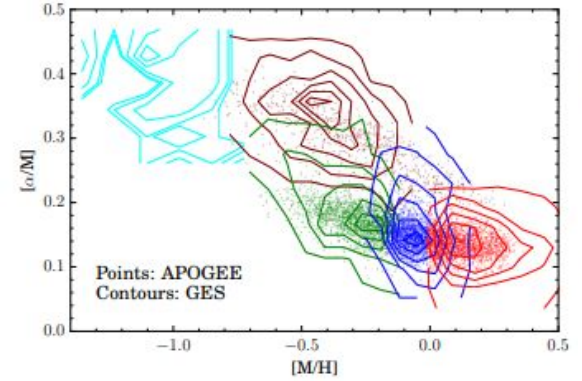


Fig. 3. GMM decomposition of *Gaia*-ESO survey and APOGEE samples in the [α/M] vs. [M/H] plane. Points depict the APOGEE working sample. Contour lines draw the density distribution of *Gaia*-ESO survey GMM data groups. A vertical shift of Δ[α/M] = 0.1 dex was applied to the APOGEE sample to obtain a better agreement between the two data sets. Color coding is set to be consistent with the one used to represent data groups throughout this paper.

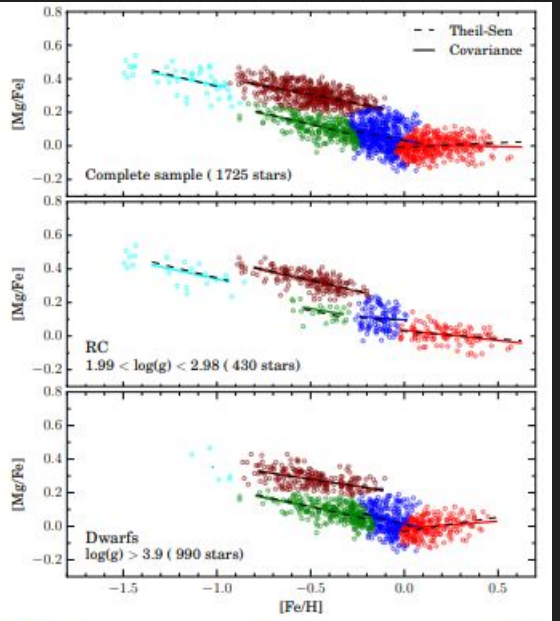


Fig. 5. GMM decomposition performed on the complete GES sample (*upper panel*), RC stars (*middle panel*), and dwarfs (*lower panel*). Solid black lines represent trend lines determined for each GMM group from the mean and covariance of the respective mode of the mixture. Dashed black lines stand for trends computed directly from the datapoints, using the Theil-Sen estimator.

It doesn't have to be colours, of course!

Successfully used to separate stars based on abundance patterns

Same clusters seen in different sub-populations

Key take-home messages

- Model selection is important, but *hard*
  - Many options, all with flaws.
- Bayesian approach uses the Bayes factor
  - Also known as evidence ratio
  - Enables multiple comparisons and naturally penalises complexity
  - Depends on choice of prior, however
  - Typically computed using Nested Sampling
- If what you want is *classification*, Mixture Models can be powerful
  - Discussed a simple implementation Gaussian Mixture Models, used for clustering
  - Depends on choice of number of components, which is another model-selection problem.