The high-dimensional and multi-modal Bayesian inference code

DIAMONDS

Application to Asteroseismology

Enrico Corsaro

enrico.corsaro@cea.fr
CEA Saclay

Bayesian Statistics

Bayes Theorem

$$\mathbf{D} = \{d_1, d_2, \ldots, d_m\}$$

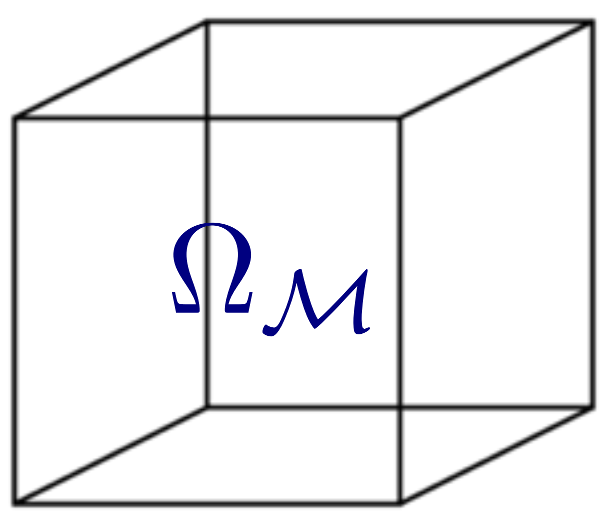
Dataset (observations)

$$\mathcal{M} = \mathcal{M}(\boldsymbol{\theta})$$

Model to be tested

$$\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$$

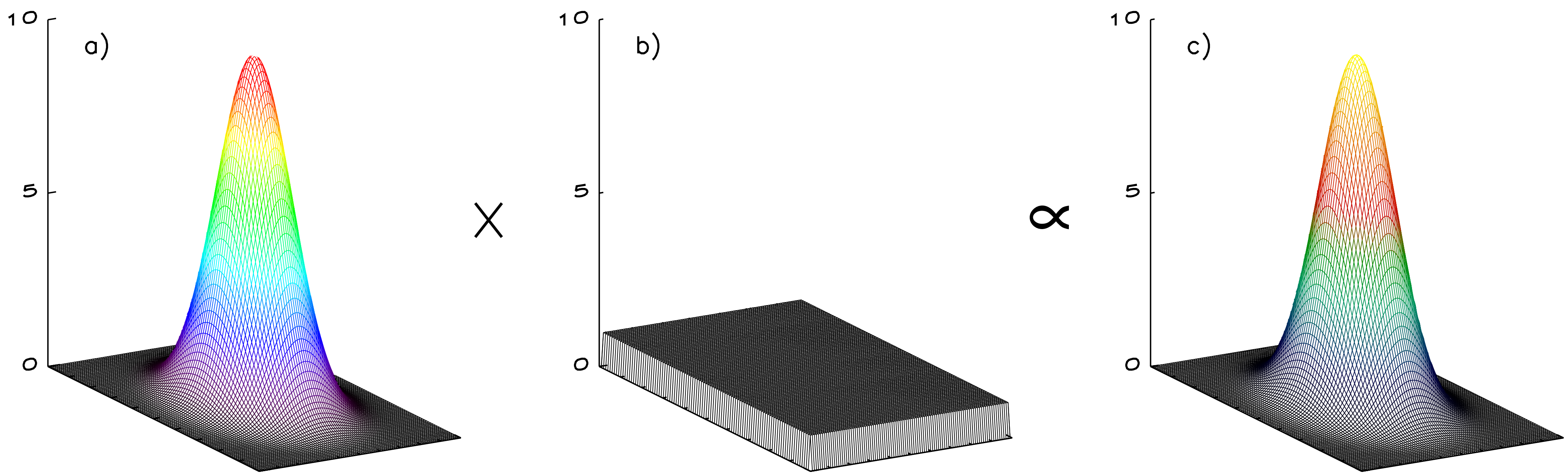
k free parameters (parameter vector)



k-dimensional **parameter space** defined by the free parameters

Bayes Theorem

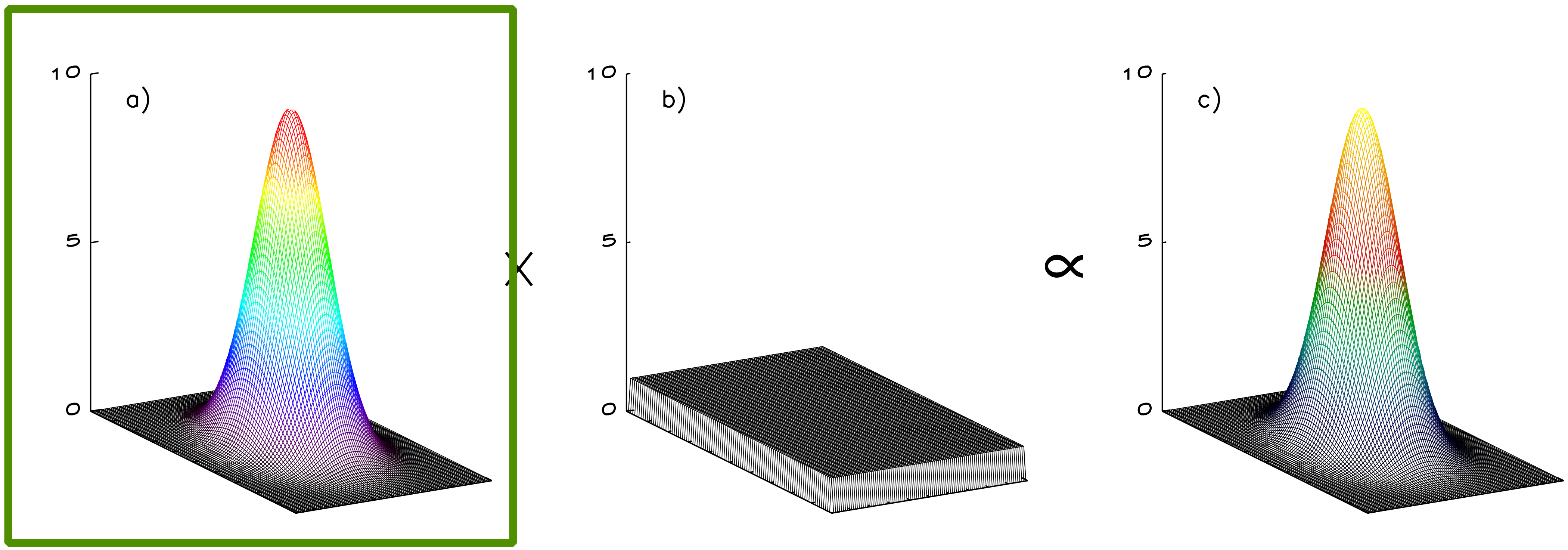
Likelihood **Prior** **Posterior**



$$p\left(\boldsymbol{\theta}\right) = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\mathcal{E}}$$

Bayes Theorem

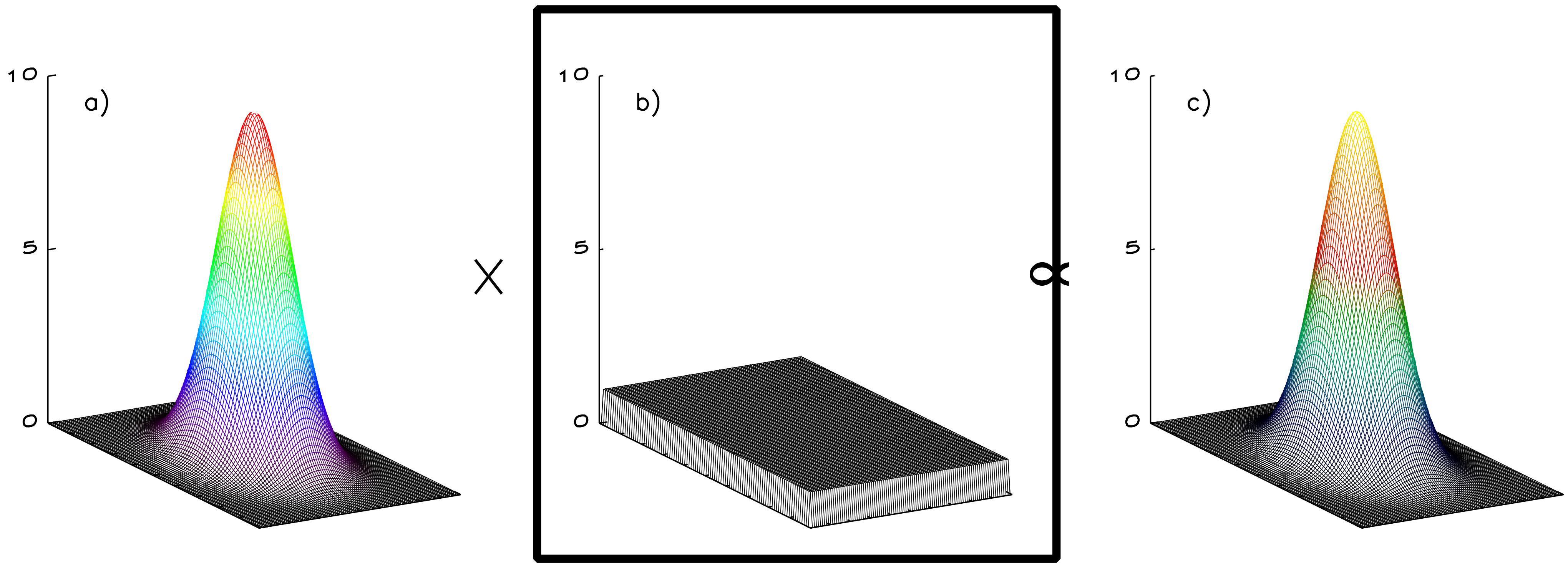
Likelihood **Prior** **Posterior**



$$p\left(\boldsymbol{\theta}\right) = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\mathcal{E}}$$

Bayes Theorem

Likelihood **Prior** **Posterior**



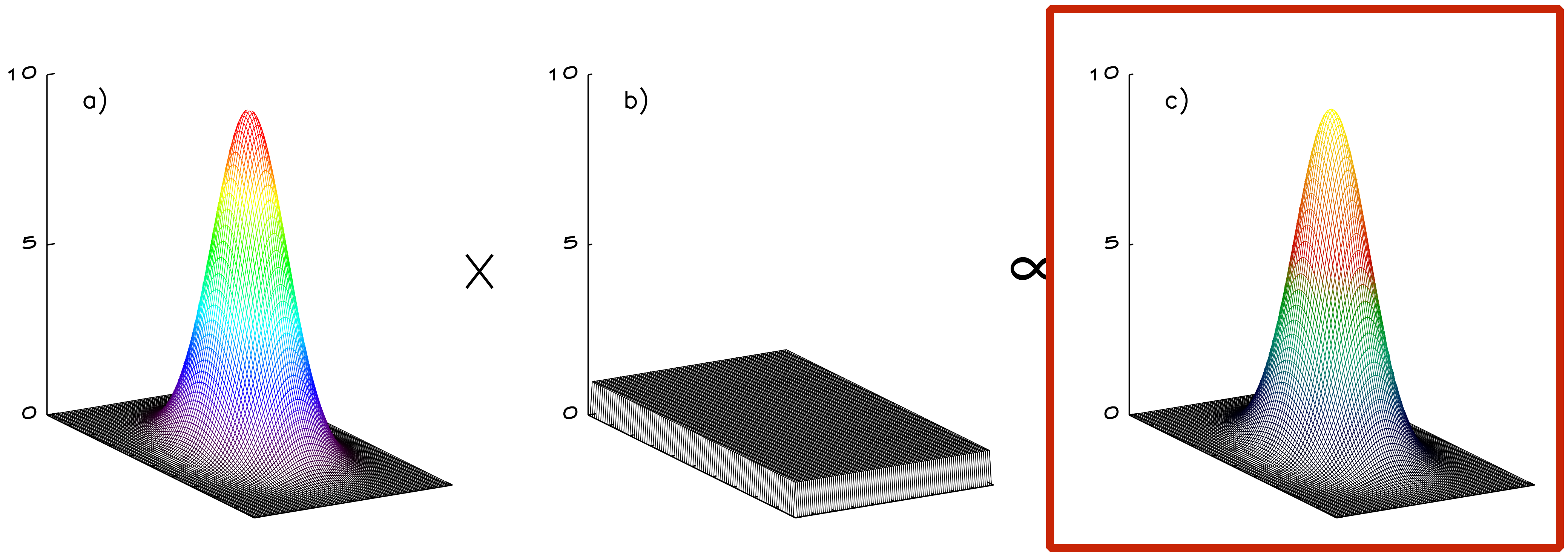
$$p\left(\boldsymbol{\theta}\right) = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\mathcal{E}}$$

Probability density function (PDF)

$$\int \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$$

Bayes Theorem

Likelihood **Prior** **Posterior**



$$p\left(\boldsymbol{\theta}\right) = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\mathcal{E}}$$

Probability density function (PDF)

$$\int p(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$$

Bayesian Inference

- The Bayesian inference of a dataset is divided in two problems:
  - **Parameter Estimation**
    Allows to obtain the estimates of all the free parameters and the corresponding error bars
  - **Model comparison**
    Provides a way to select the best model to represent the observations among different possible ones

Parameter estimation

- k-dimensional parameter space

$$\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$$ k free parameters (parameter vector)

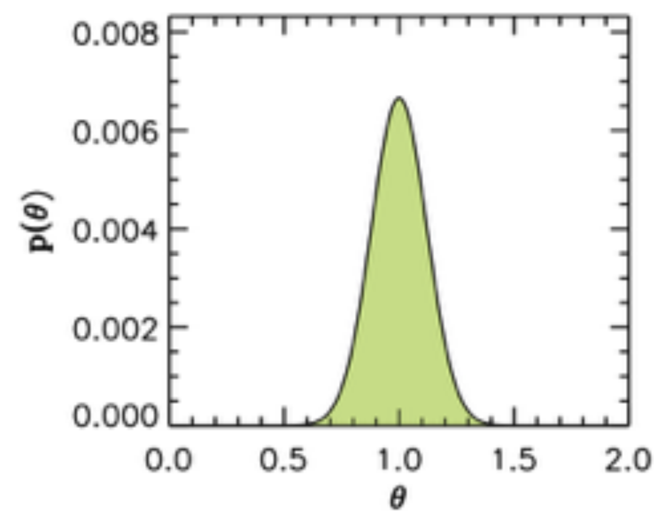
- The posterior PDF is a function of k parameters

$$p(\theta_1) = \int p(\boldsymbol{\theta})\, d\theta_2 \ldots d\theta_k$$

θ_1

Mean
Mode
Median
Variance
Credible Intervals

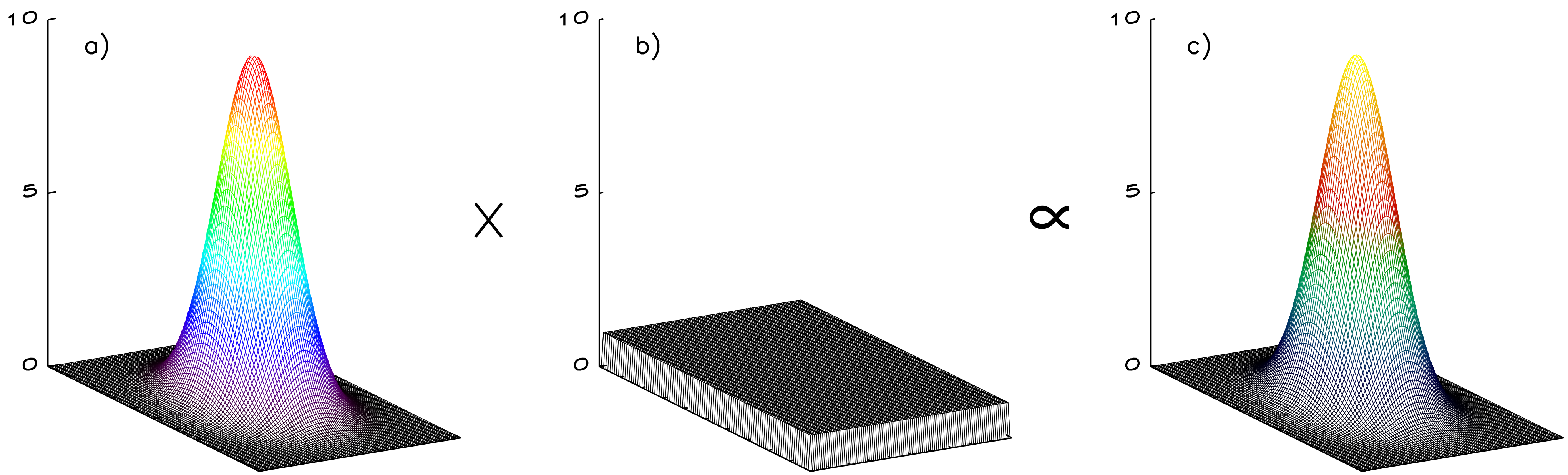
Marginal PDF



- To obtain the PDF of a single parameter we can marginalize the posterior PDF

Model comparison

Likelihood **Prior** **Posterior**

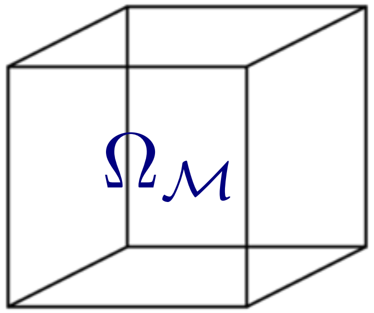


$$p\left(\boldsymbol{\theta}\right) = \frac{\mathcal{L}\left(\boldsymbol{\theta}\right)\pi\left(\boldsymbol{\theta}\right)}{\mathcal{E}}$$

Model comparison

- **Bayesian Evidence** is an a-dimensional quantity given as a k-dimensional integral over the entire parameter space (does not exist in frequentist approach!)

$$\mathcal{E} = \int \mathcal{L}(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

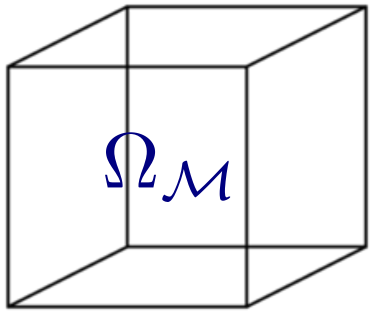


$$p(\boldsymbol{\theta}) = \frac{\mathcal{L}(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\mathcal{E}}$$

Model comparison

- **Bayesian Evidence** is an a-dimensional quantity given as a k-dimensional integral over the entire parameter space (does not exist in frequentist approach!)

$$\mathcal{E} = \int \mathcal{L}\left(\boldsymbol{\theta}\right) \pi \left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}$$



$\mathcal{E}$ ↑

Fit quality ↑

free parameters ↓

<u>WEIGHT</u>: simple models are preferred (Occam's razor)

Problems

- For **k > 3** no more analytical solutions to the marginalization problem (hence also the computation of the Bayesian Evidence integral)
- Numerical integration needed but for higher dimensions (**k ~ 20**) is not enough (too approximated)
- Numerical sampling techniques (e.g. **Monte Carlo**) are approximate by definition, so lot of samples are required.
- Sampling algorithm can get stuck into a local maximum and never be able to explore all the parameter space (e.g. Eggbox). Lot of ad-hoc improvements required, depending on the application.
- Computational time and number of samples to be used can be a real problem. Big limitations to complex fitting problems.

The basic algorithm

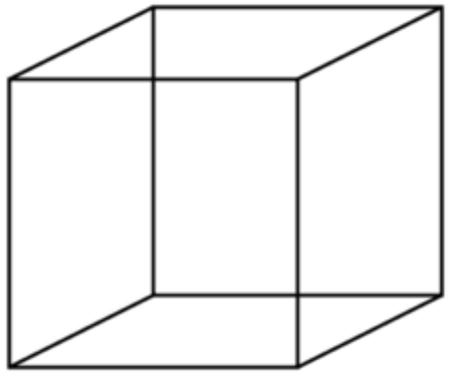
- **Nested Sampling Monte Carlo (NSMC)**
  Skilling 2004
  - For k free parameters to estimate, Bayesian Evidence is a k-dimensional integral

$$\mathcal{E} = \int \mathcal{L}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

Bayes' Theorem

$$p(\boldsymbol{\theta}) = \frac{\mathcal{L}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\mathcal{E}}$$

  - Convert evidence into a one-dimensional integral



$$\mathcal{E} = \int_0^1 \mathcal{L}(X) \, dX$$

$$dX = \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

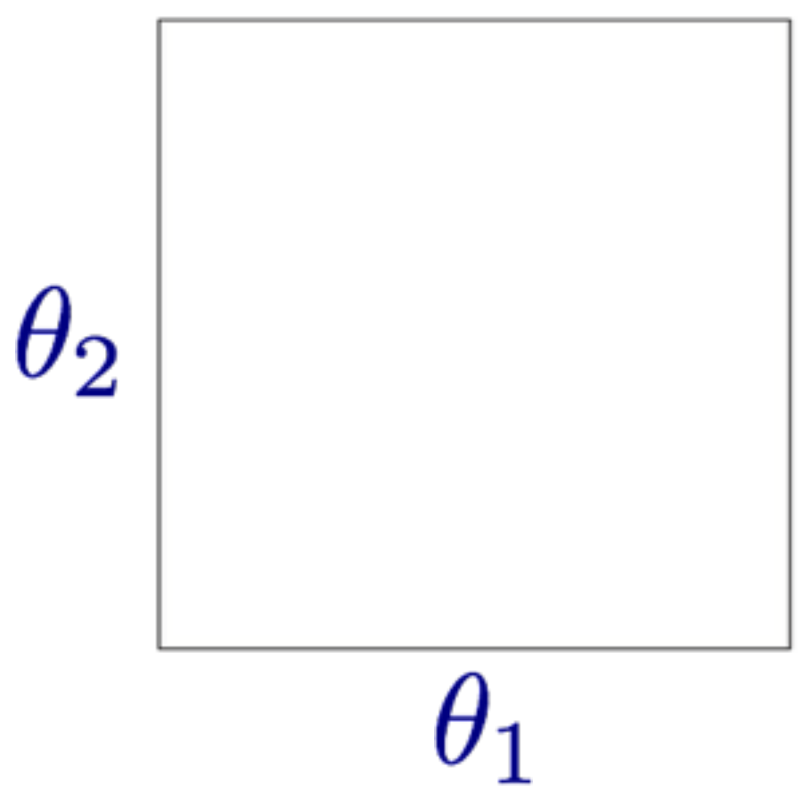
small portion of prior volume (prior mass)

Nested Sampling

in a nutshell…

Bayesian Evidence

$$\mathcal{E} = \int_0^1 \mathcal{L}(X)\, dX$$

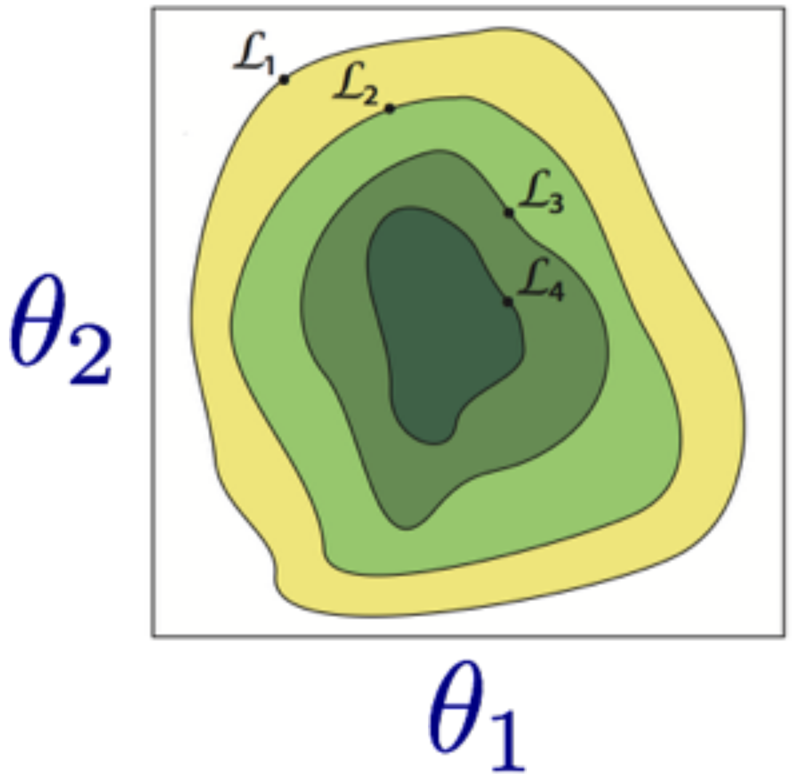


Nested Sampling

in a nutshell…

Bayesian Evidence

$$\mathcal{E} = \int_0^1 \mathcal{L}(X)\, dX$$

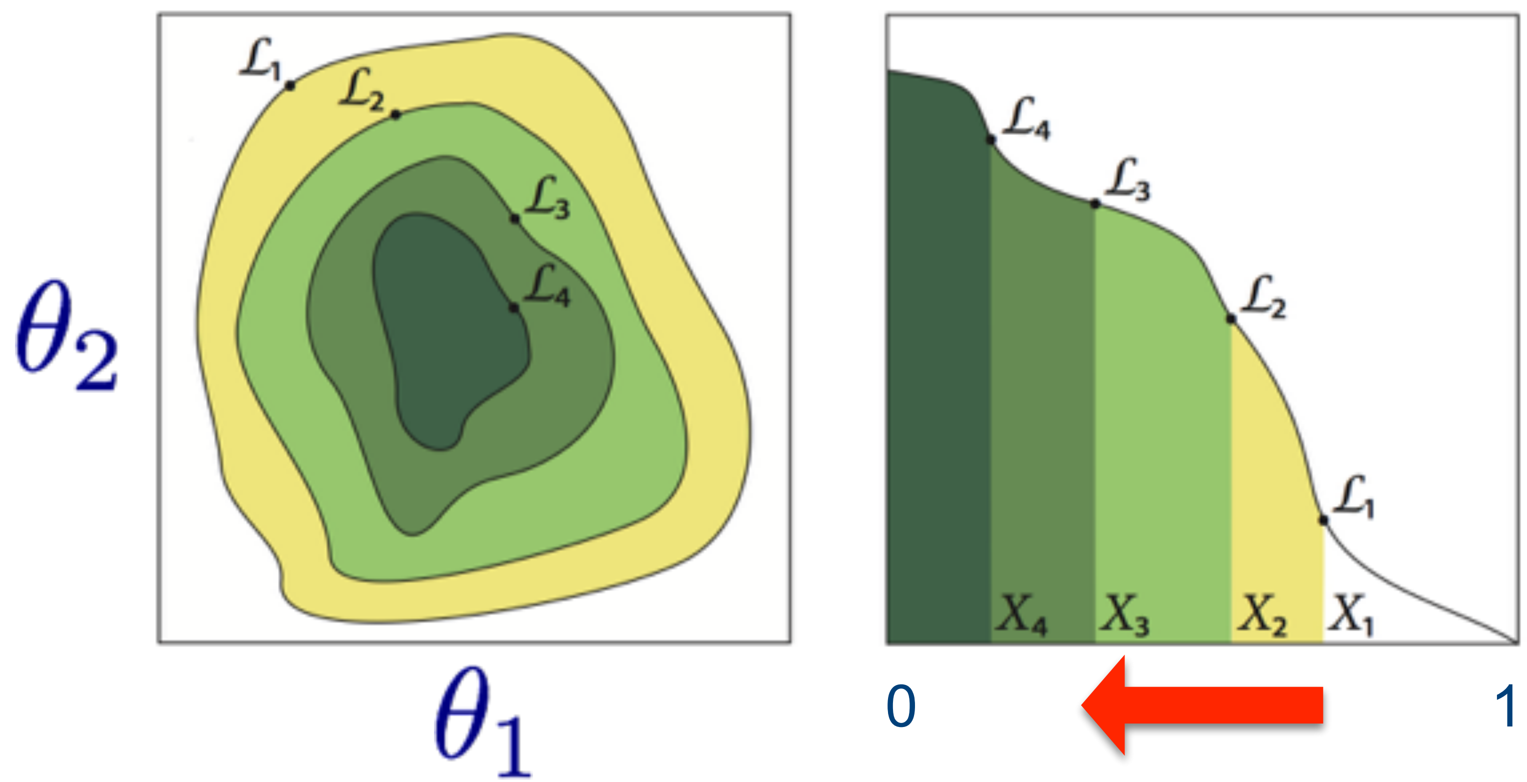


Nested Sampling

in a nutshell…

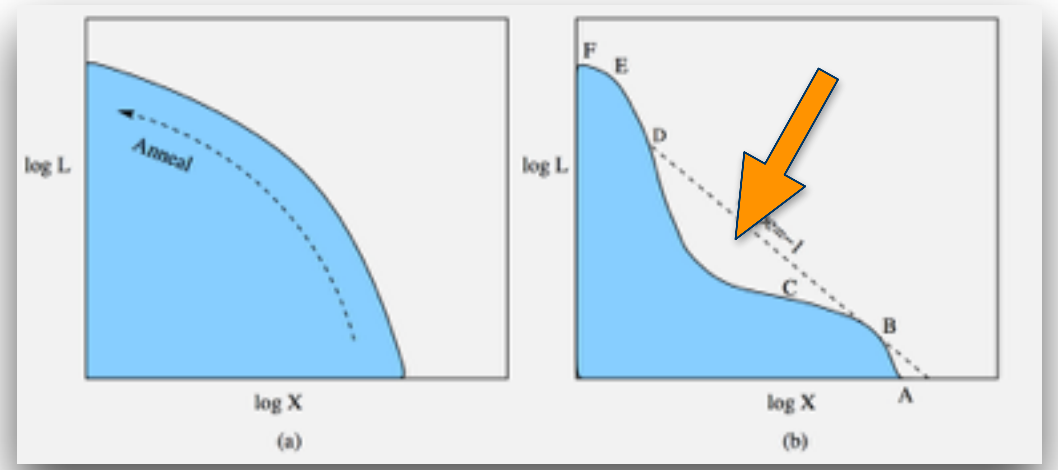
Bayesian Evidence

$$\mathcal{E} = \int_0^1 \mathcal{L}(X)\, dX \quad \longrightarrow \quad \mathcal{E} = \sum_{i=0}^{M} \mathcal{L}_i \Delta X_i$$



Nested Sampling
in a nutshell…

- **ADVANTAGES** with respect to Markov chain Monte Carlo:
  1. Typically **~100 times fewer** samples than thermodynamic integration to calculate evidence to same accuracy + error bar
  2. **Direct** solution to model comparison problems
  3. No troubles with phase changes in likelihood (**multi modal distributions**)



DIAMONDS™
BAYESIAN SOFTWARE

A&A 571, A71 (2014)
DOI: 10.1051/0004-6361/201424181

Astronomy & Astrophysics

DIAMONDS: A new Bayesian nested sampling tool*

Application to peak bagging of solar-like oscillations

E. Corsaro and J. De Ridder

Instituut voor Sterrenkunde, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium
e-mail: emncorsaro@gmail.com, joris.deridder@ster.kuleuven.be

Received 11 May 2014 / Accepted 8 August 2014

What is DIAMONDS ?

high-DImensional And multi-MOdal NesteD Sampling

Corsaro & De Ridder 2014 A&A, 571, 71

- C++11 code for **inference** problems in a Bayesian framework:
  - Dataset to fit
  - Model to test
  - Estimate the free parameters of the model

What is DIAMONDS ?

high-DImensional And multi-MOdal NesteD Sampling

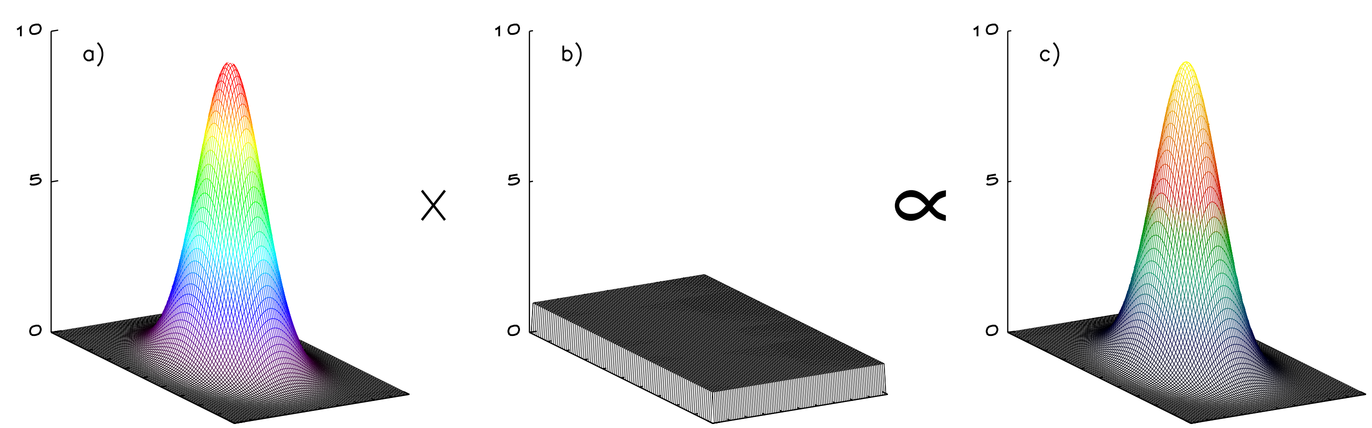
Corsaro & De Ridder 2014 A&A, 571, 71

- C++11 code for **inference** problems in a Bayesian framework:
  - Dataset to fit (**Likelihood**)
  - Model to test (**Prior**)
  - Estimate the free parameters of the model (**Posterior**)

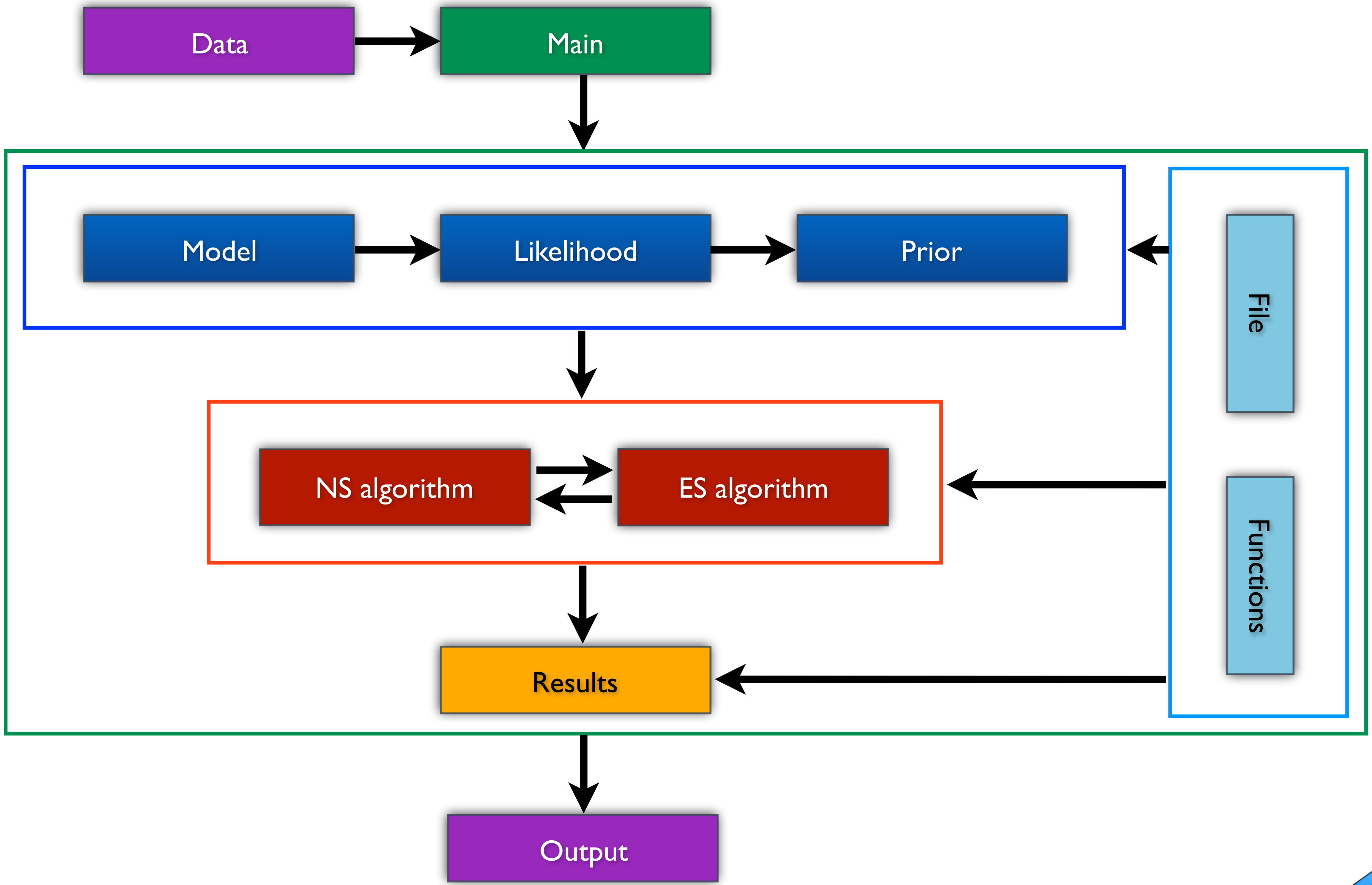
Likelihood **Prior** **Posterior**

Bayes' Theorem

$$p(\boldsymbol{\theta}) = \frac{\mathcal{L}(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\mathcal{E}}$$

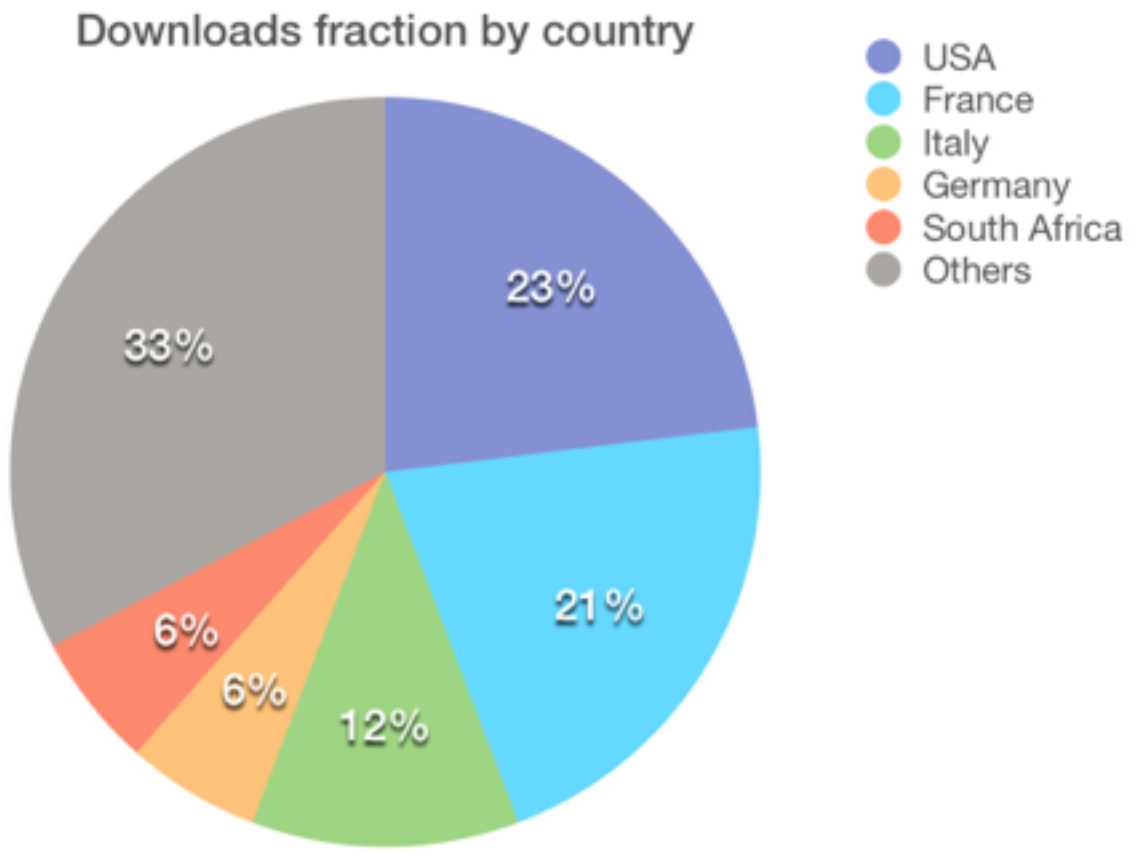


Working scheme



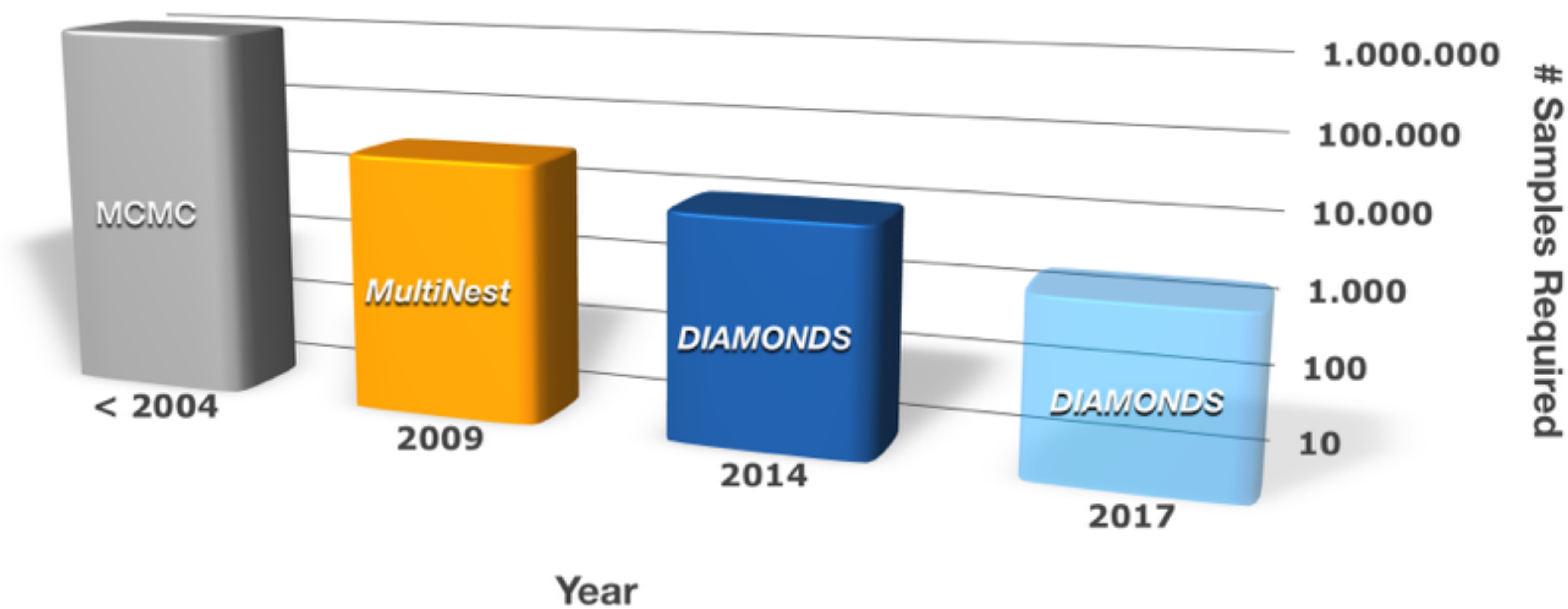
What makes DIAMONDS so appealing? (1)

- Basic core **public** available (now released v. 1.1) with usable demos
- General for **any** application involving Bayesian Inference
- **Bayesian evidence** (essential for model comparison problems) is a direct output
- Very powerful in identifying multiple (degenerate) solutions, also in high-dimensions
- Code implementation is **flexible** and easy to upgrade (replace modules, add new ones)
- Different types of **prior** distributions and **likelihood** functions already provided
- Attracted more than **50 users** from many world's institutions and different fields of physics



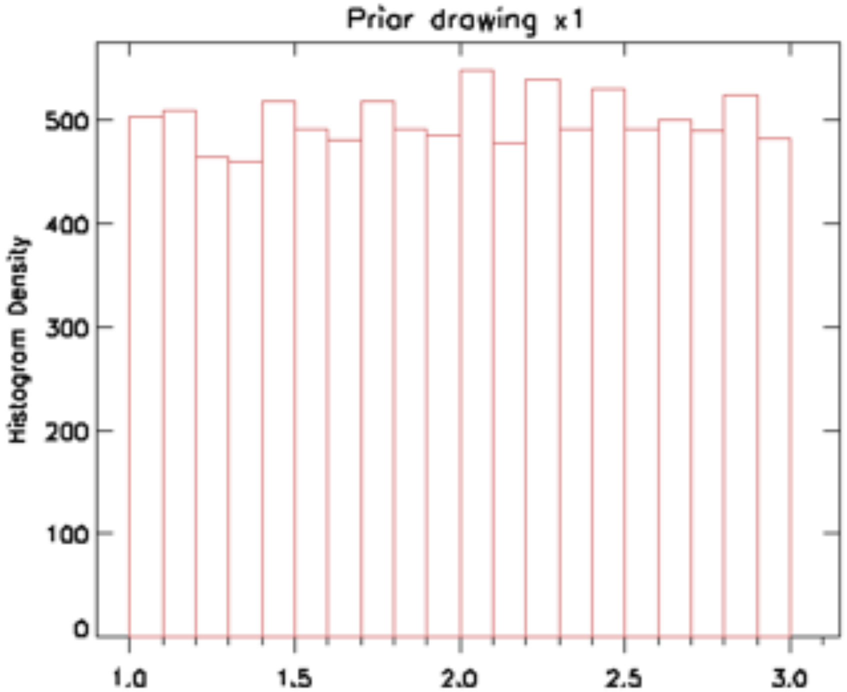
What makes DIAMONDS so appealing? (2)

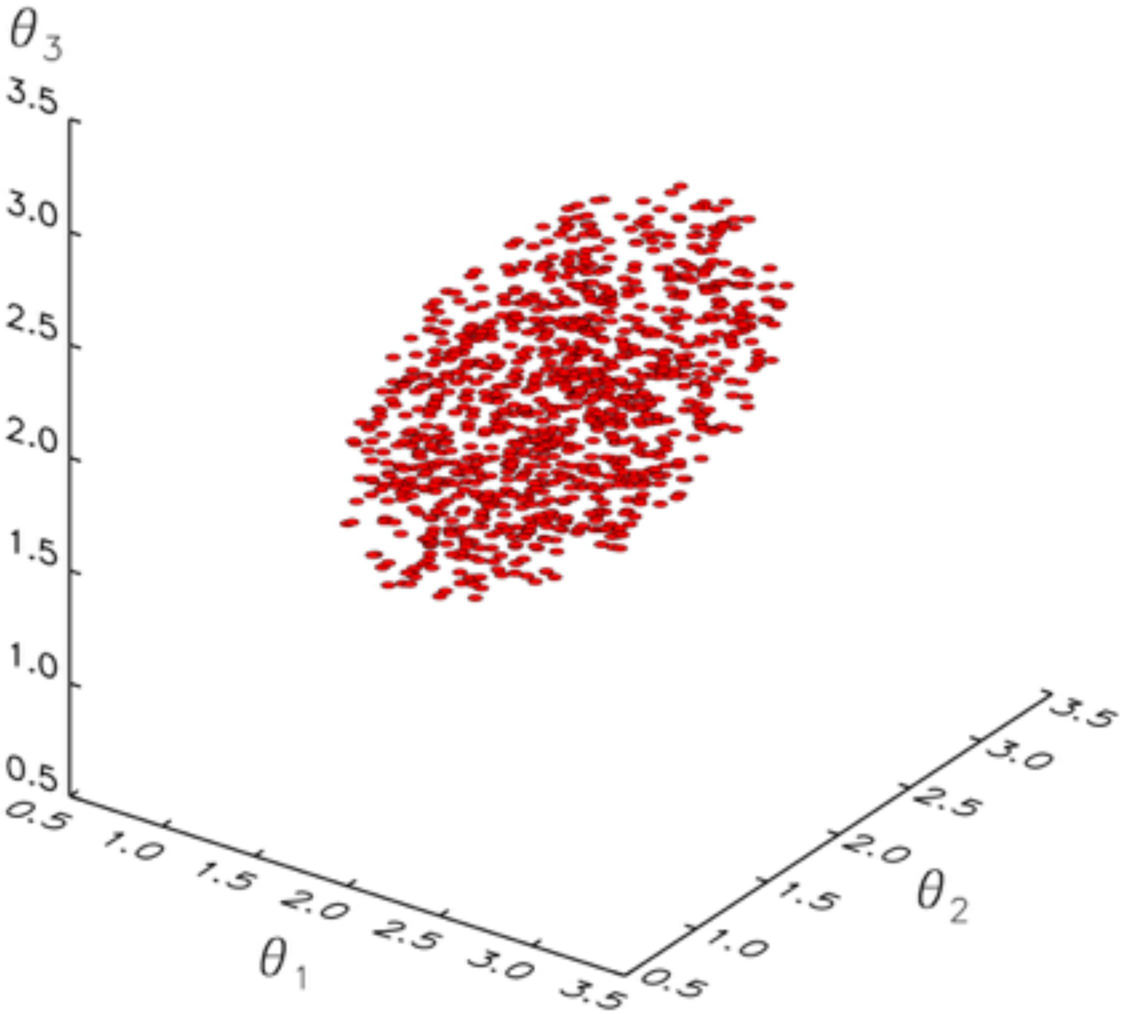
- **Overtakes** other existing MCMC, NSMC codes (e.g. MultiNest, POLYCHORD)
- Foreseen upgrade with full multi-core parallelization by early 2017 (v. 2.0)



Prior distributions

Uniform

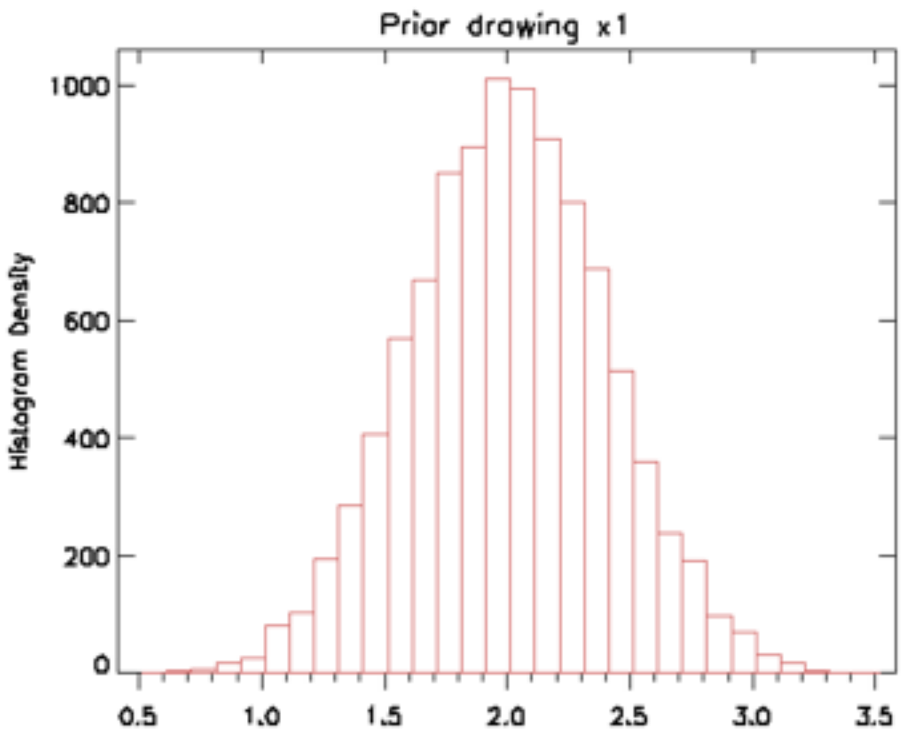


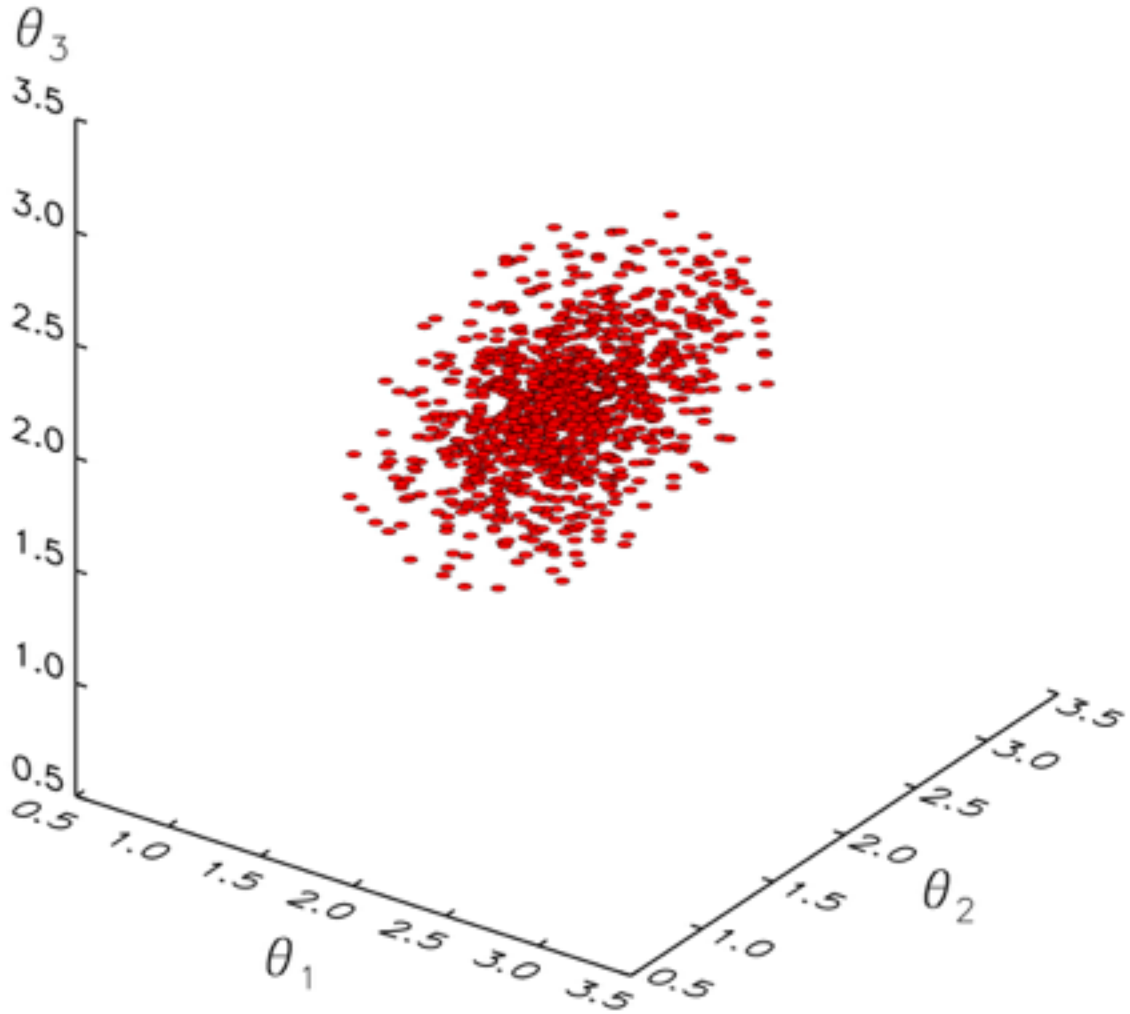


Corsaro & De Ridder 2014 A&A, 571, 71

Prior distributions

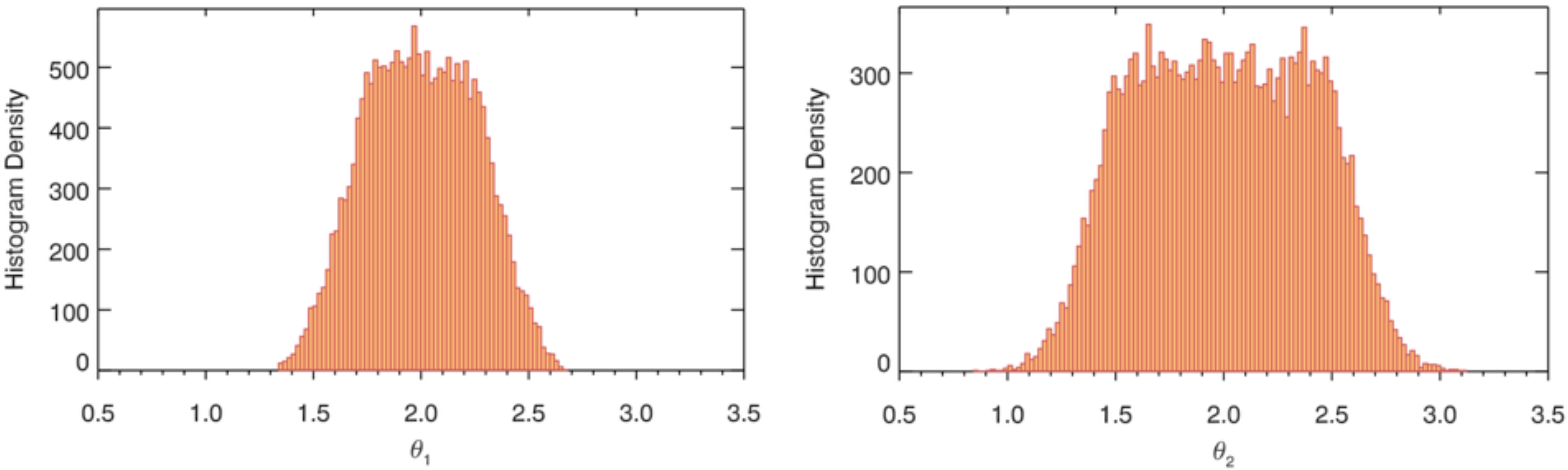
Gaussian





Corsaro & De Ridder 2014 A&A, 571, 71

Prior distributions

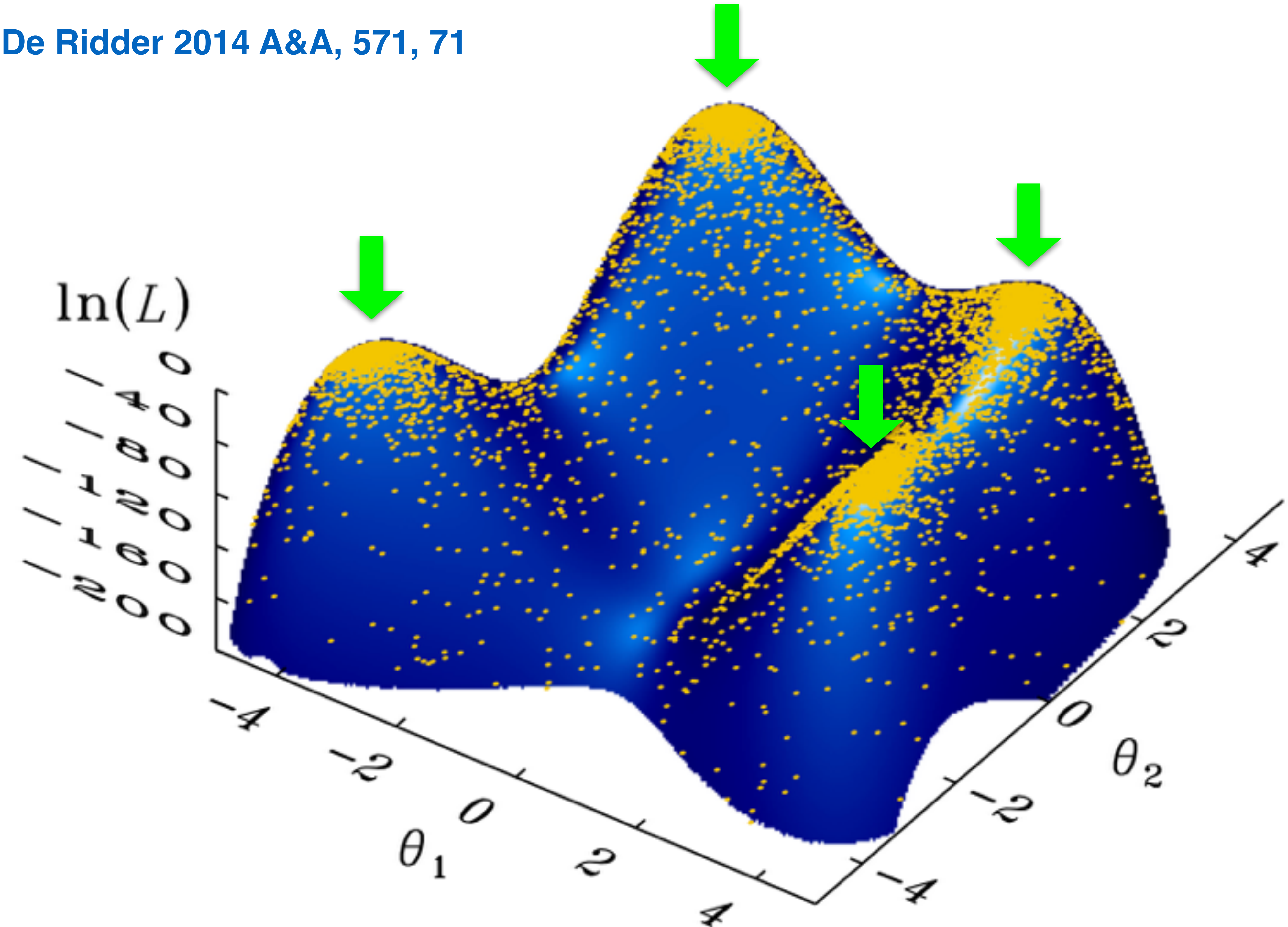


Super Gaussian

Corsaro & De Ridder 2014 A&A, 571, 71

Sampling efficiency demos

Corsaro & De Ridder 2014 A&A, 571, 71



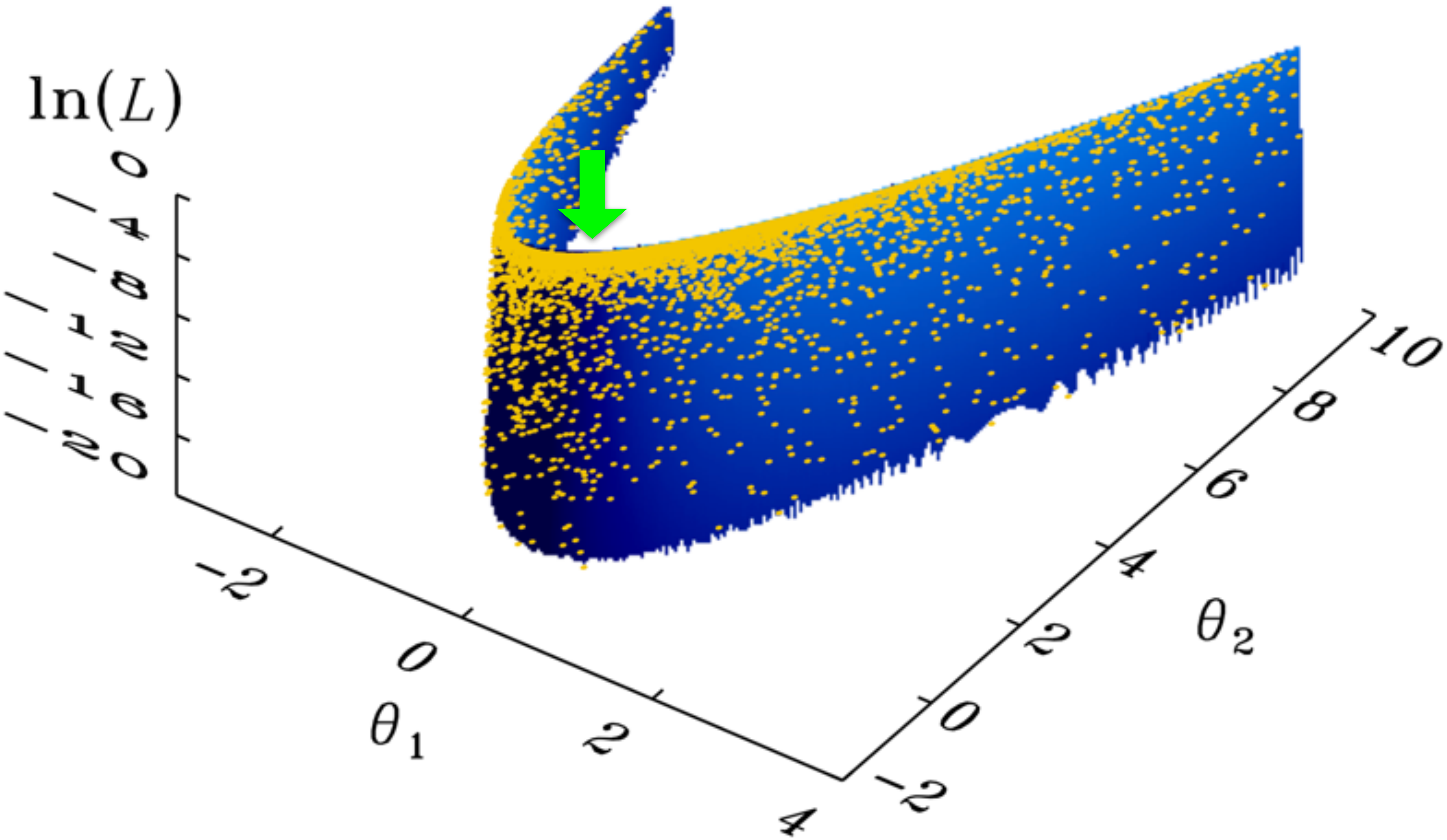
Himmelblau's Function

N = 8485 Samples

Enrico Corsaro - enrico.corsaro@cea.fr 8 December 2015

Sampling efficiency demos

Corsaro & De Ridder 2014 A&A, 571, 71

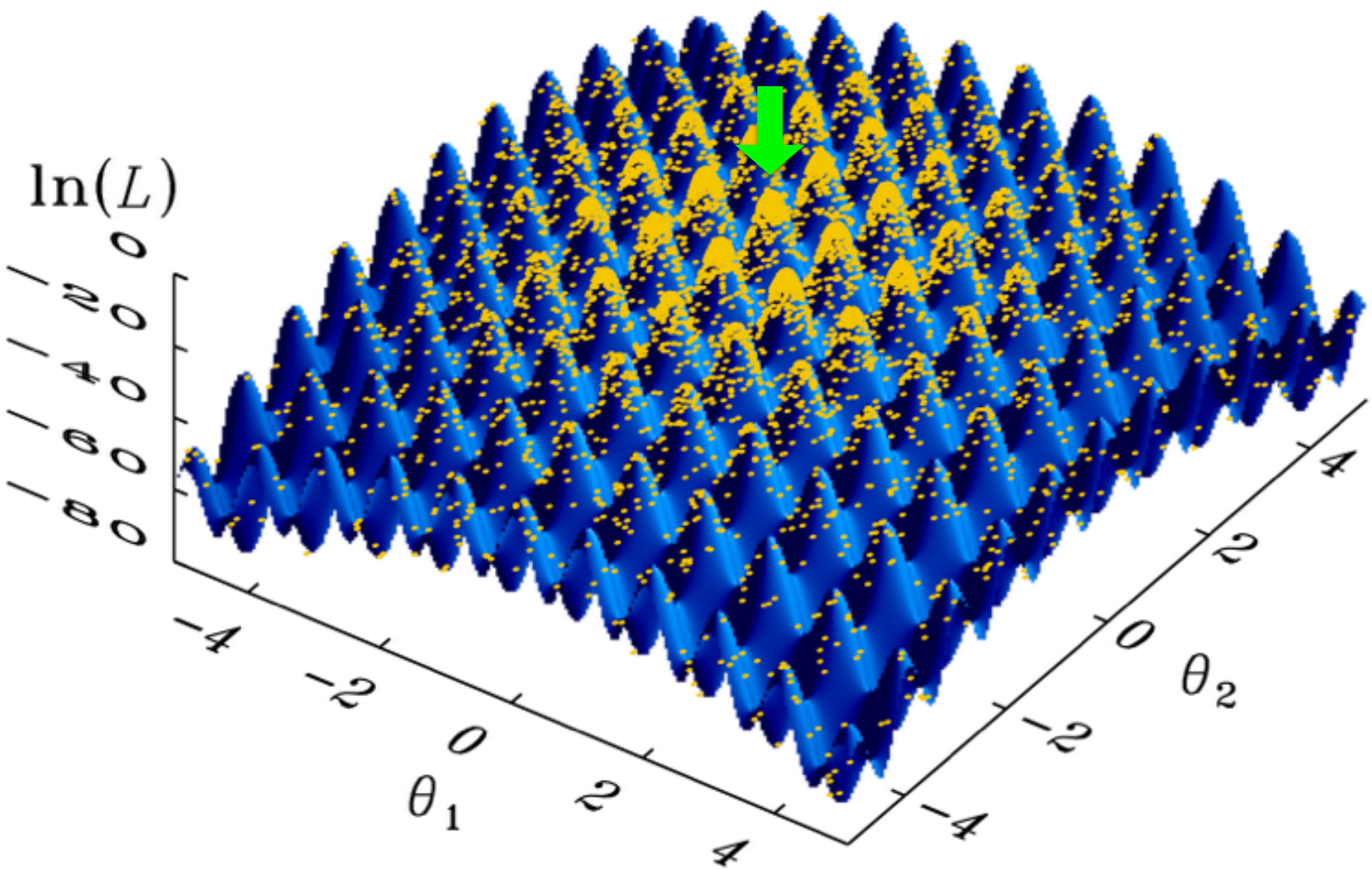


Rosenbrock's Function

N = 8558 Samples

Sampling efficiency demos

Corsaro & De Ridder 2014 A&A, 571, 71

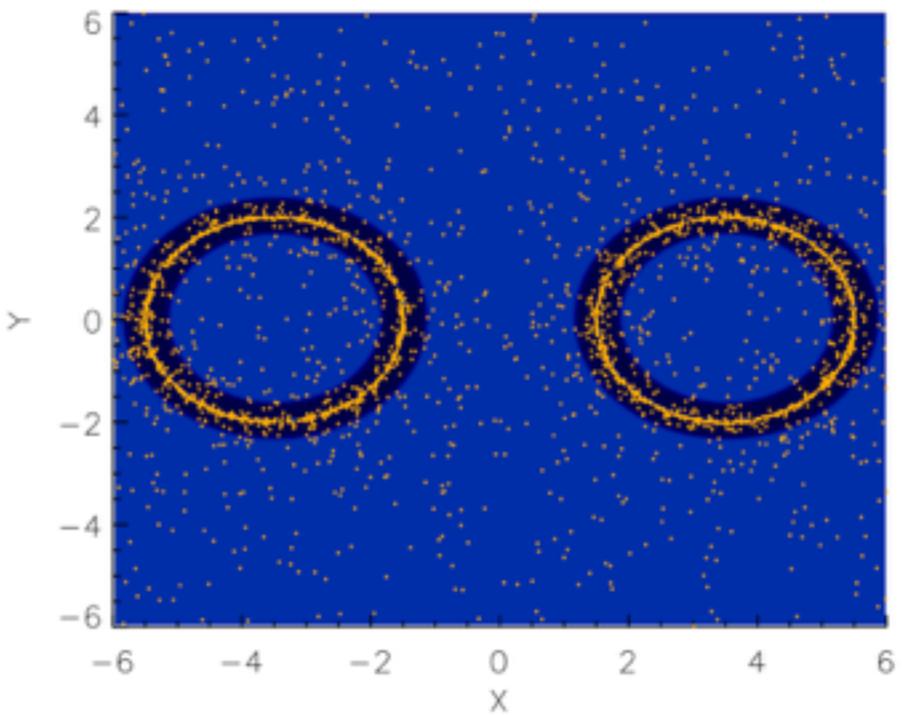


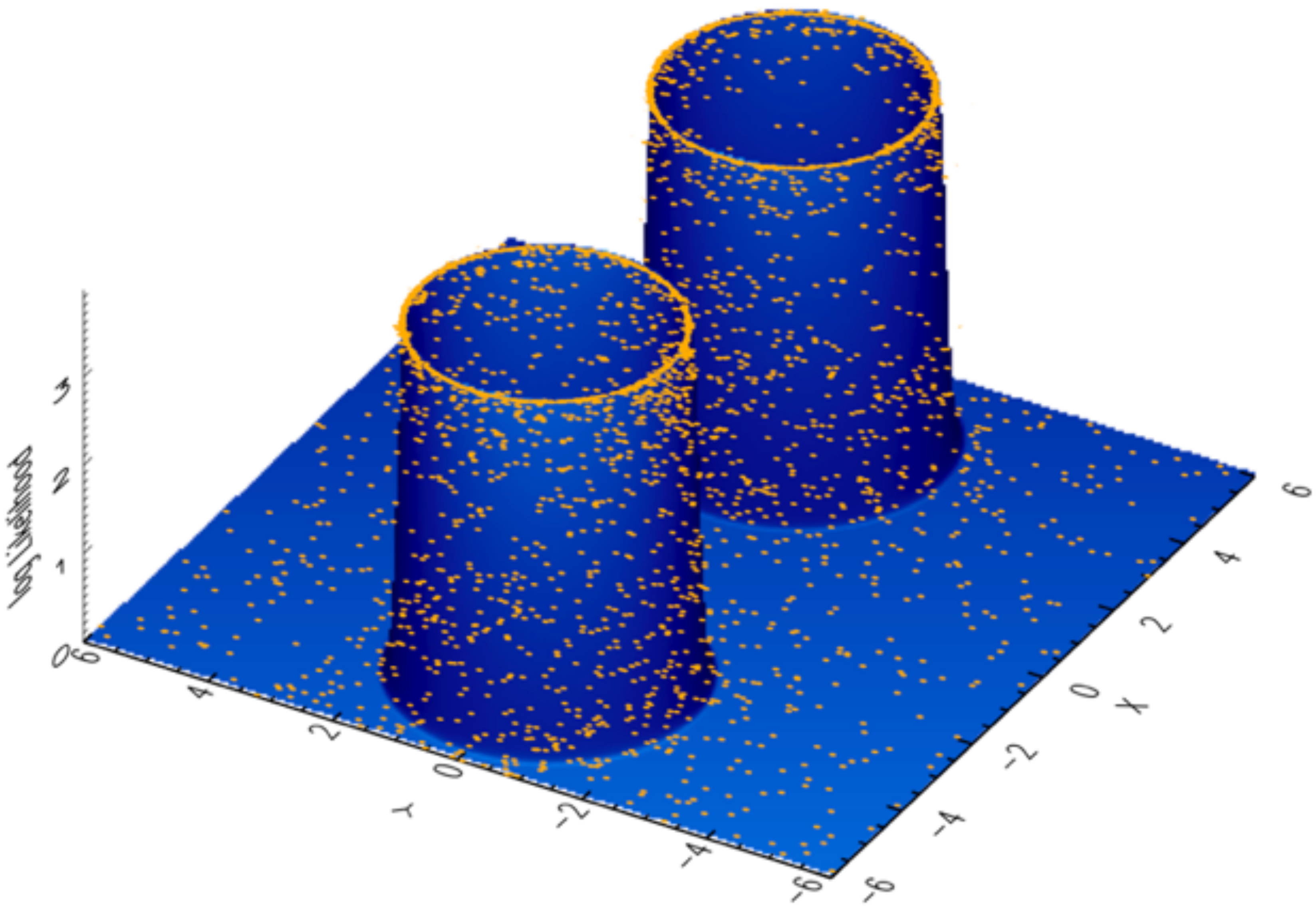
Rastrigin's Function

N = 10648 Samples

Sampling efficiency demos

Corsaro & De Ridder 2014 A&A, 571, 71



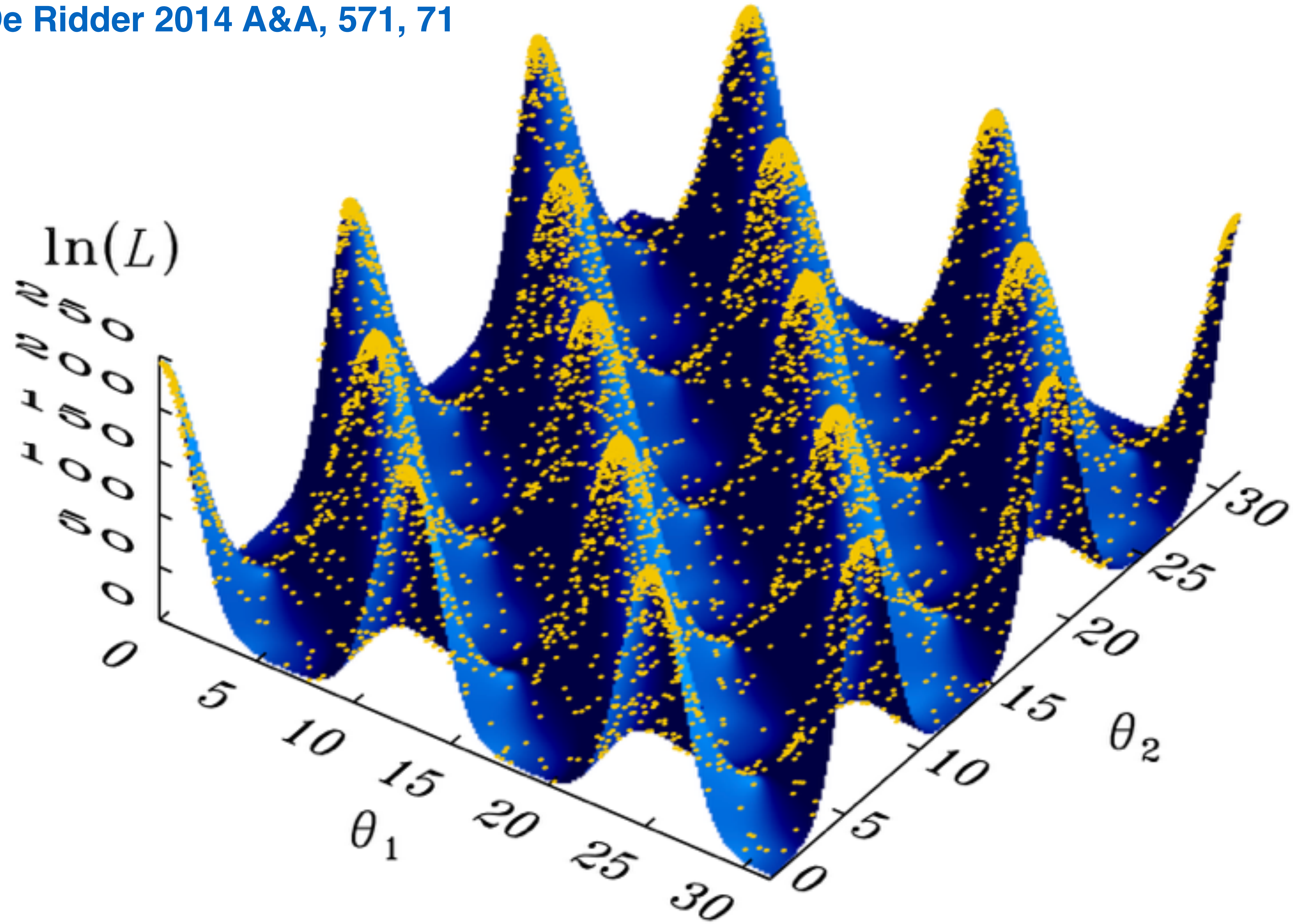


Gaussian Shell Function

N = 3100 Samples

Sampling efficiency demos

Corsaro & De Ridder 2014 A&A, 571, 71

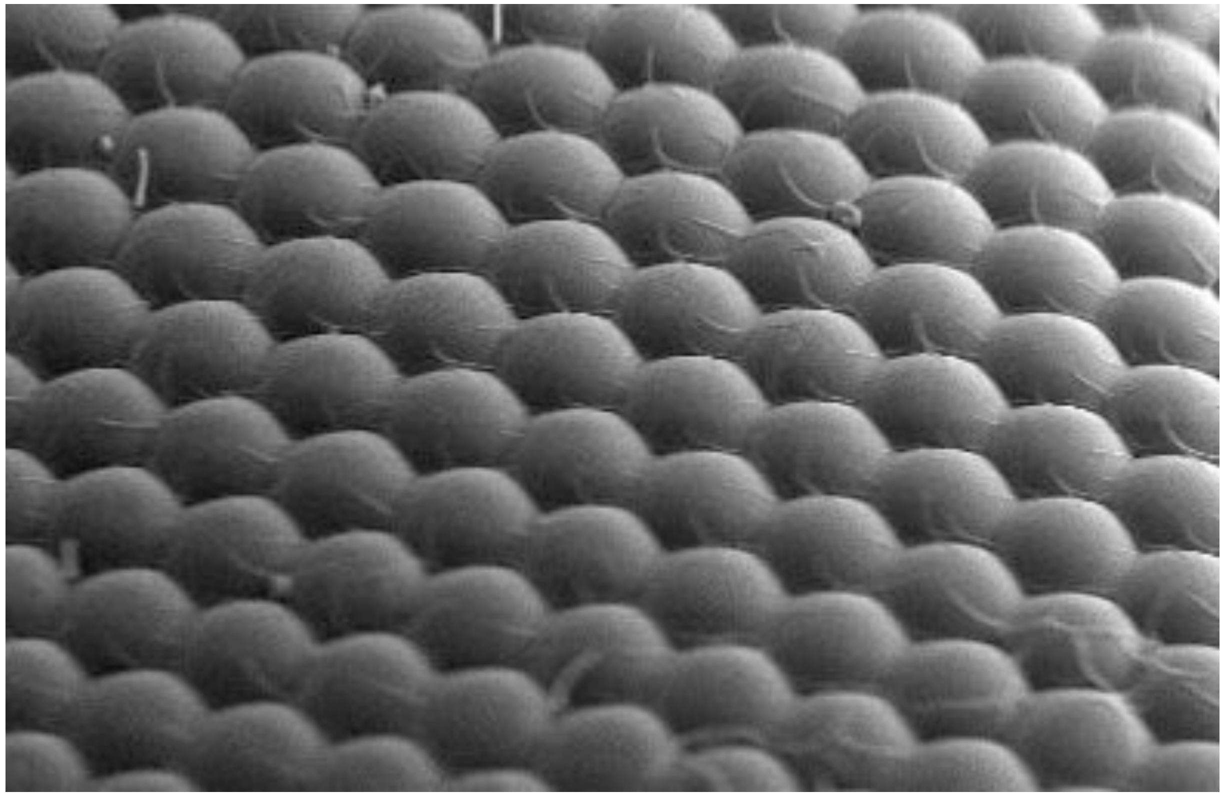


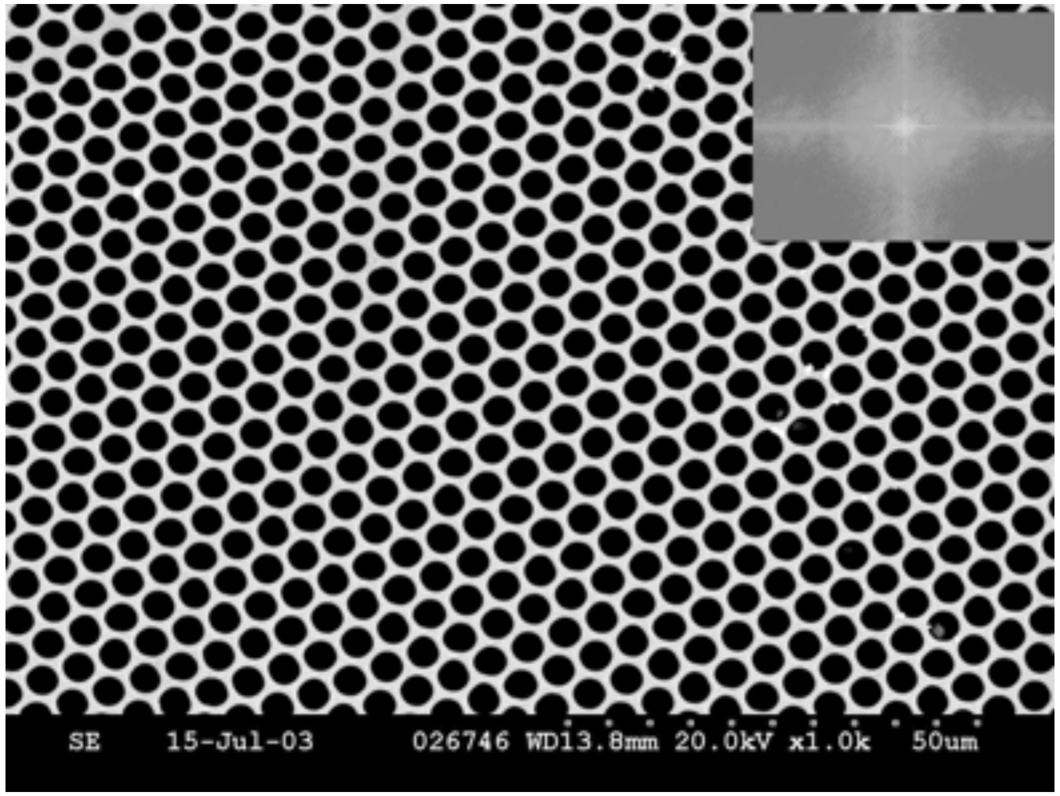
Eggbox Function

N = 8207 Samples

Examples for real applications

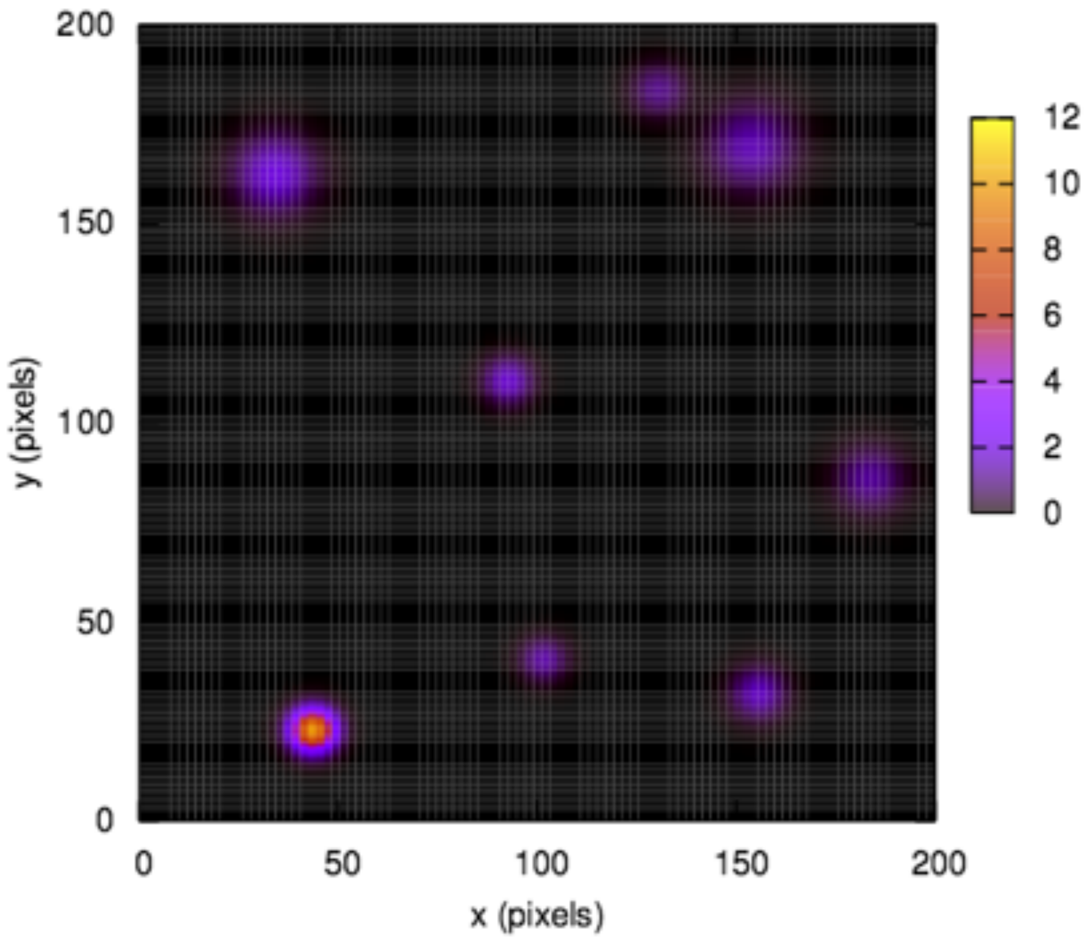
Scanning Electron Microscopy (SEM)
e.g. detecting the position of individual atoms

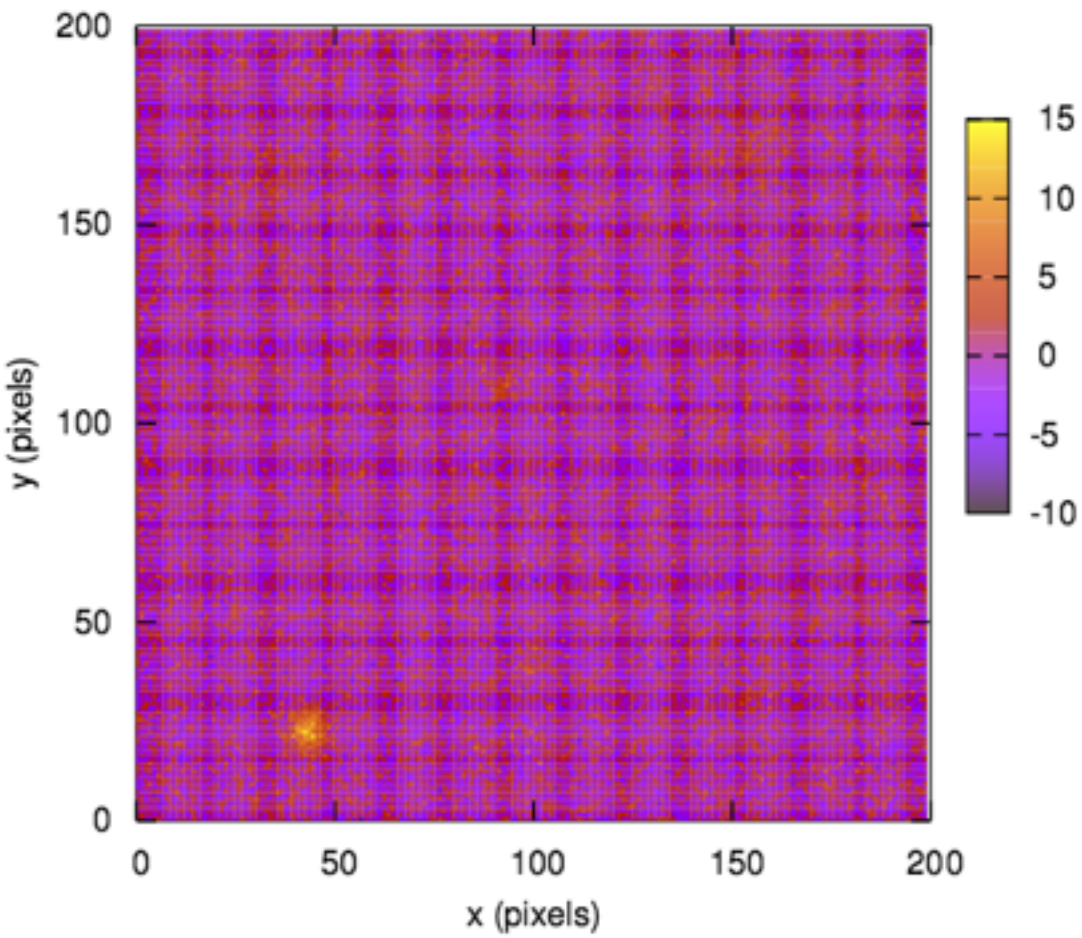




Examples for real applications

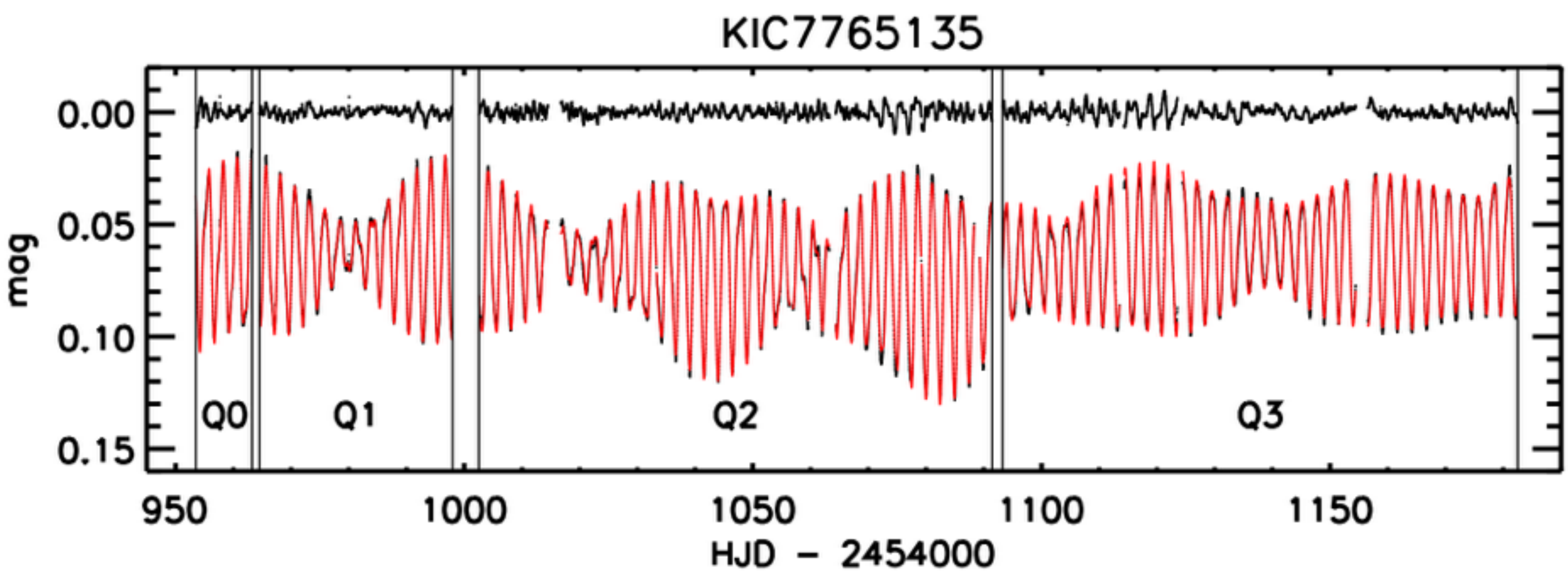
Detecting signal from a noisy background
e.g. detecting SZ effect in CBR maps





Examples for real applications

Fitting very complex time-series shapes
e.g. spot modeling for differential rotation in active stars



Asteroseismology

Why do we need DIAMONDS?

1. Tackling **high-dimensional** and/or **multi-modal** fitting problems at high speed (otherwise very difficult, if not impossible, to solve with standard methods and available computational power)

2. Easy and direct solution to model comparison problems

For example?

Asteroseismology!

but also…

exoplanetary science, solar physics, cosmology, high-energy physics, etc.

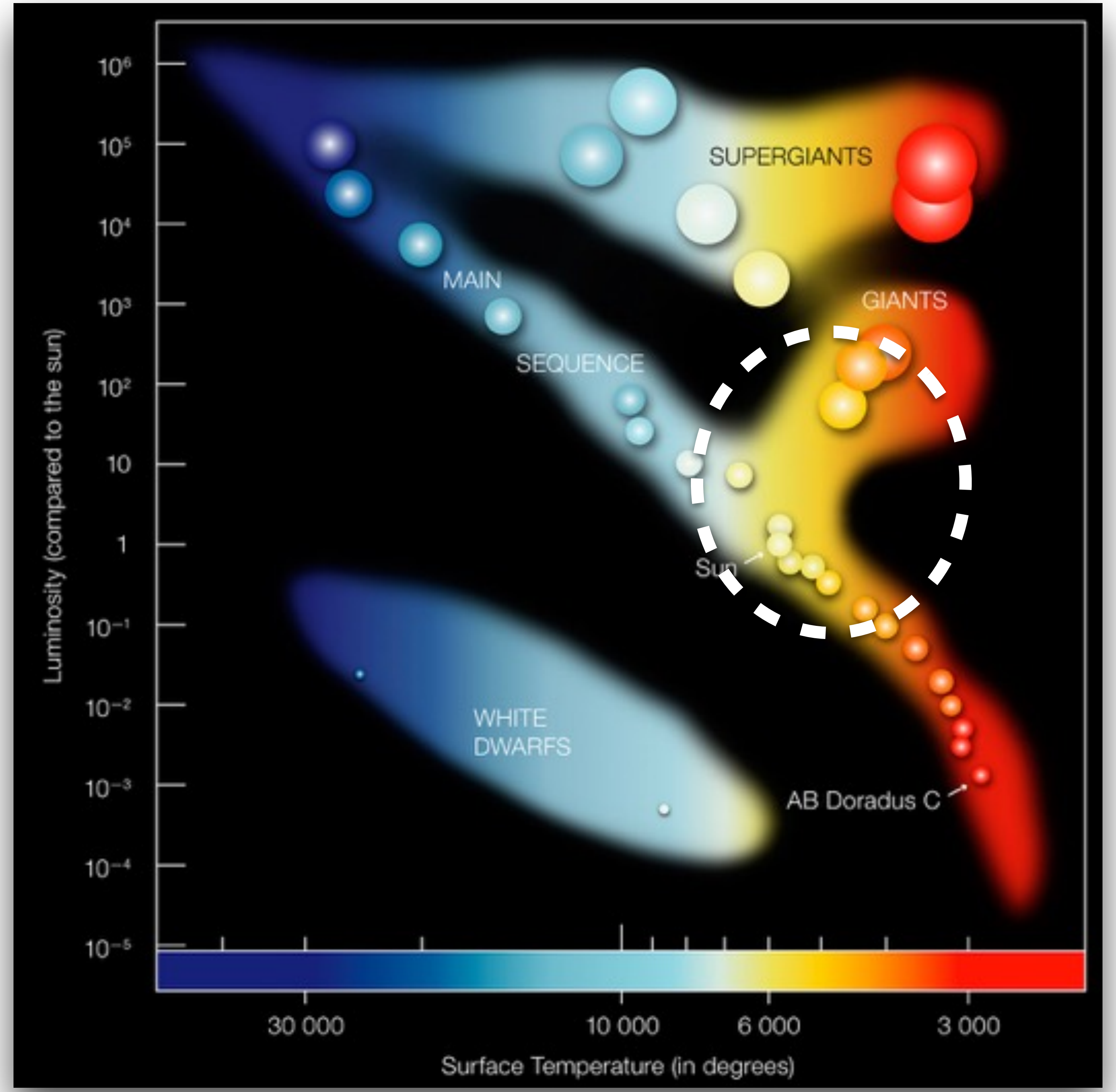
What is Asteroseismology?

The analysis of stellar oscillations to probe stellar structure, dynamics, and evolution

- The **most powerful** available approach to look inside the stars!
- Our main example: **the Sun** (helioseismology)
- Many stars oscillate similarly to the Sun (**solar-like**): about **40,000 known to date** and growing every year

Solar-like oscillators

- Spectral types F-K, can be observed at galactic scale distances (up to about 10 kpc)
- The most common, hence constitute a statistically useful sample for population studies
- Often host Earth-sized planets and offer extended habitability zones, hence crucial to study the conditions for life
- Cover all epochs of star formation in the different regions of the Galaxy (thin and thick disks, and the halo)
- Show narrow spectral lines that provide more accurate and precise element abundances than for other stars.



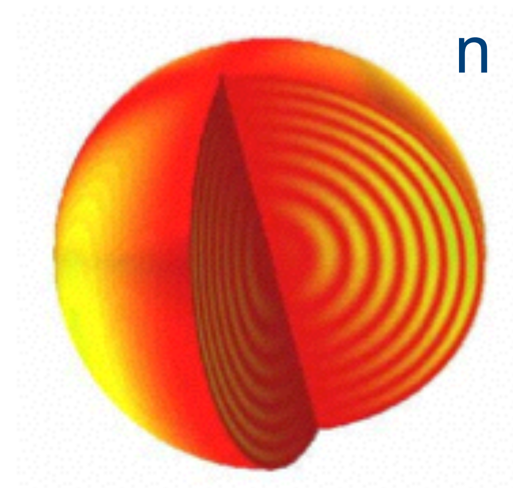
Solar-like oscillations

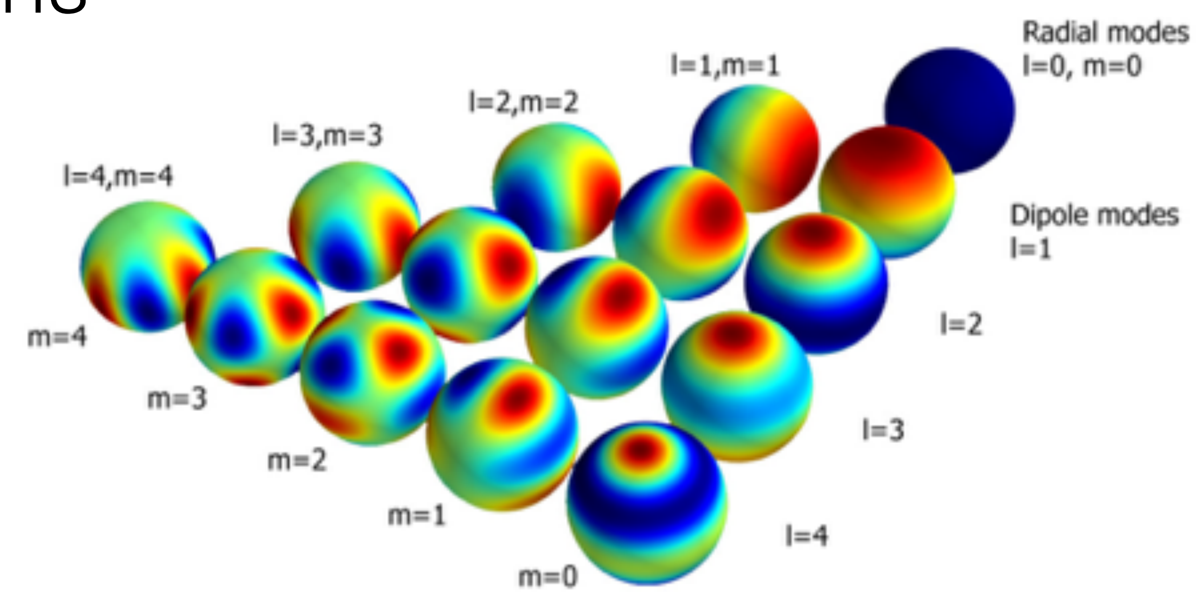
- Acoustic waves from surface convection in low- and intermediate-mass stars (***p* modes**)

- Produce tiny brightness variations (**10^{-6} - 10^{-3} mag**)

- Each oscillation mode can be identified by three quantum numbers

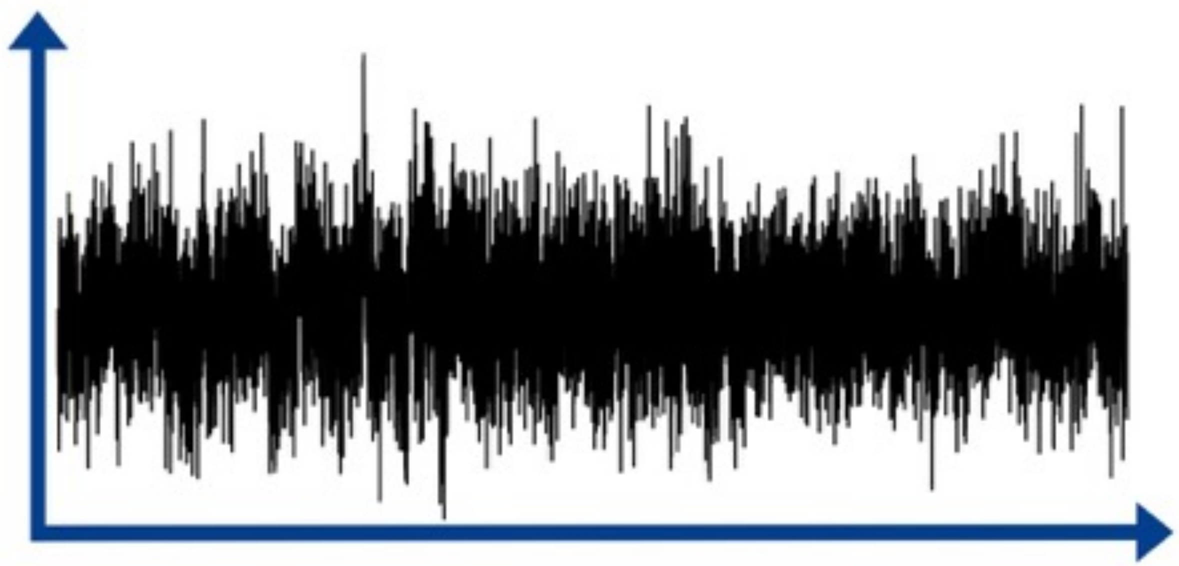
$$\nu_{\mathrm{n,l,m}}$$





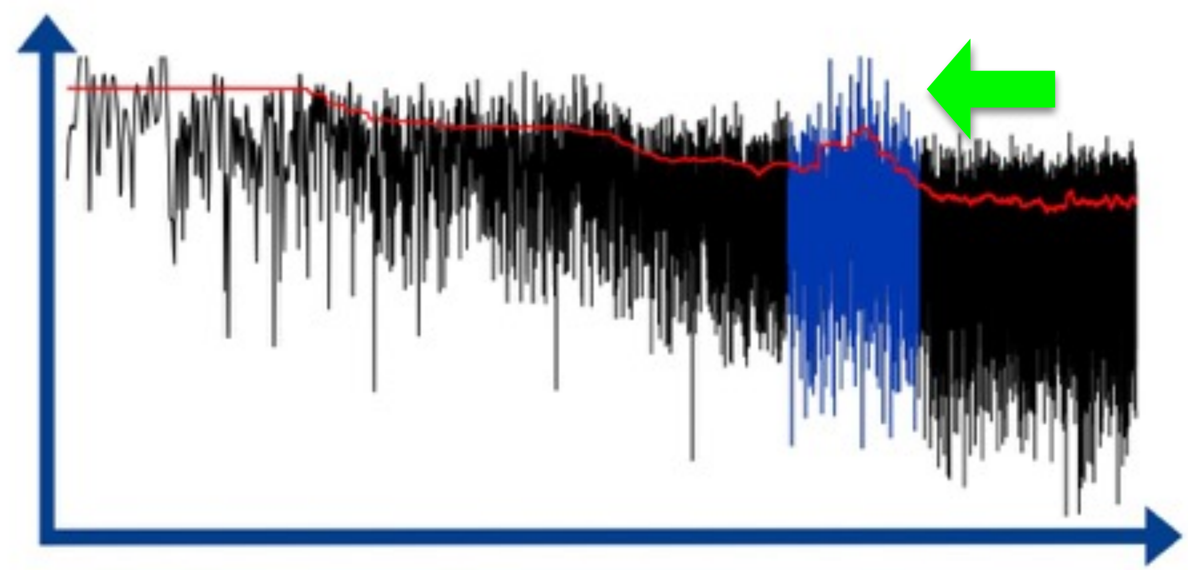
Beck & Kallinger S&W 2013

Time-series analysis



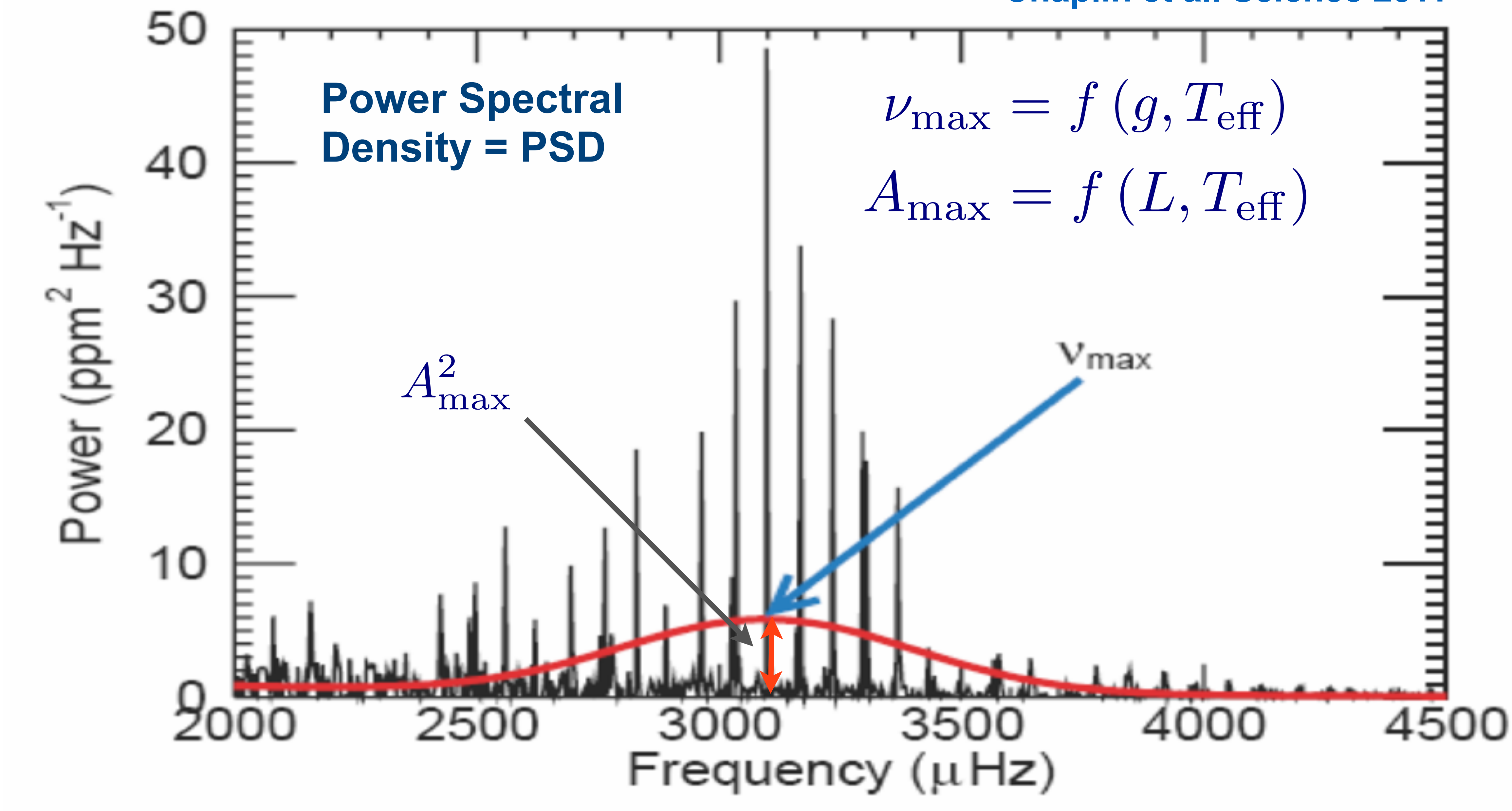
CoRoT Kepler, K2 TESS

Fourier Transform



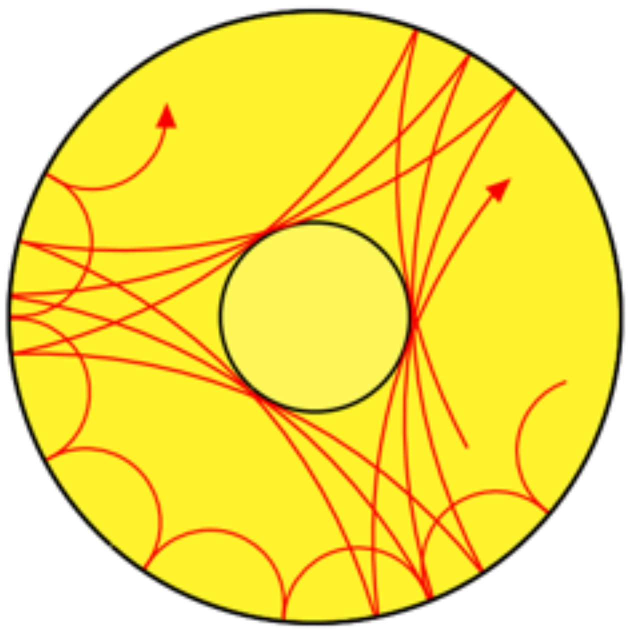
Global parameters

Chaplin et al. Science 2011



Global parameters

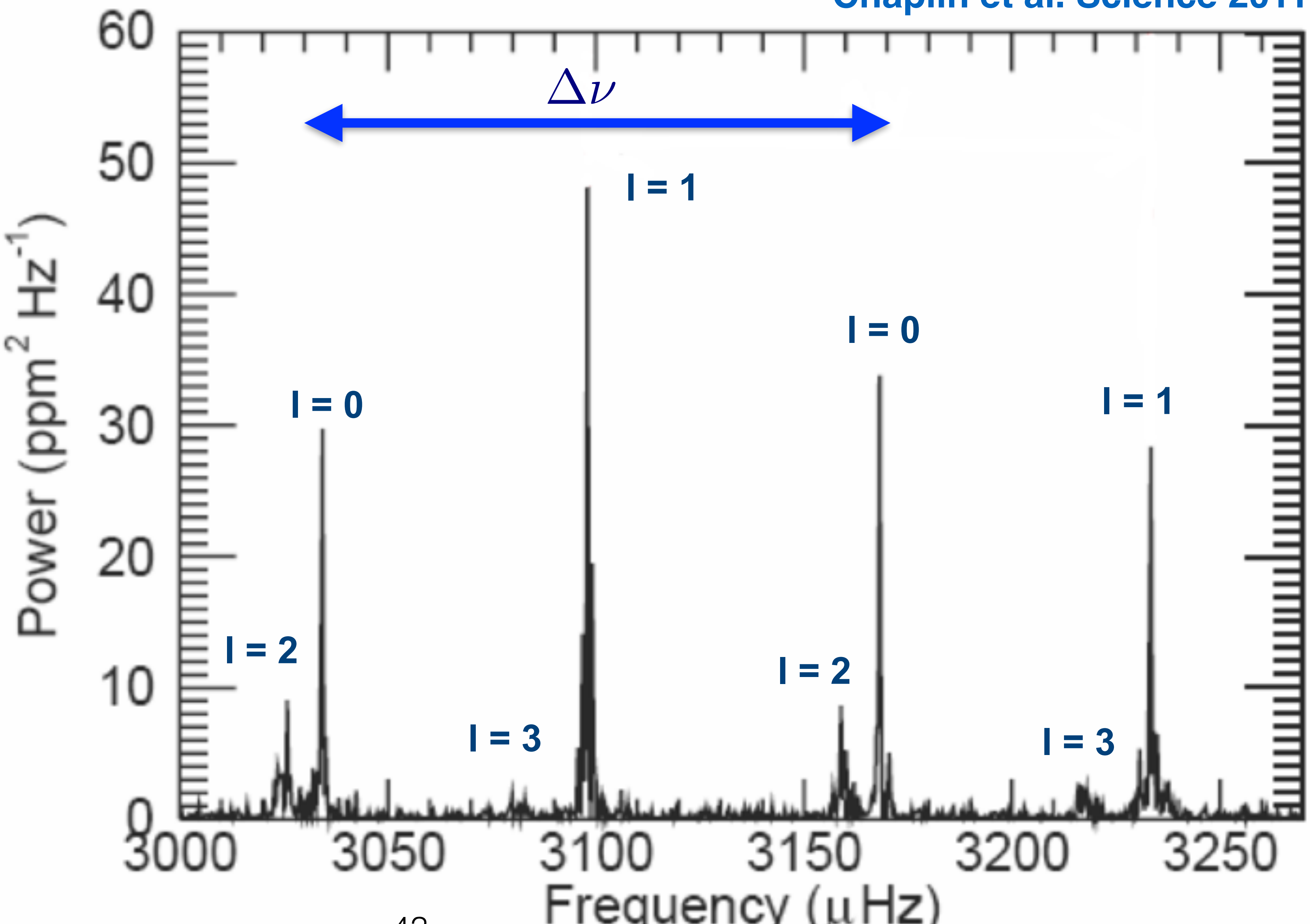
$$\Delta\nu \propto \bar{\rho}$$



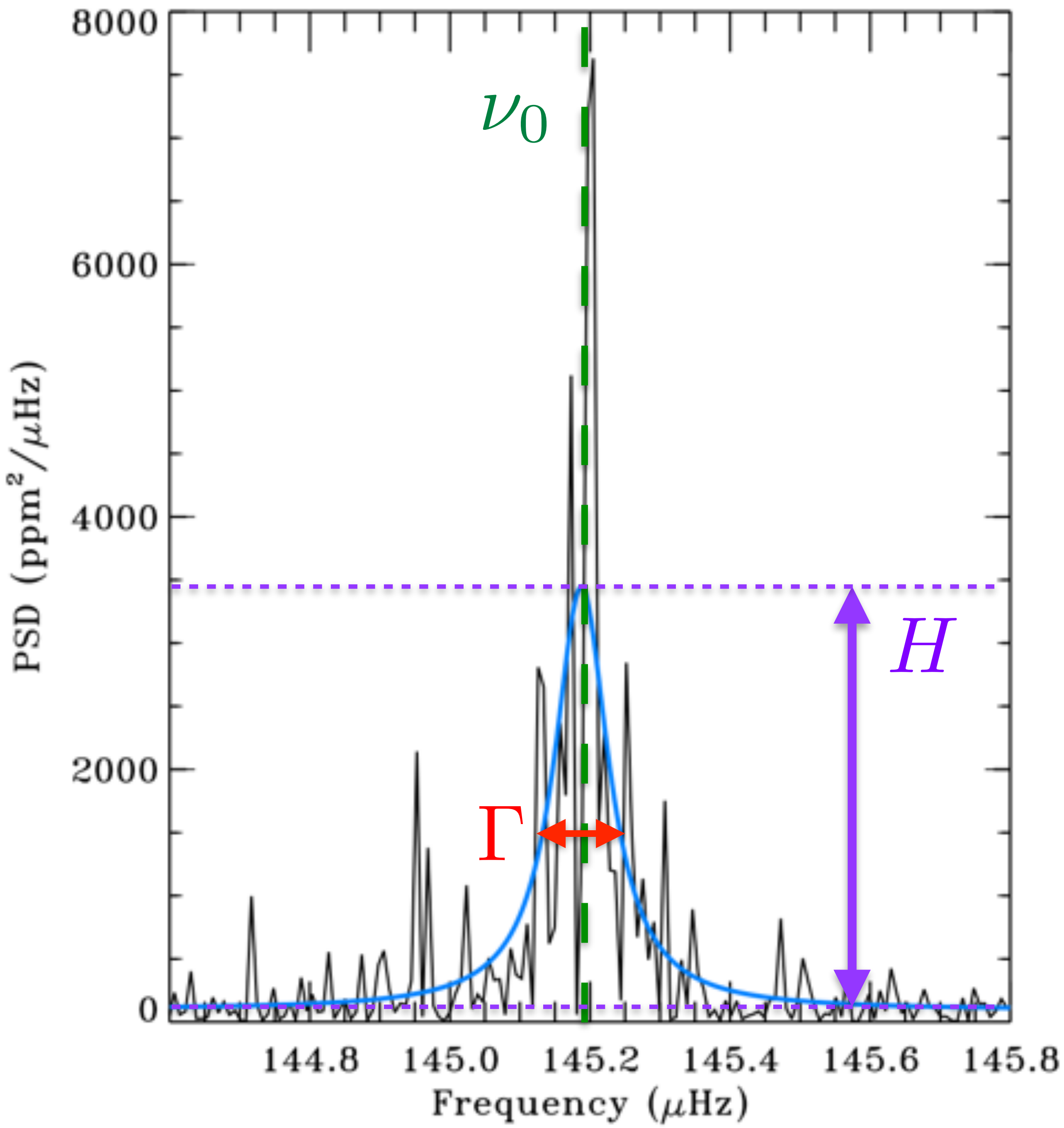
↑ $\mathrm{Age}, A_{\mathrm{max}}$

↓ $\nu_{\mathrm{max}}, \Delta\nu$

Chaplin et al. Science 2011



Fine-structure of *p* modes



Damped oscillation

$\nu_{\mathrm{n,l,m}}$

Lorentzian profile

$$T_{\mathrm{obs}} \gg \tau$$

$$\Gamma \propto \tau^{-1}$$

$$A^2 = \pi H \Gamma$$

$$\nu_0, \Gamma, A$$

Why do we need this?

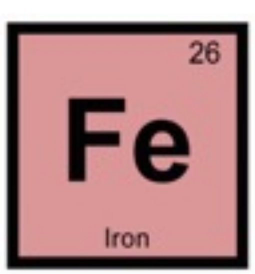
Constrain and understand to the best level possible

+ Spectroscopy

Physical Properties & Internal Structure

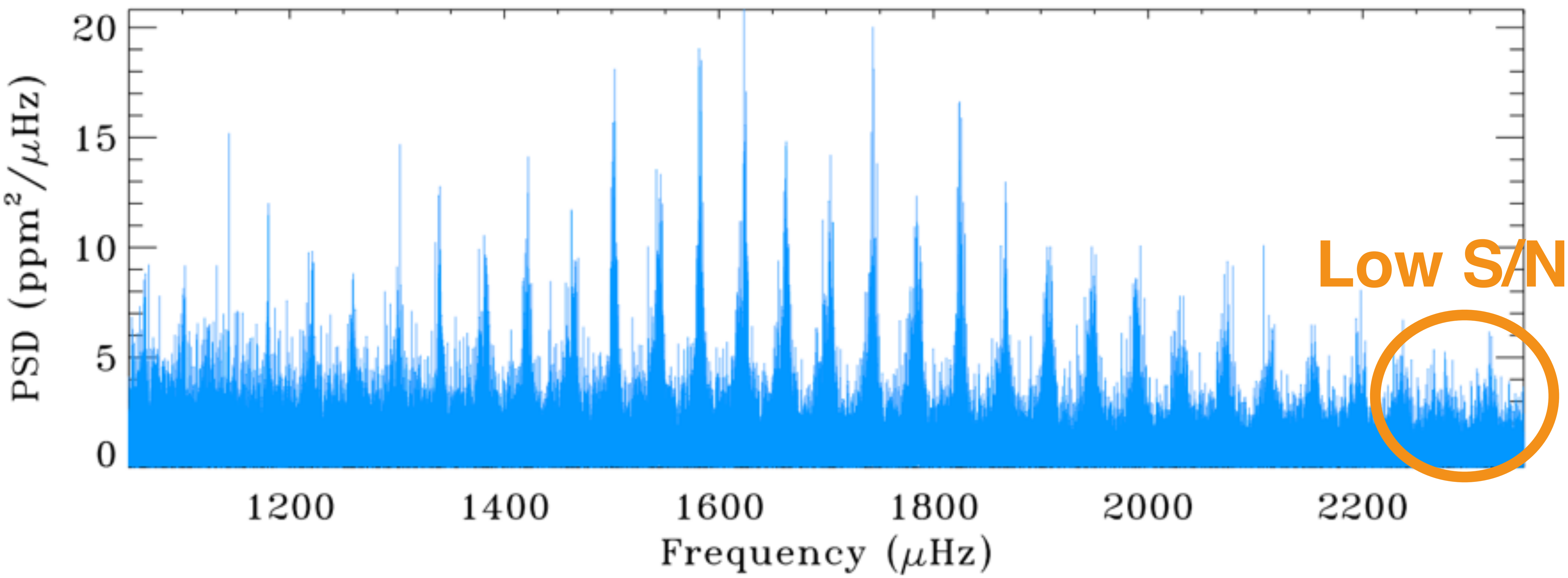
Mass, Radius to few percent precision

Position of BCZ, HeII zone
Evolutionary stage



Metallicity effect

Problems



- **Problem 1**: big dataset + fitting numerous oscillation modes (peaks) per star (can be more than 100)
- **Problem 2**: testing if a peak is real or not (noise)

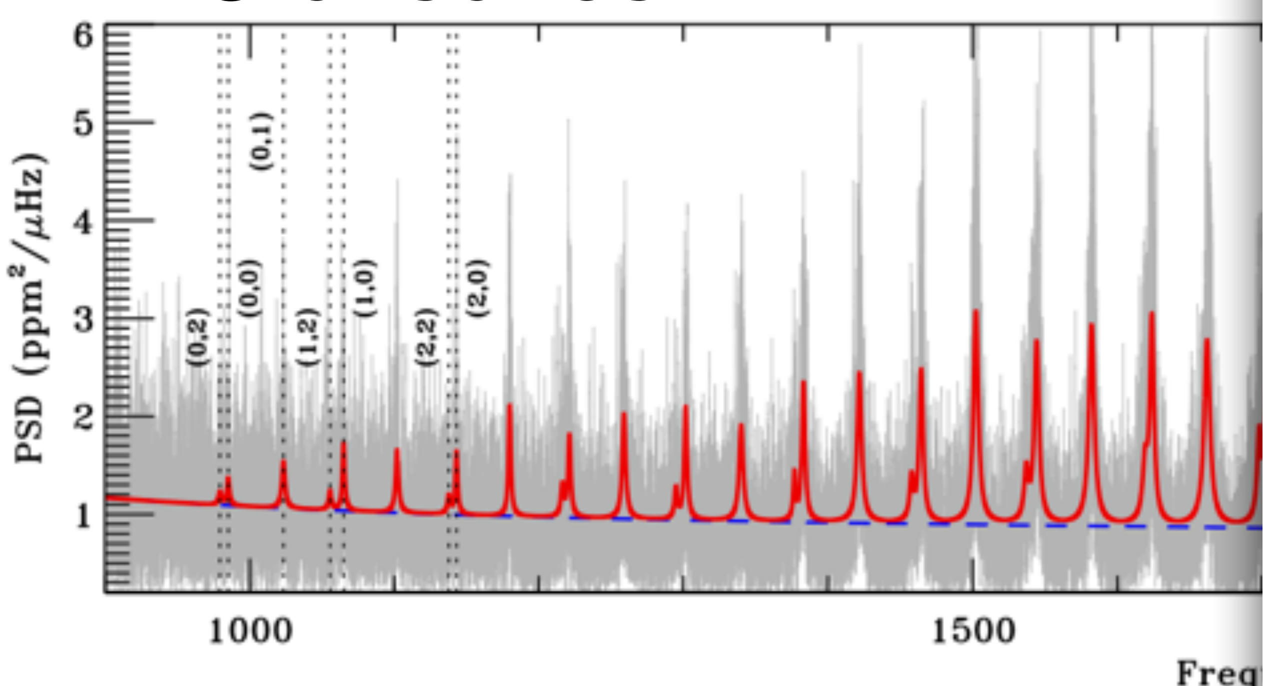
Problem 1

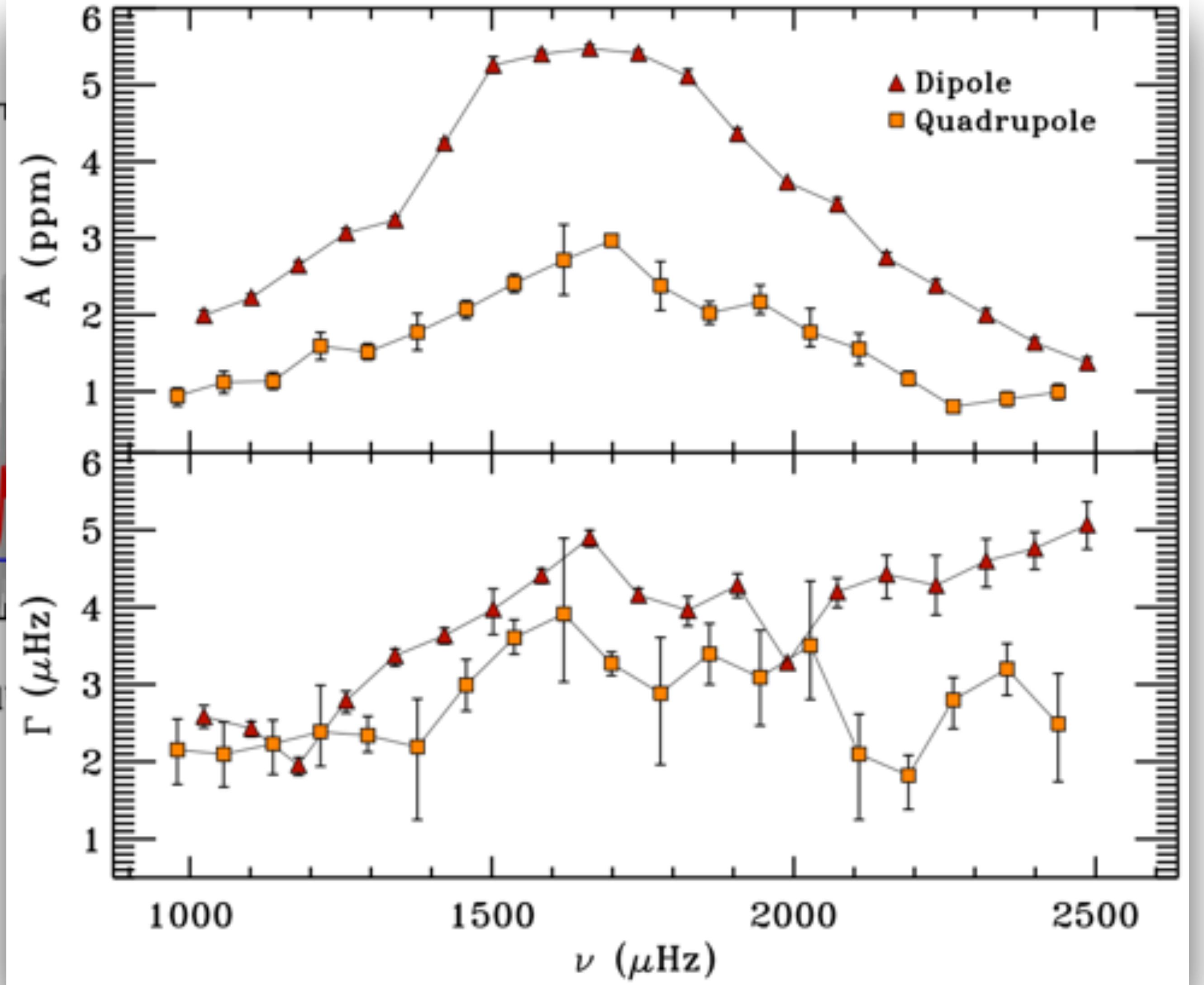
Solving a high-dimensional fitting problem

High-dimensional Model

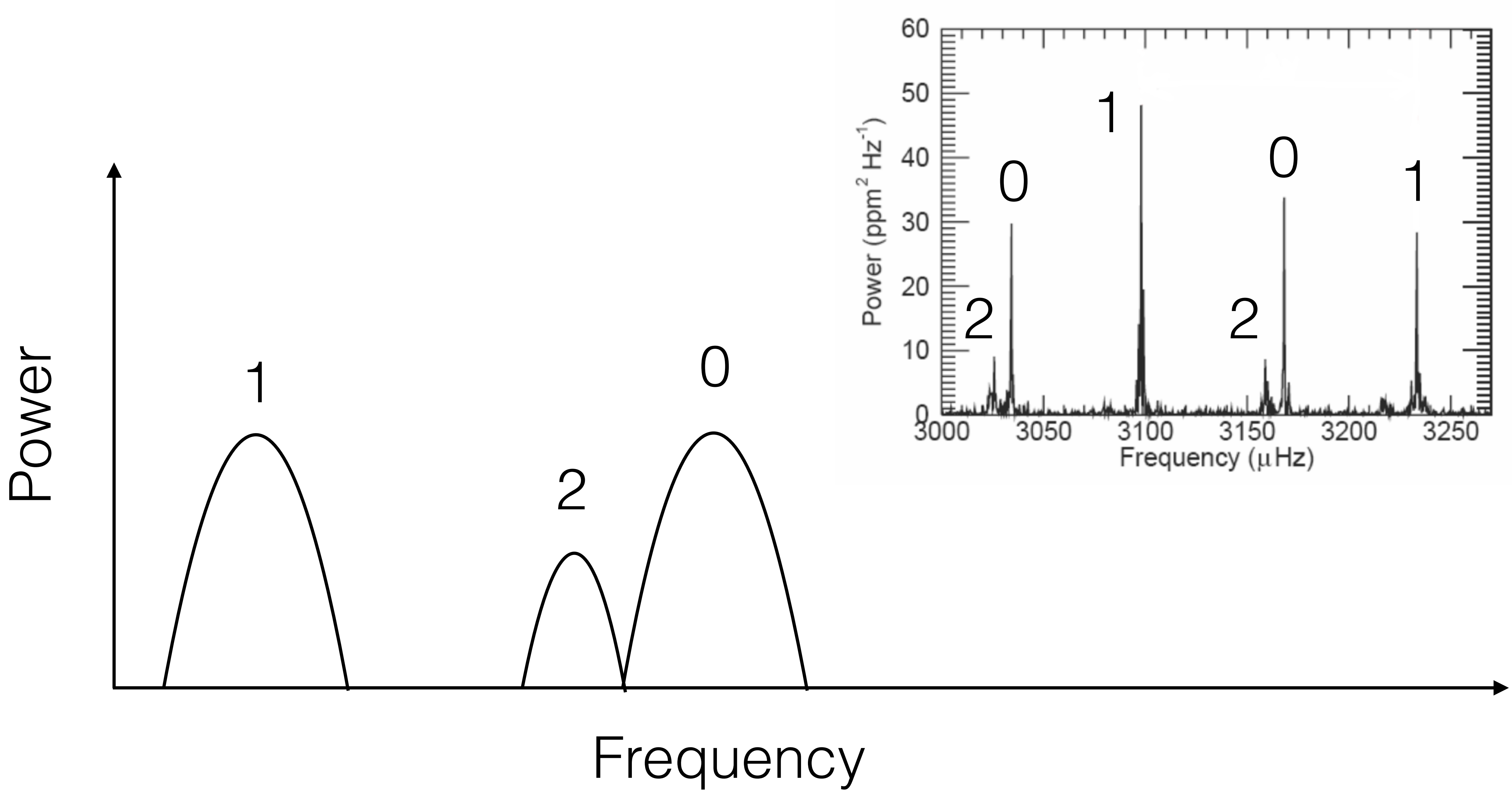
About 180 free parameters!
Computational time increases a lot

KIC 9139163

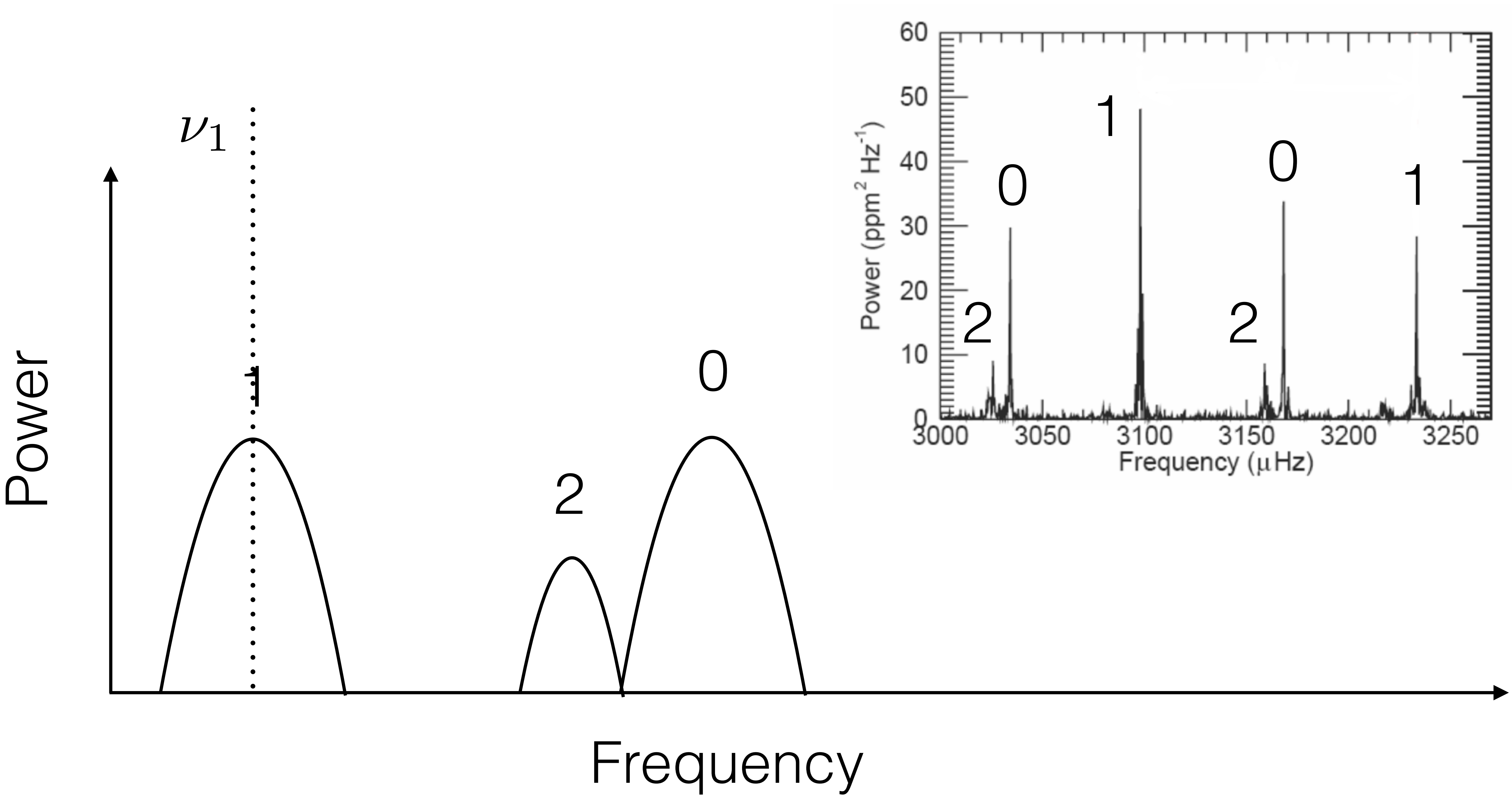




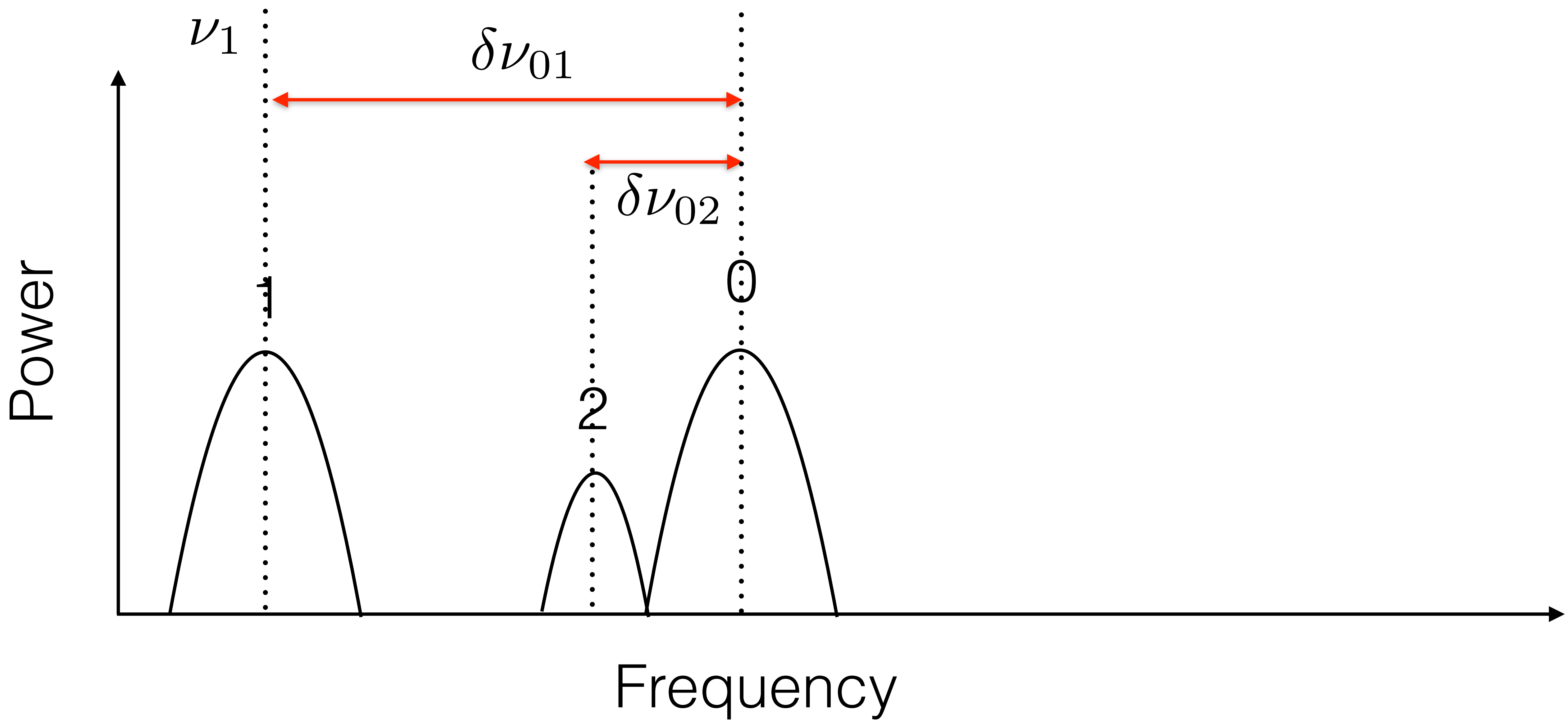
Multi-modal Model



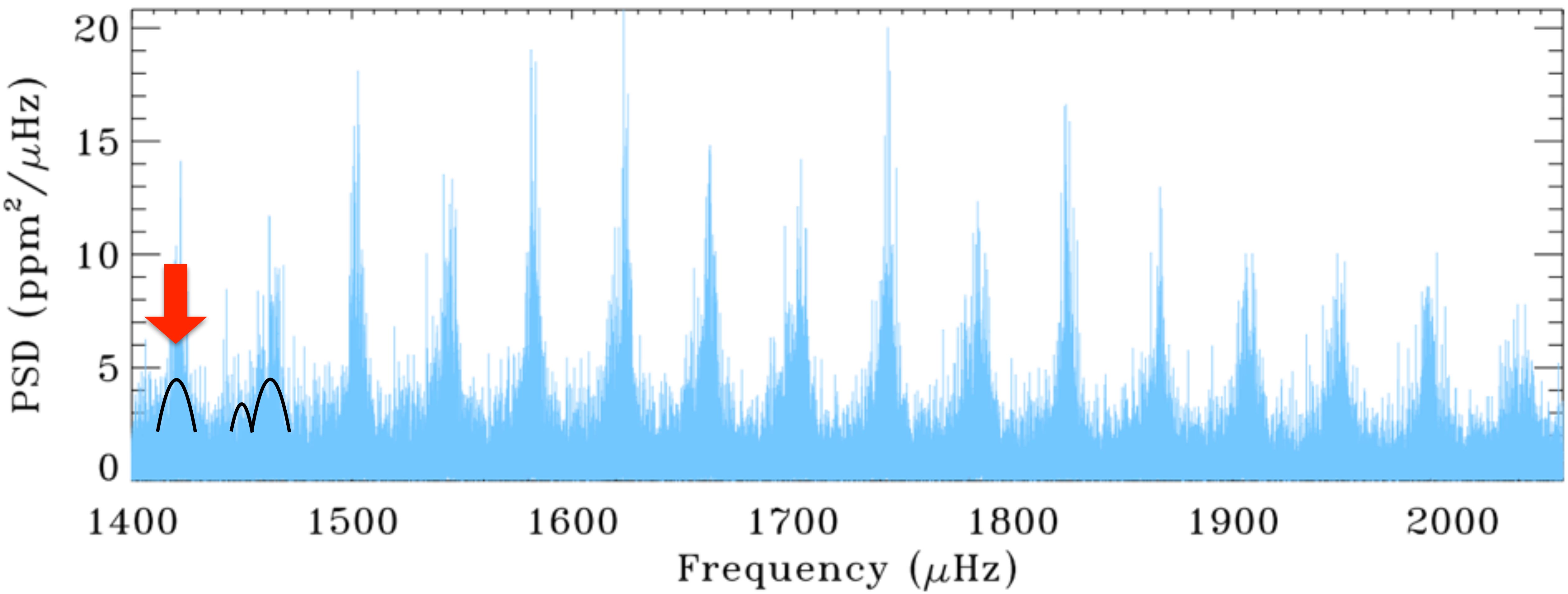
Multi-modal Model



Multi-modal Model



Multi-modal Model

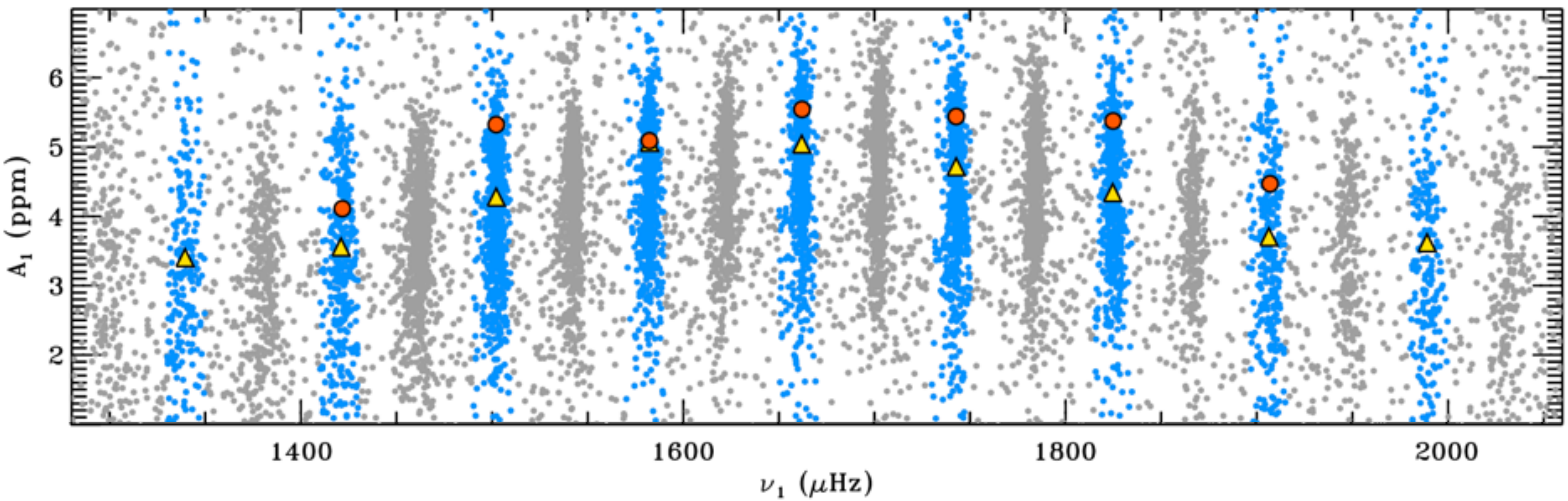


Corsaro & De Ridder 2014 A&A, 571, 71

Results

Multi-modal inference problem on 9 consecutive radial orders (27 peaks)

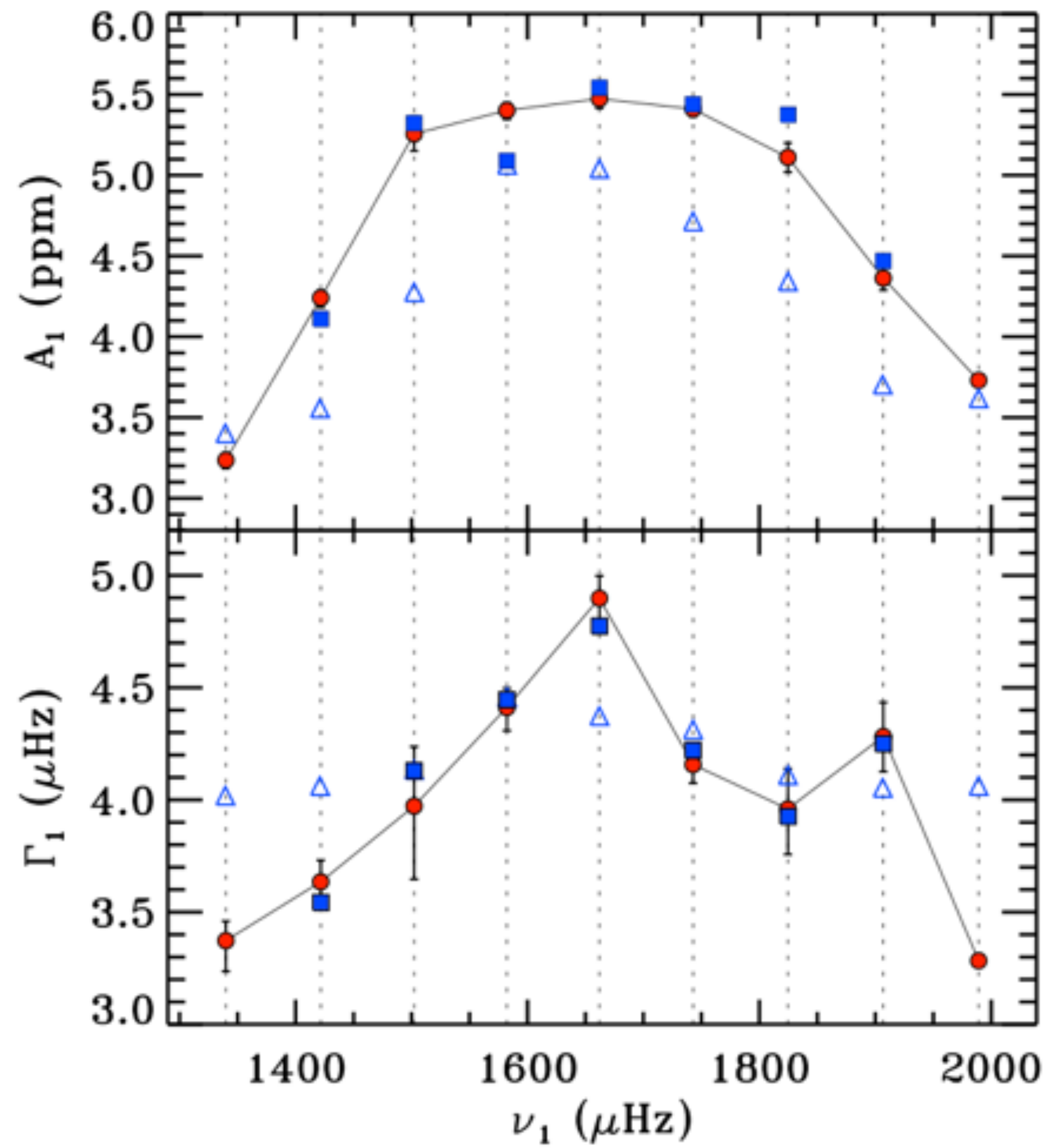
Only 9 free parameters!



Corsaro & De Ridder 2014 A&A, 571, 71

Comparison

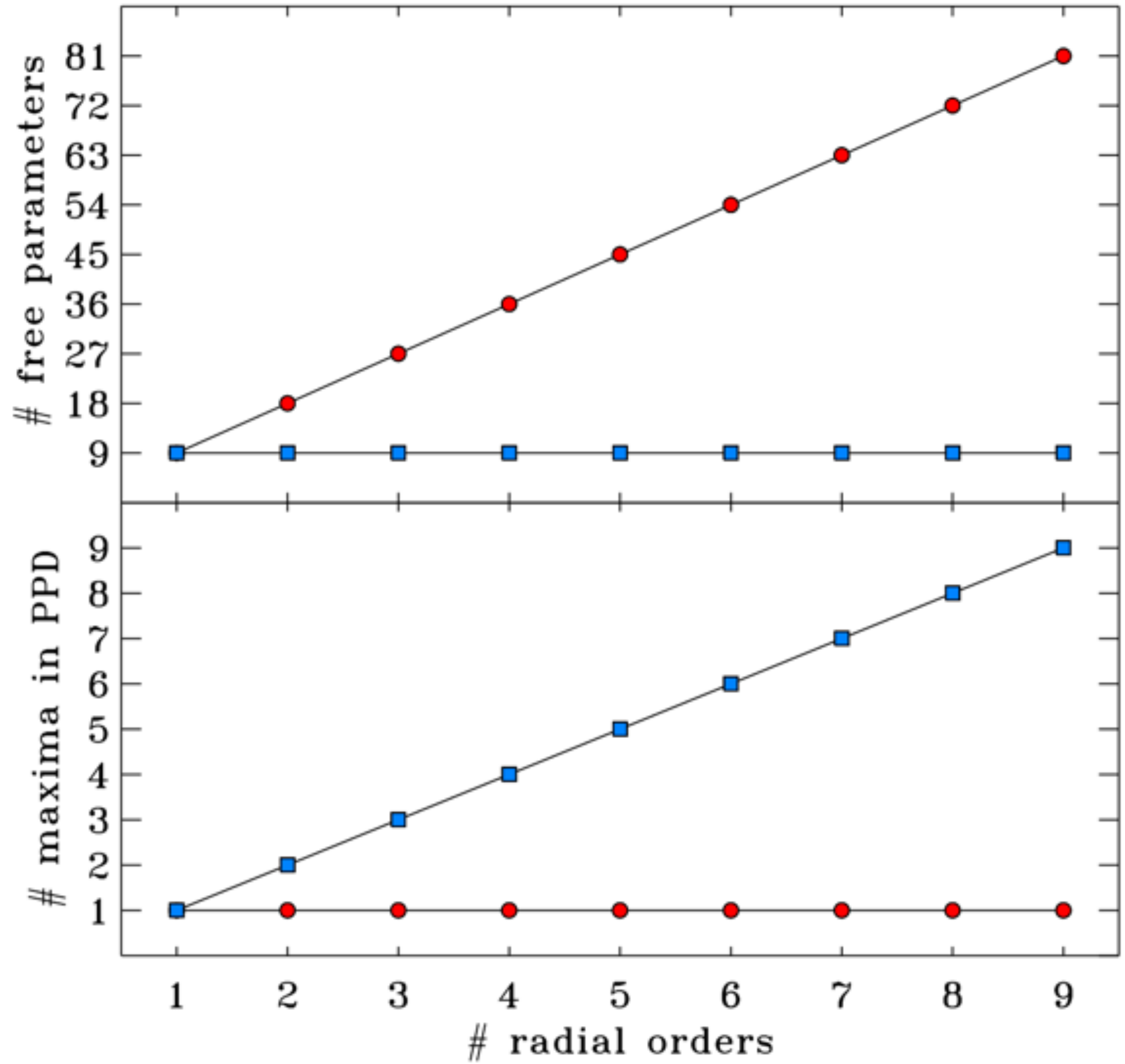
Red: uni-modal fit
Blue: multi-modal fit



Corsaro & De Ridder 2014 A&A, 571, 71

Comparison

Red: uni-modal fit
Blue: multi-modal fit



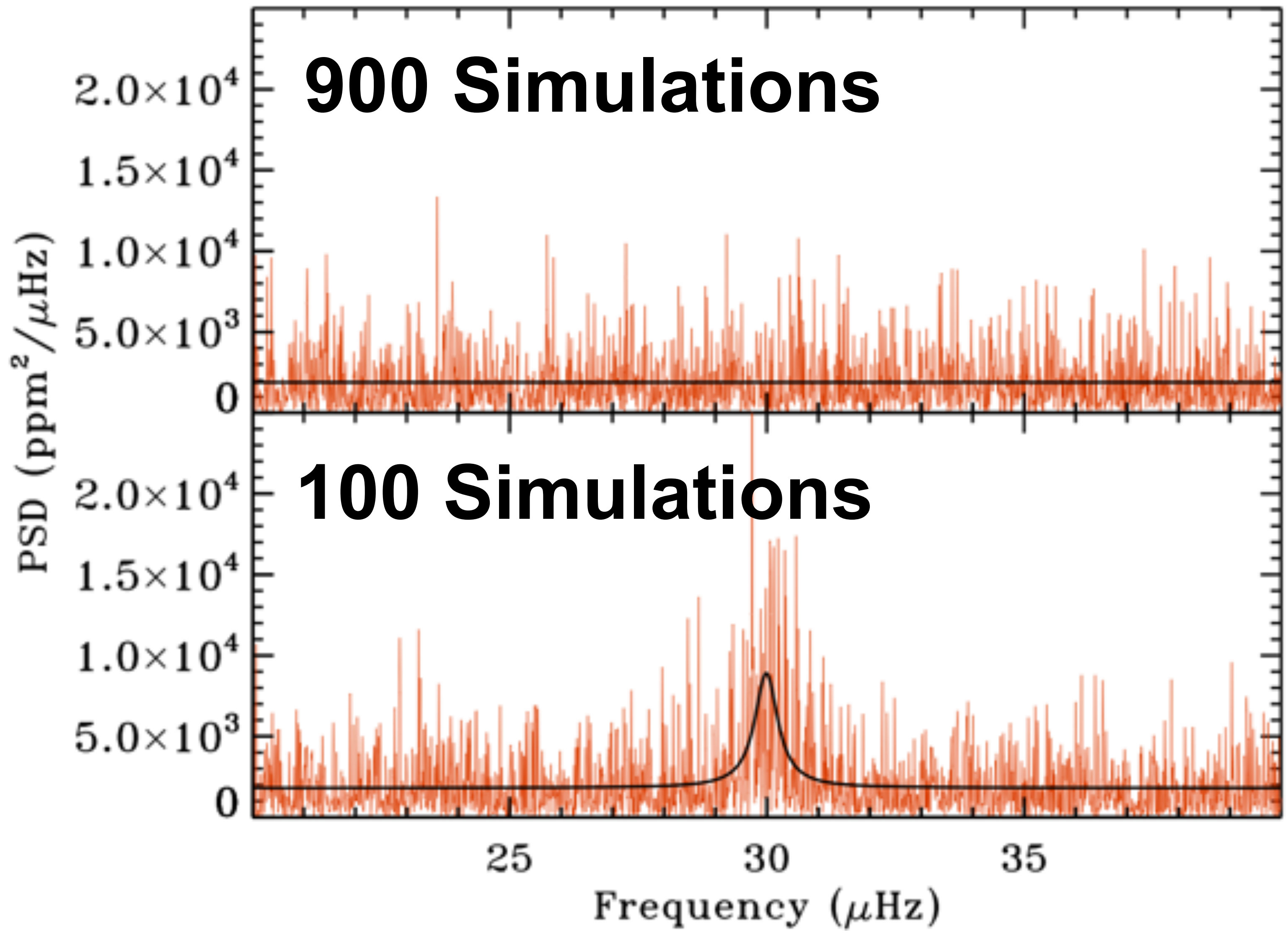
Corsaro & De Ridder 2014 A&A, 571, 71

Problem 2

Test the significance of an oscillation peak

Peak Significance Criterion

- Simulations test
- **1000** artificial chunks of PSD
- Blind search for those with a peak



Bayes' factor

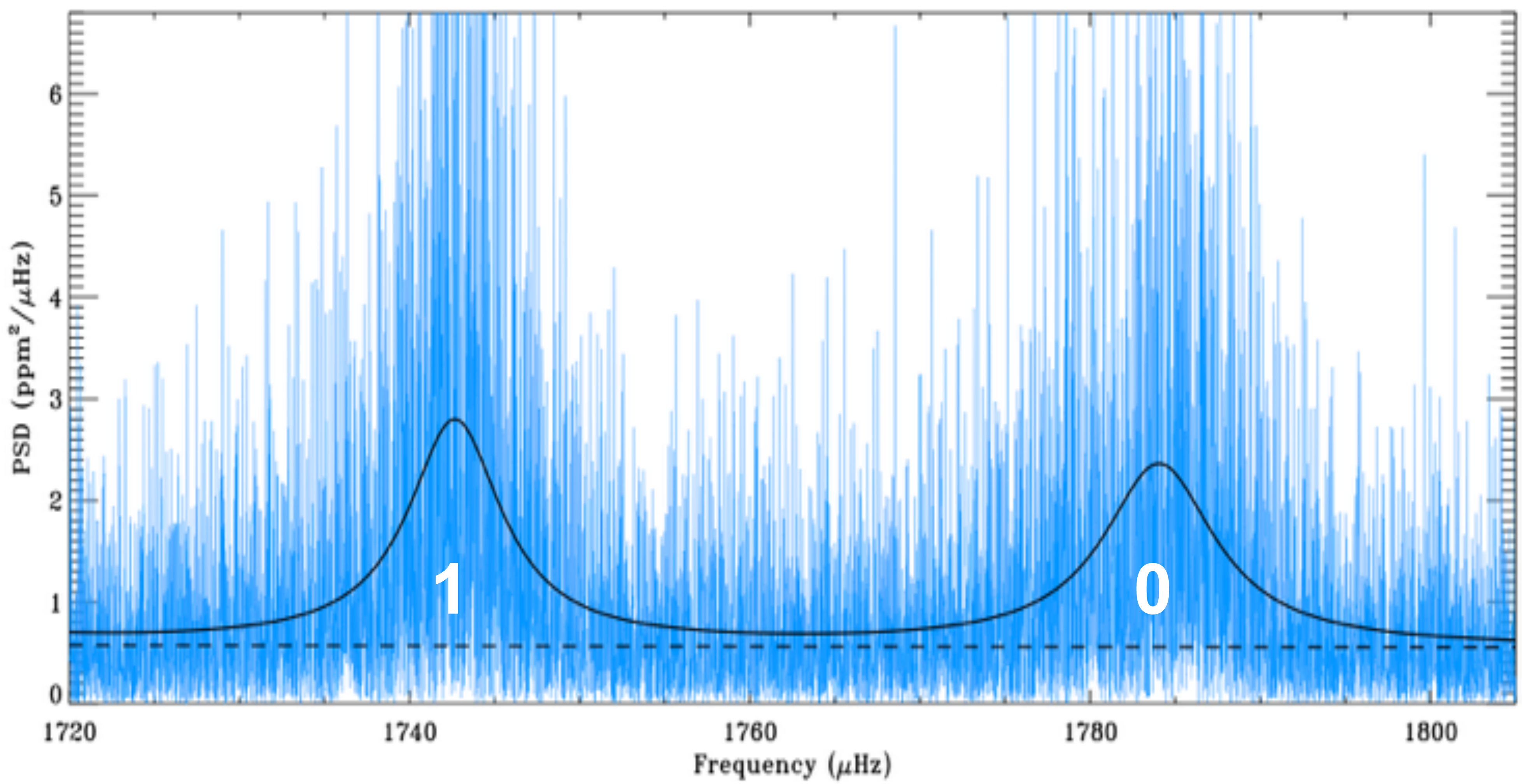
$$B_{yes,no} = \frac{\mathcal{E}_{yes}}{\mathcal{E}_{no}}$$

$$B_{yes,no} \sim 150$$

Strong Evidence (Jeffreys' scale)

All peaks found!

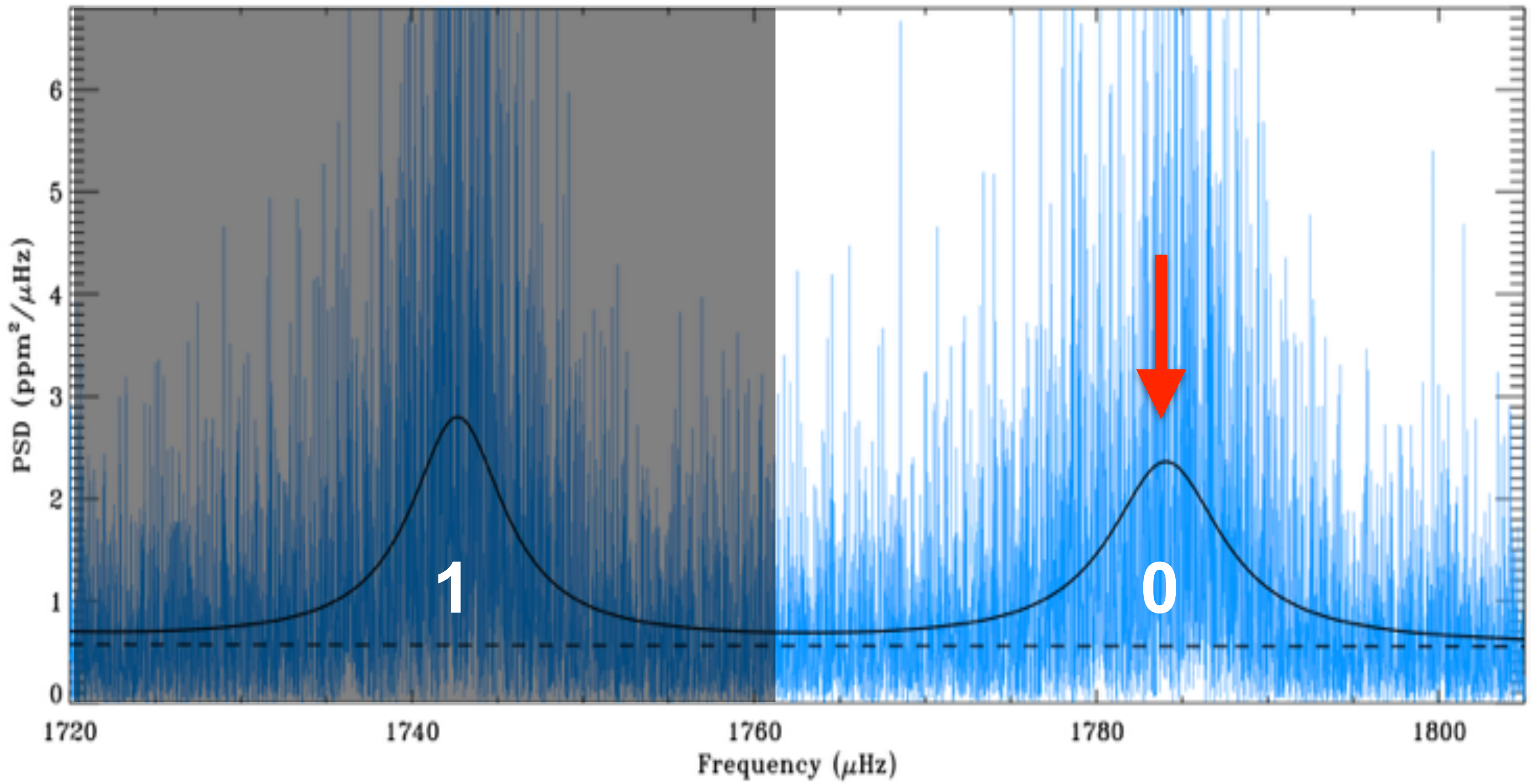
Peak significance



Peak significance

$\mathcal{M}_{\ell=0}$ **Only ℓ = 0**

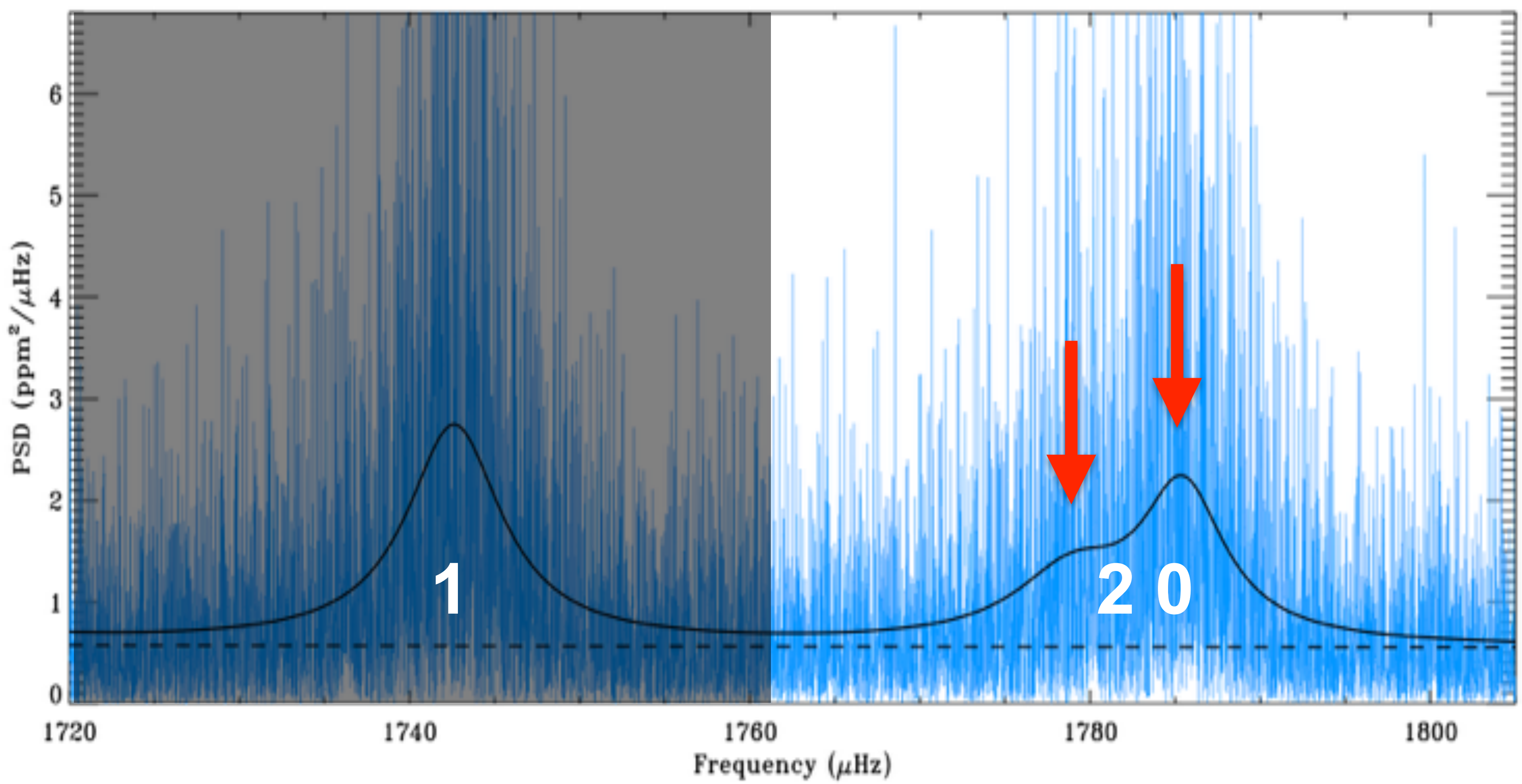
$\mathcal{E}_{\ell=0}$ **Bayesian Evidence**



Peak significance

$\mathcal{M}_{\ell=2}$ **Both ℓ = 2 and ℓ = 0**

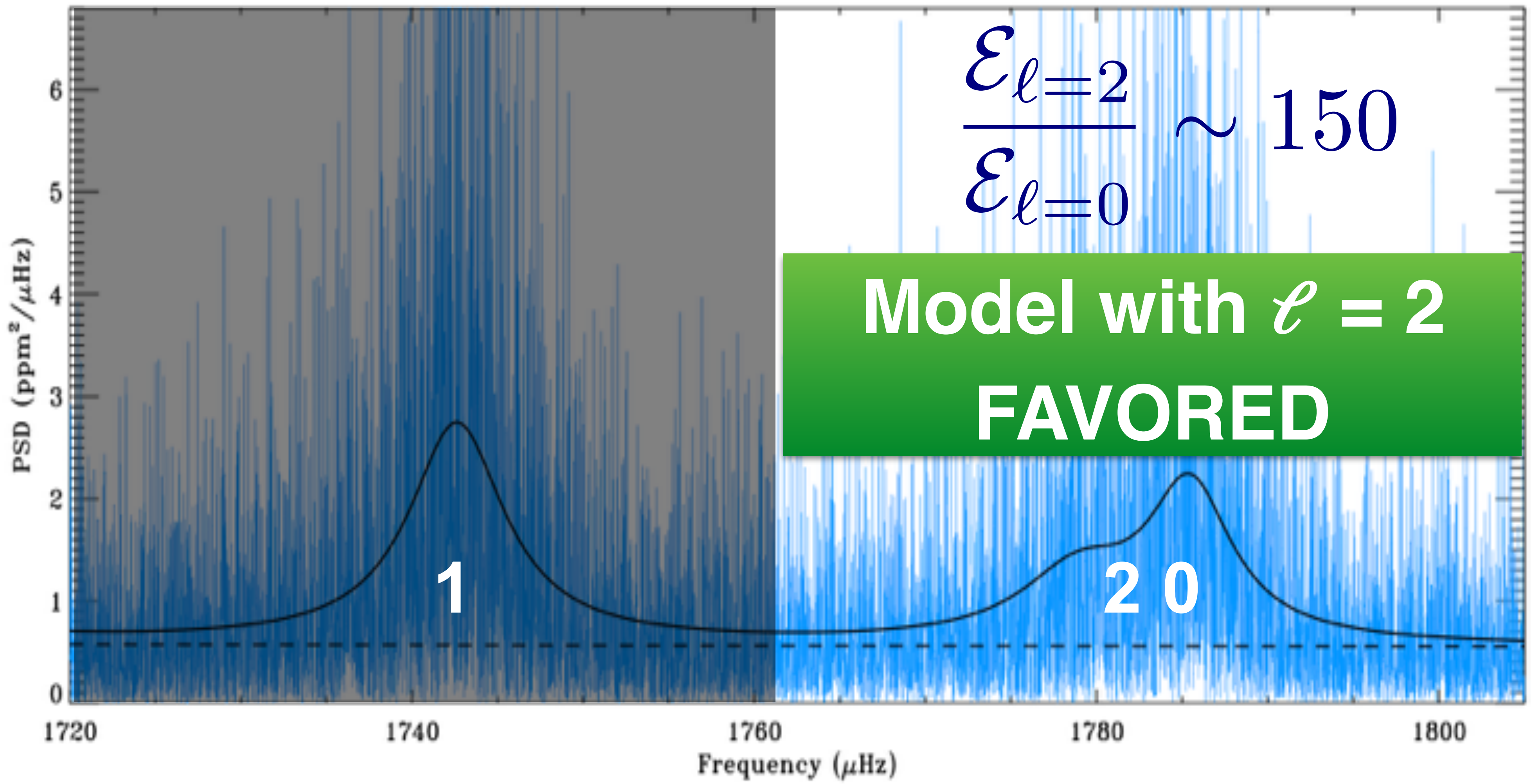
$\mathcal{E}_{\ell=2}$ **Bayesian Evidence**



Peak significance

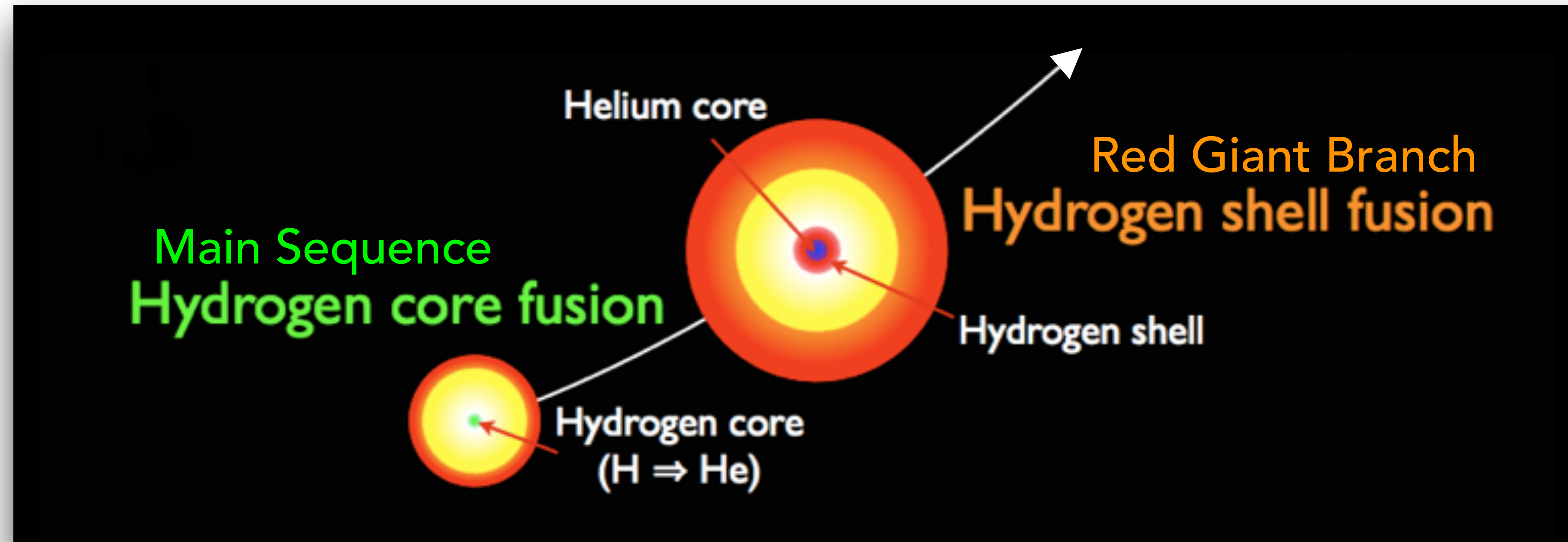
$\mathcal{M}_{\ell=2}$ **Both ℓ = 2 and ℓ = 0**

$\mathcal{E}_{\ell=2}$ **Bayesian Evidence**

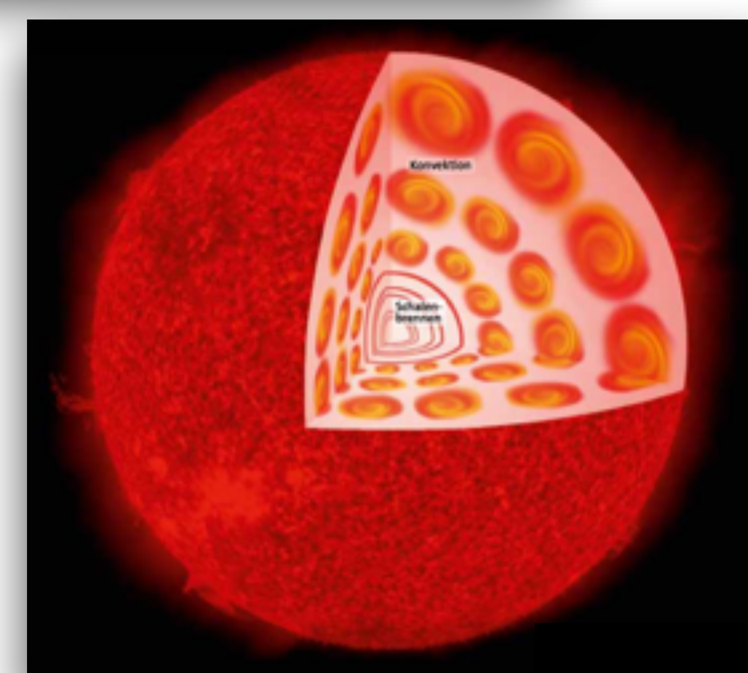


Oscillations in red giant stars

RGB oscillations

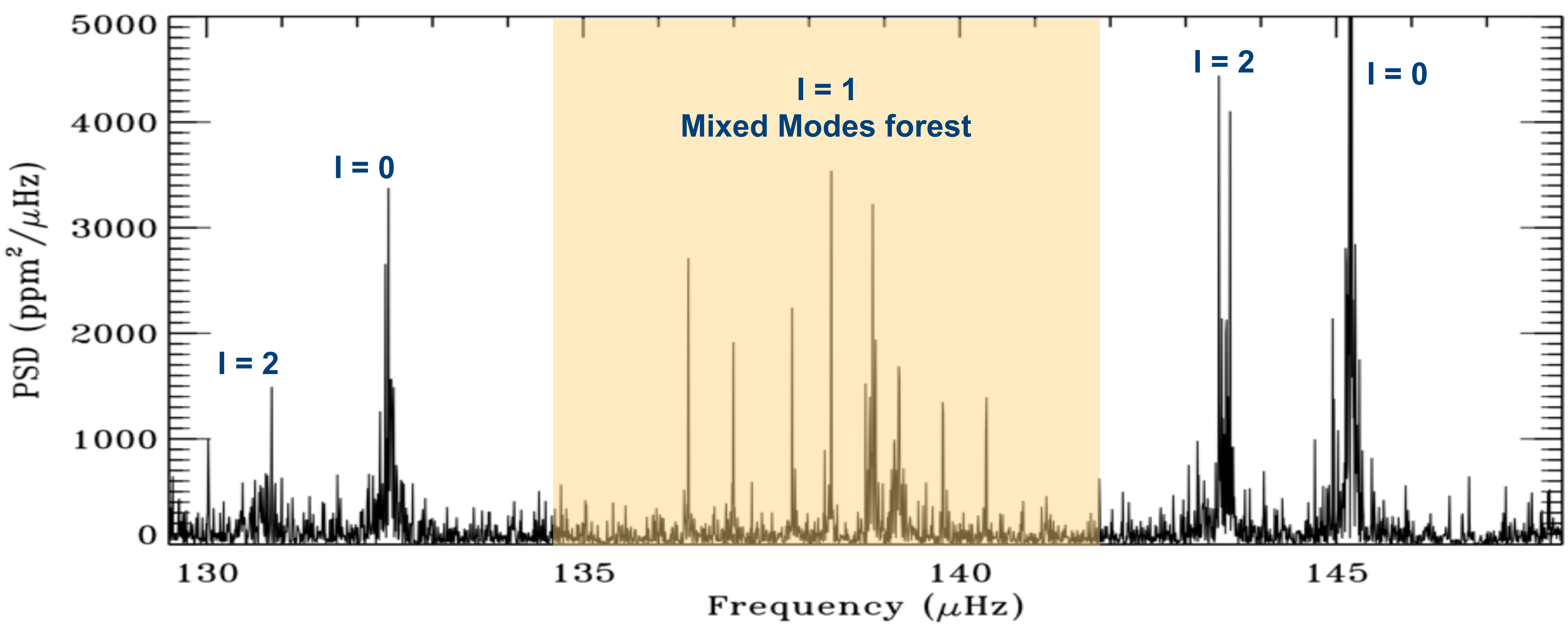


© Thomas Kallinger



p modes couple with gravity modes (*g* modes) from radiative interior (mixed modes)

RGB oscillations



Many oscillation modes per star (up to about 100)!

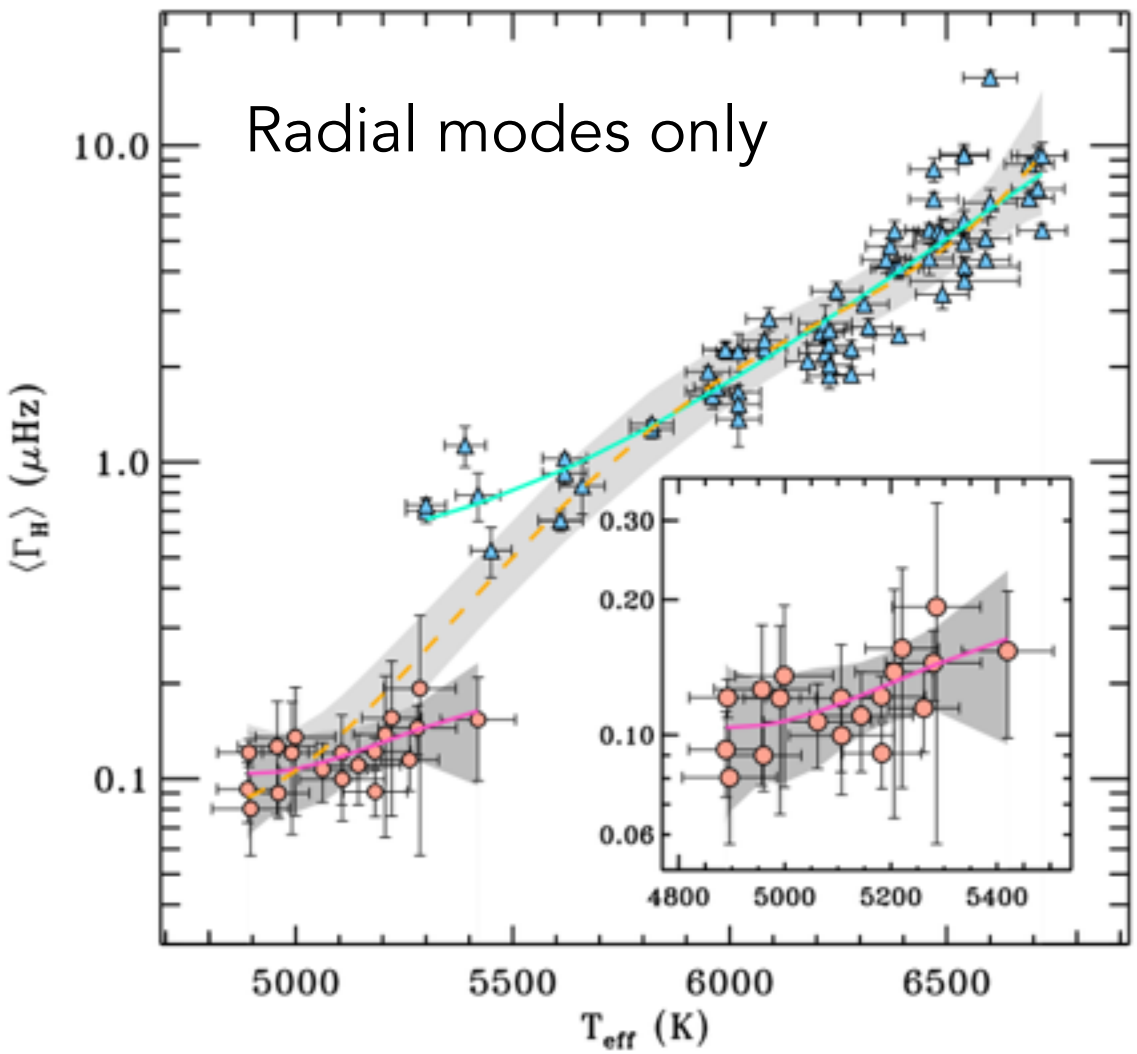
Results on 19 RGB stars

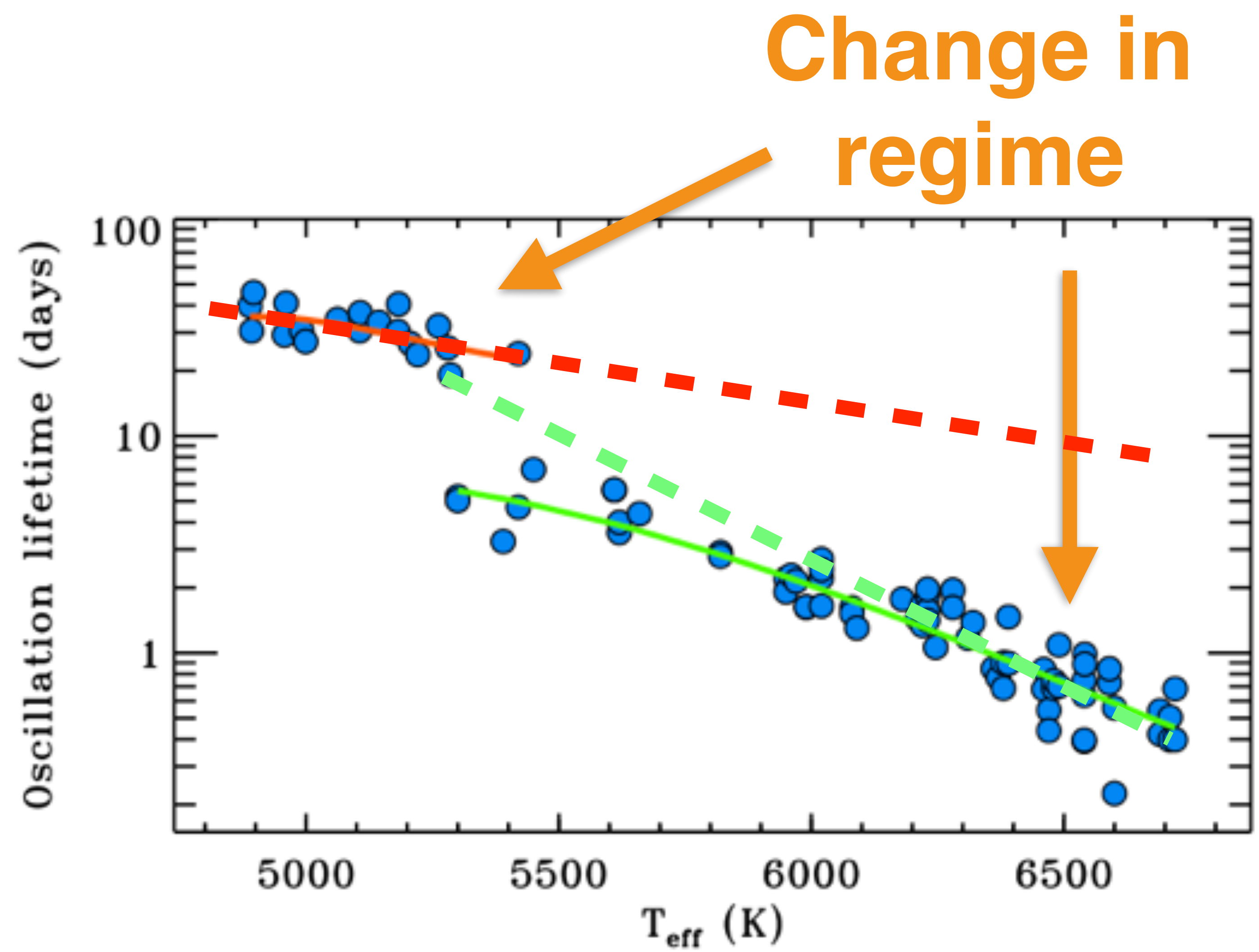
Corsaro, De Ridder, García 2015 A&A, 579, 83

- **1618** oscillation modes extracted
- **612** peaks tested (38%) with Bayesian model comparison
- **380** peaks detected (62% of tested peaks)
- Internal rotation detected in **14** stars

Mode lifetimes

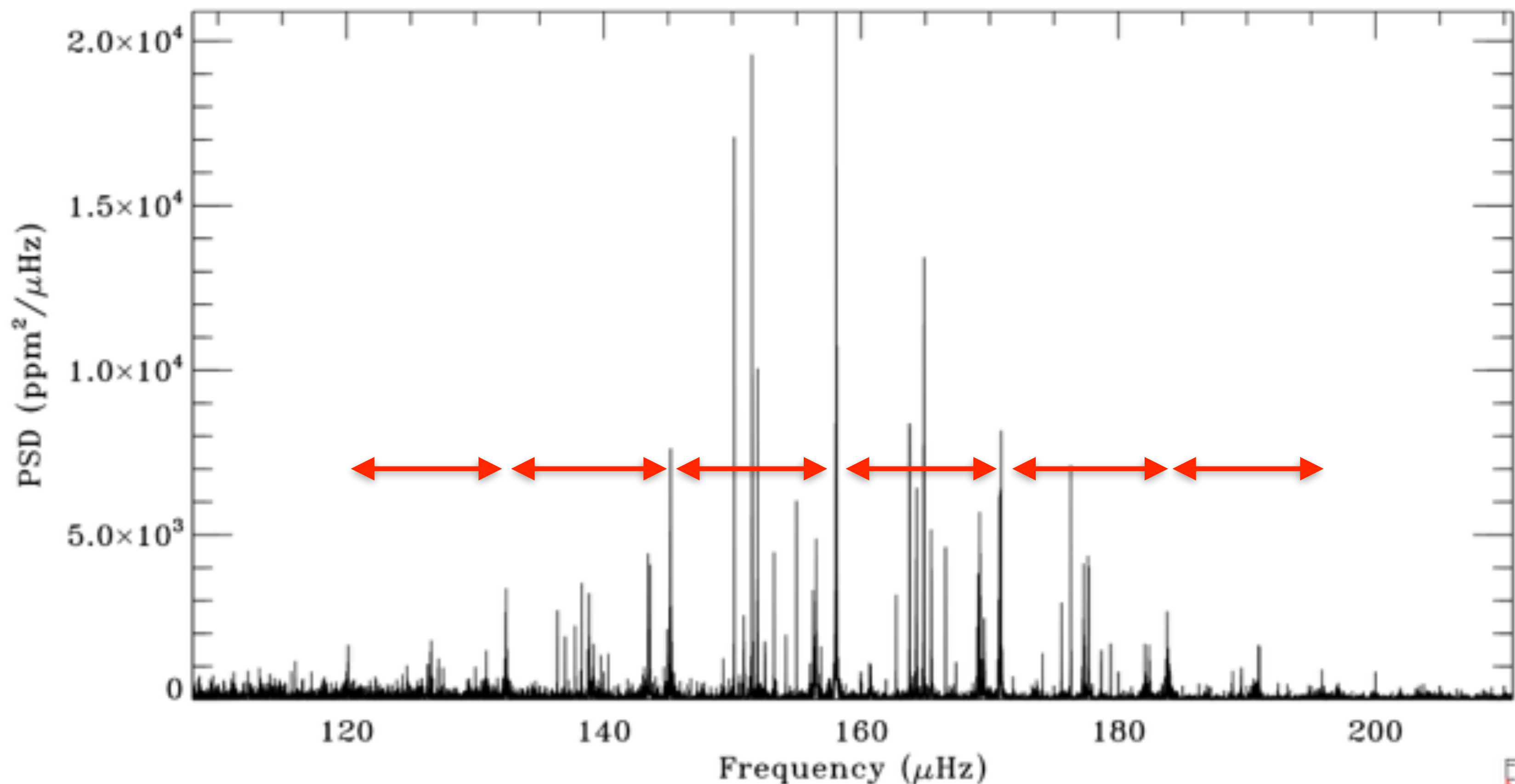
Corsaro, De Ridder, García 2015 A&A, 579, 83



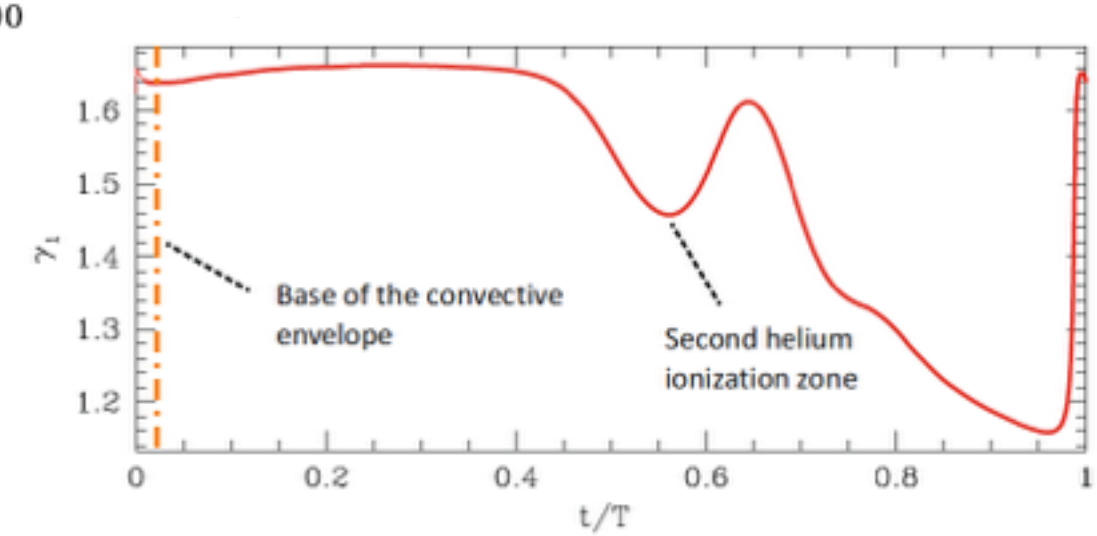


Acoustic glitches

Corsaro, De Ridder, García 2015 A&A, 578, 76

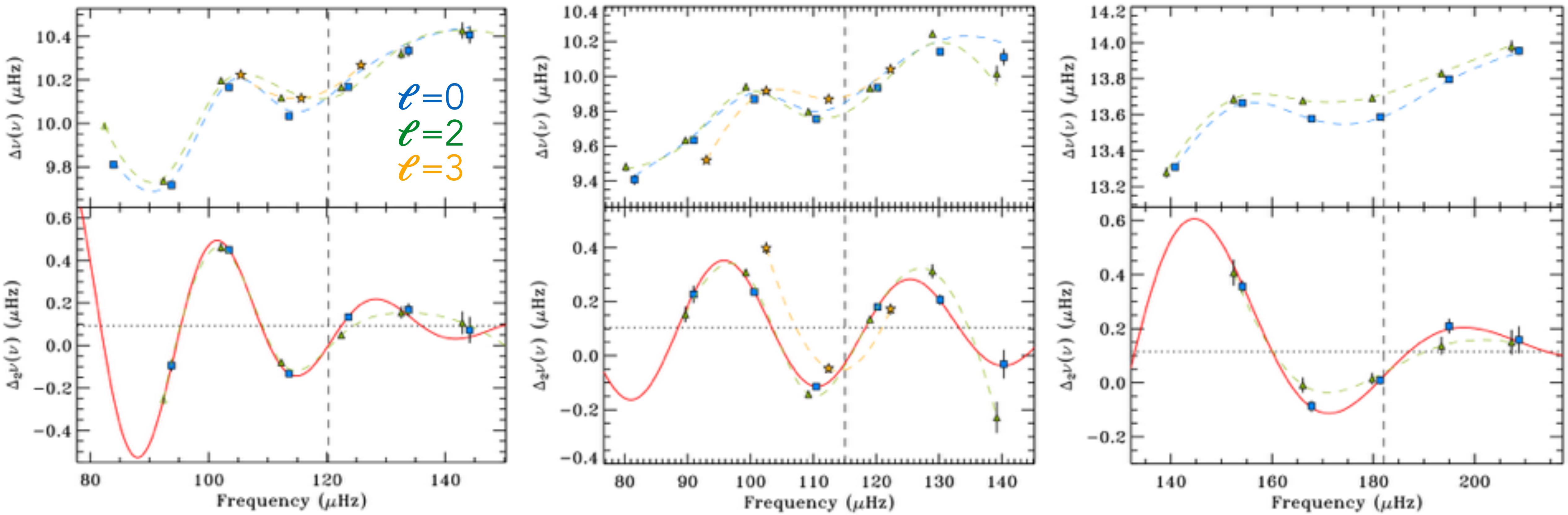


Regions of sharp structure variation inside the star induce oscillatory behavior in the mode frequencies



Signature of HeII zone

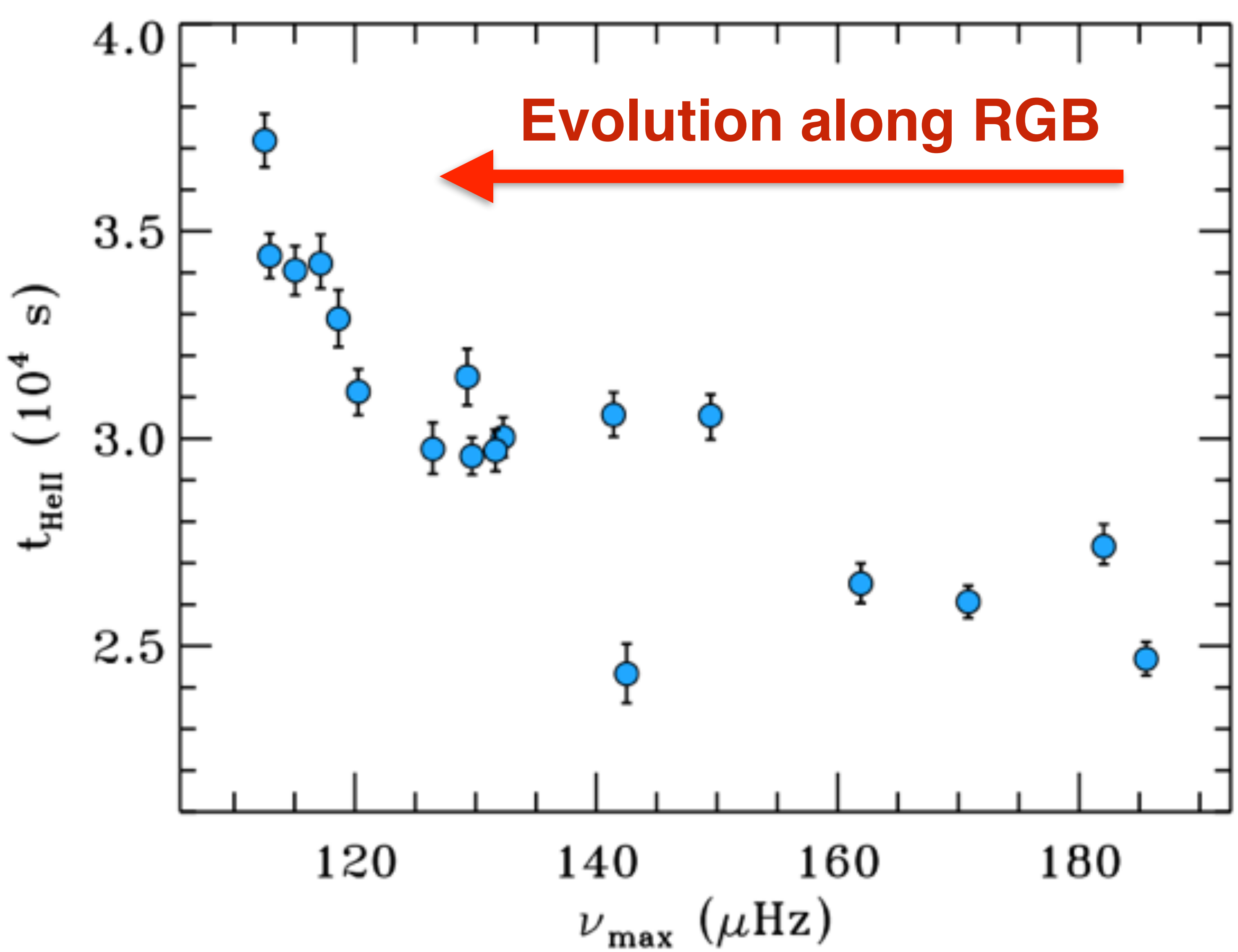
Corsaro, De Ridder, García 2015 A&A, 578, 76



- The position of HeII zone is constrained up to 2% precision!
- Amplitudes up to 6%, can give estimate of He abundance in convective envelope

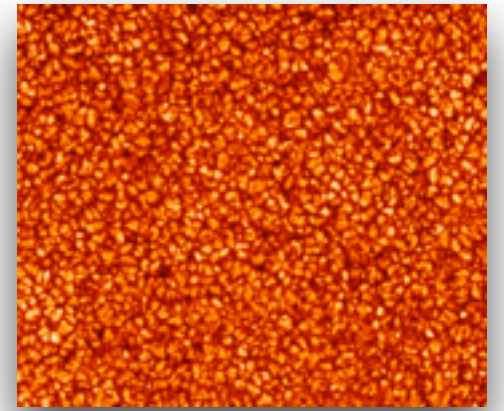
Evolution of HeII zone

Corsaro, De Ridder, García 2015 A&A, 578, 76



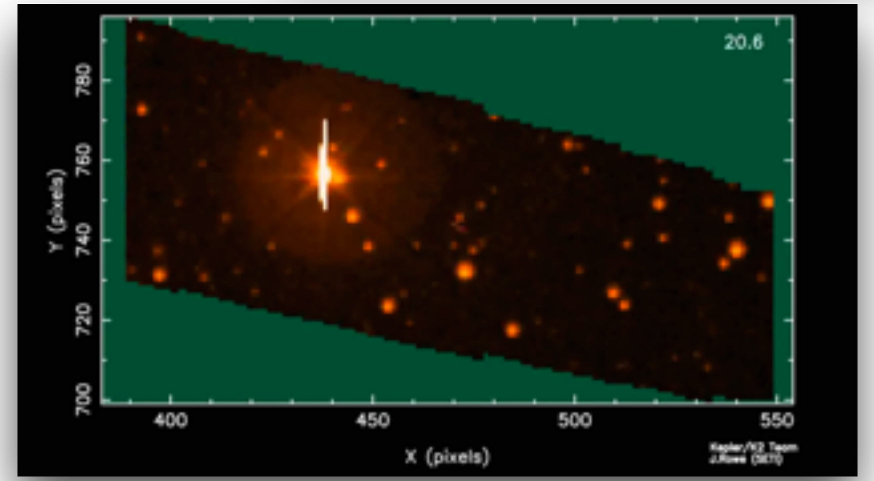
Ongoing research

- Analysis of granulation and oscillation properties of the Sun with GOLF & VIRGO + correlation with magnetic activity

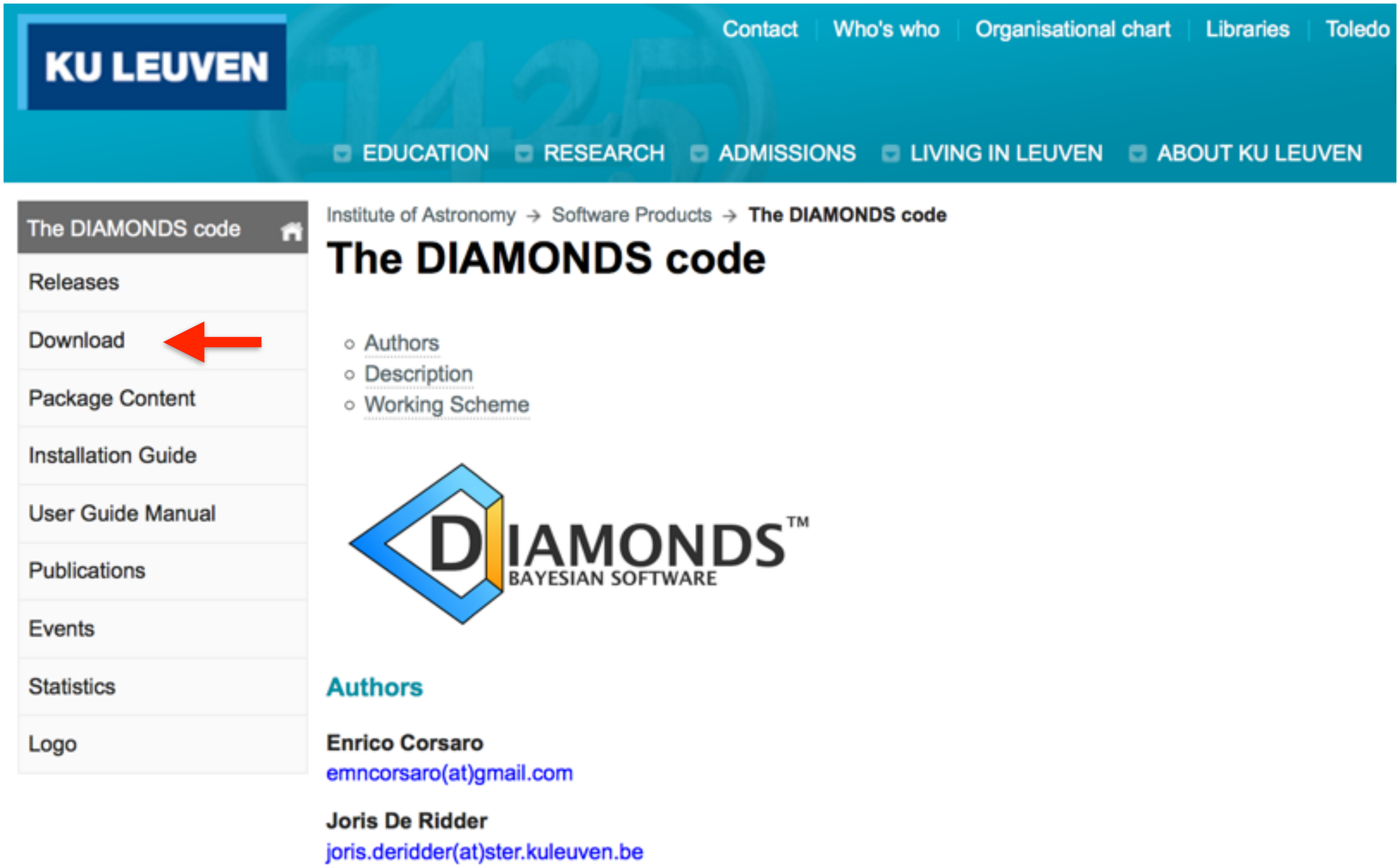


- Full characterization of red giant stars in NASA Kepler open clusters NGC 6791, NGC 6811, NGC 6819

- Analysis of solar oscillations reflected from Neptune’s atmosphere, observed by K2

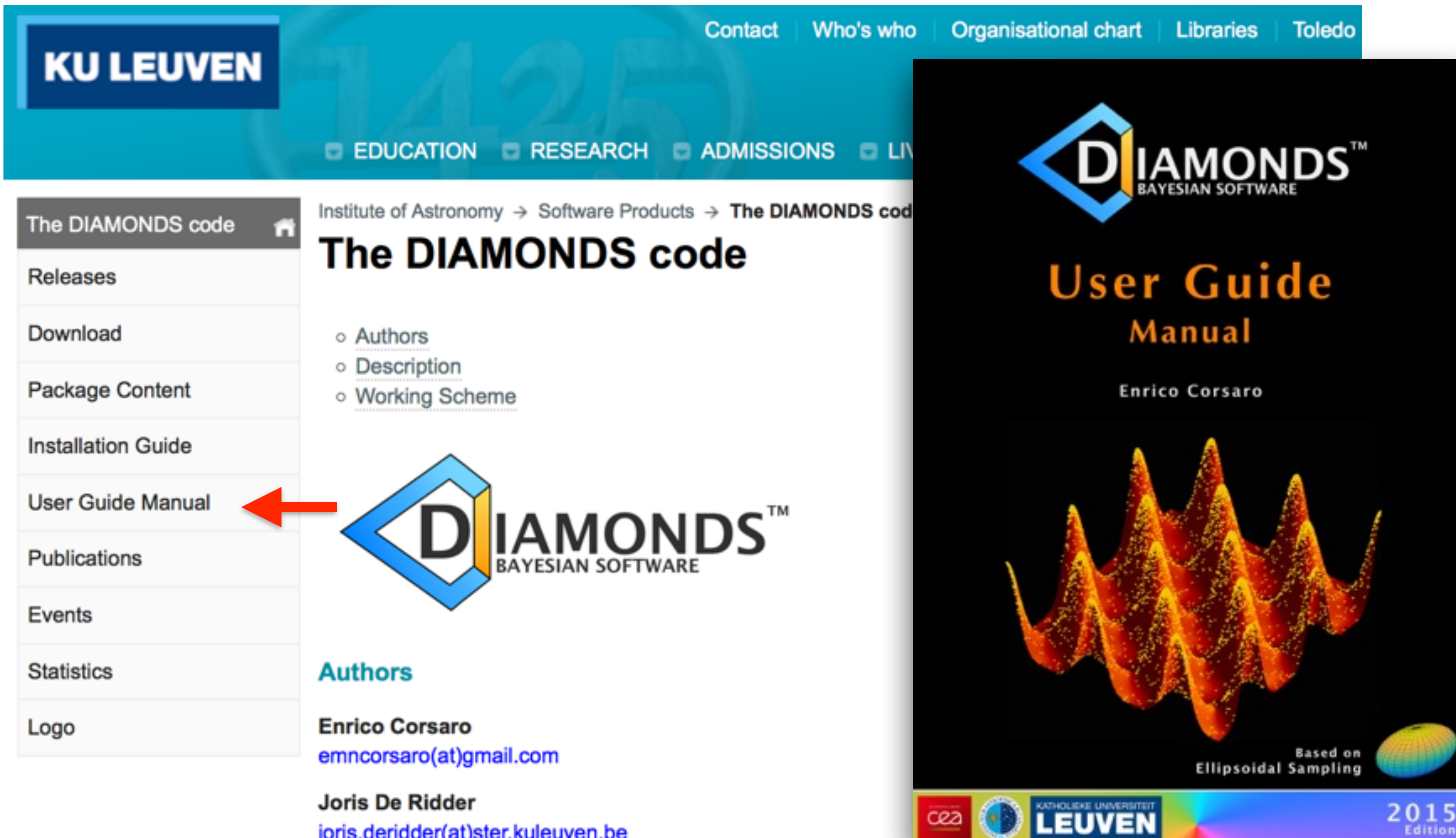


Download DIAMONDS™ BAYESIAN SOFTWARE



https://fys.kuleuven.be/ster/Software/Diamonds/

Download DIAMONDS™
BAYESIAN SOFTWARE



https://fys.kuleuven.be/ster/Software/Diamonds/

Thank you!