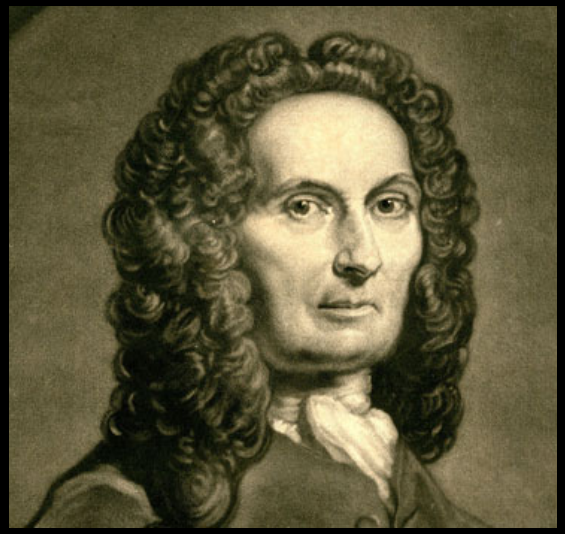
Spectral fitting methods

Workshop at the
Max Planck Institute for extraterrestrial Physics
24-25.09.2019
organised by
Johannes Buchner, J Michael Burgess, Joern Wilms

Welcome!

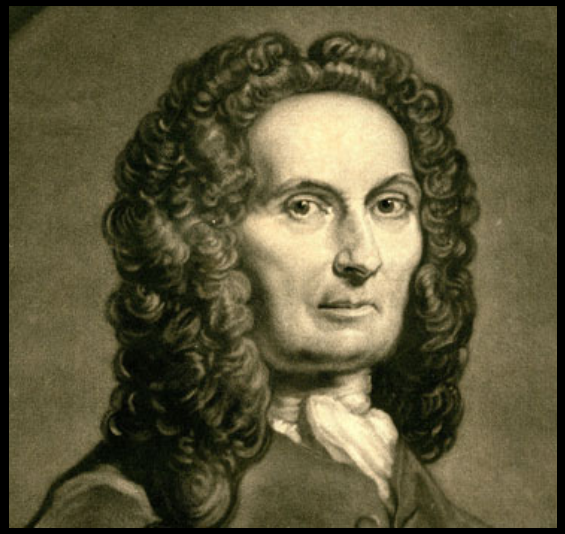


Probabilities of discrete events

Probabilities of hypotheses

Scientific application

Welcome!



Probabilities of discrete events

Poisson distribution
(by Abraham de Moivre)

Probabilities of hypotheses

Bayesian inference
(by Pierre-Simon Laplace)

Scientific application

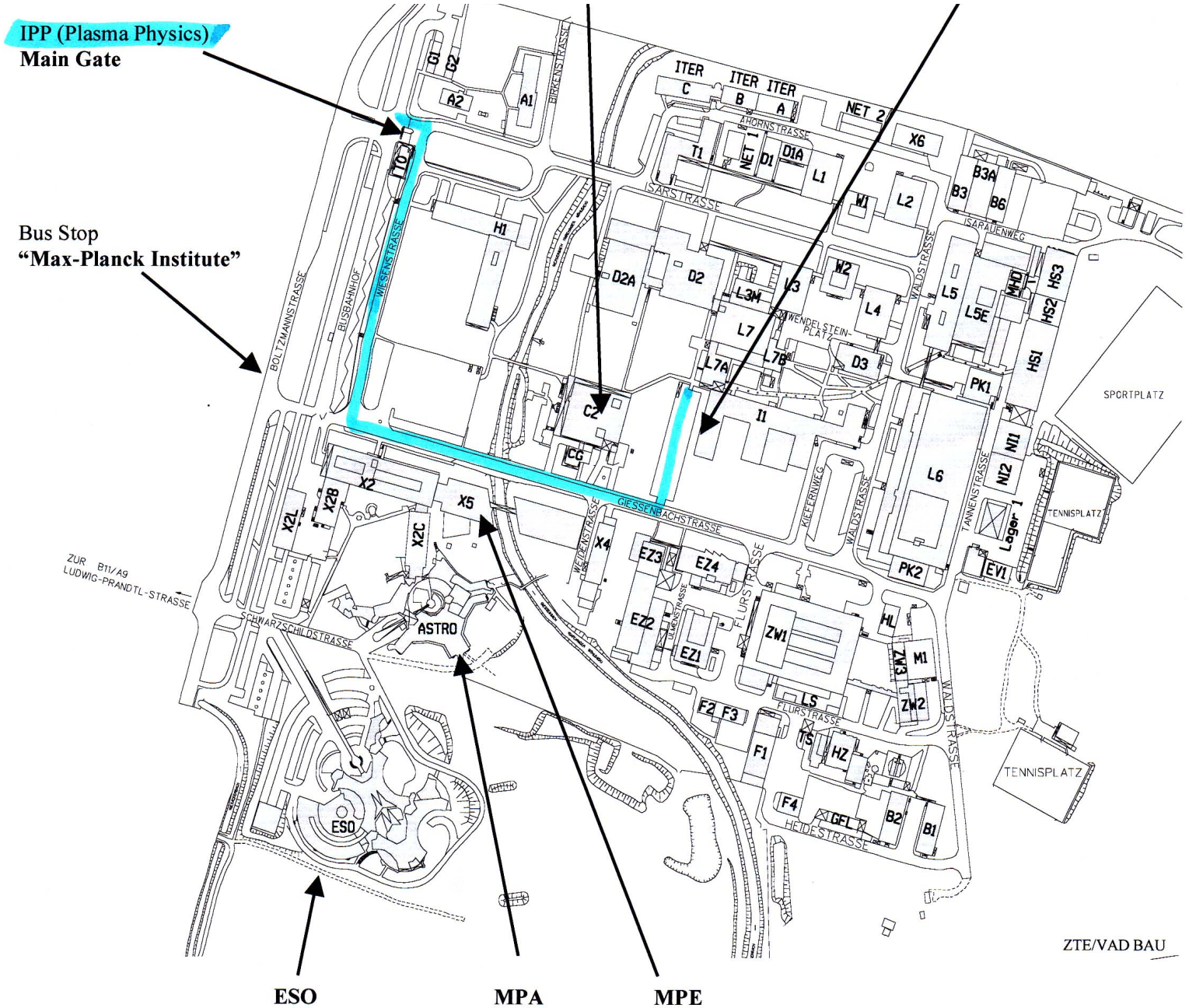
Horse kicks
(by Ladislaus Bortkiewicz)

Stigler's law of eponymy
(Robert K. Merton)

Welcome!

- Practical information
  - organisers
  - wifi
  - local information
  - rough agenda & goal
  - dinner
- First content block

Local information



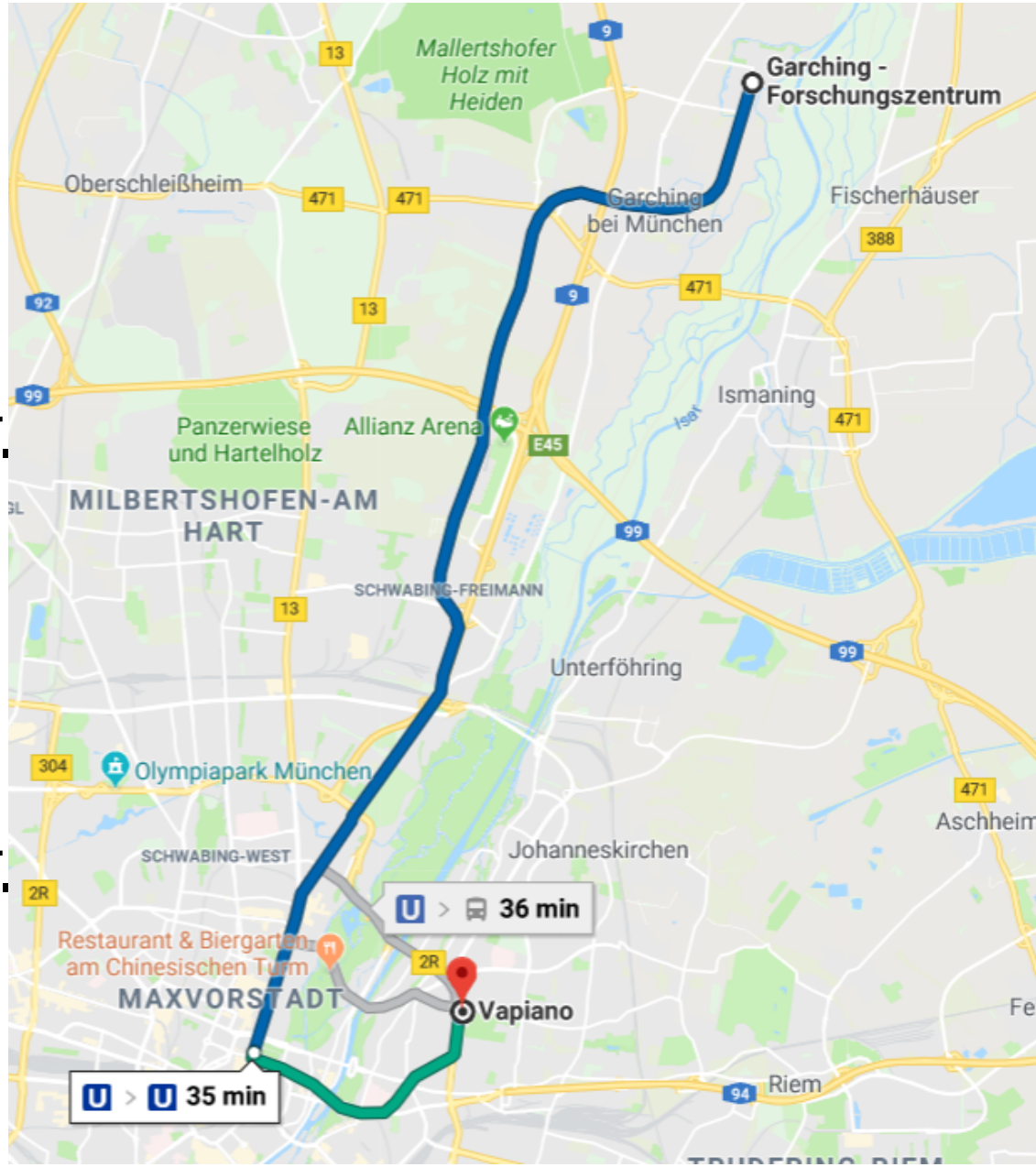
Agenda

- Morning
  - Measurement process & statistics involved
  - Background treatment
  - Local best fits
- Lunch Cantine
- Afternoon
  - Global fits & probability distributions
  - Model comparison
  - Coffee 15:30
  - Combining information
  - Beyond X-ray spectra
  - Discussion & Questions
- End: 17:30. Dinner 19:00
- Morning
  - Extended sources, calibration
  - Discussion & Questions
  - Poisson knowledge & practical pointers

Goal:
what methods exist
what are their benefits & limitations
what to pay attention to

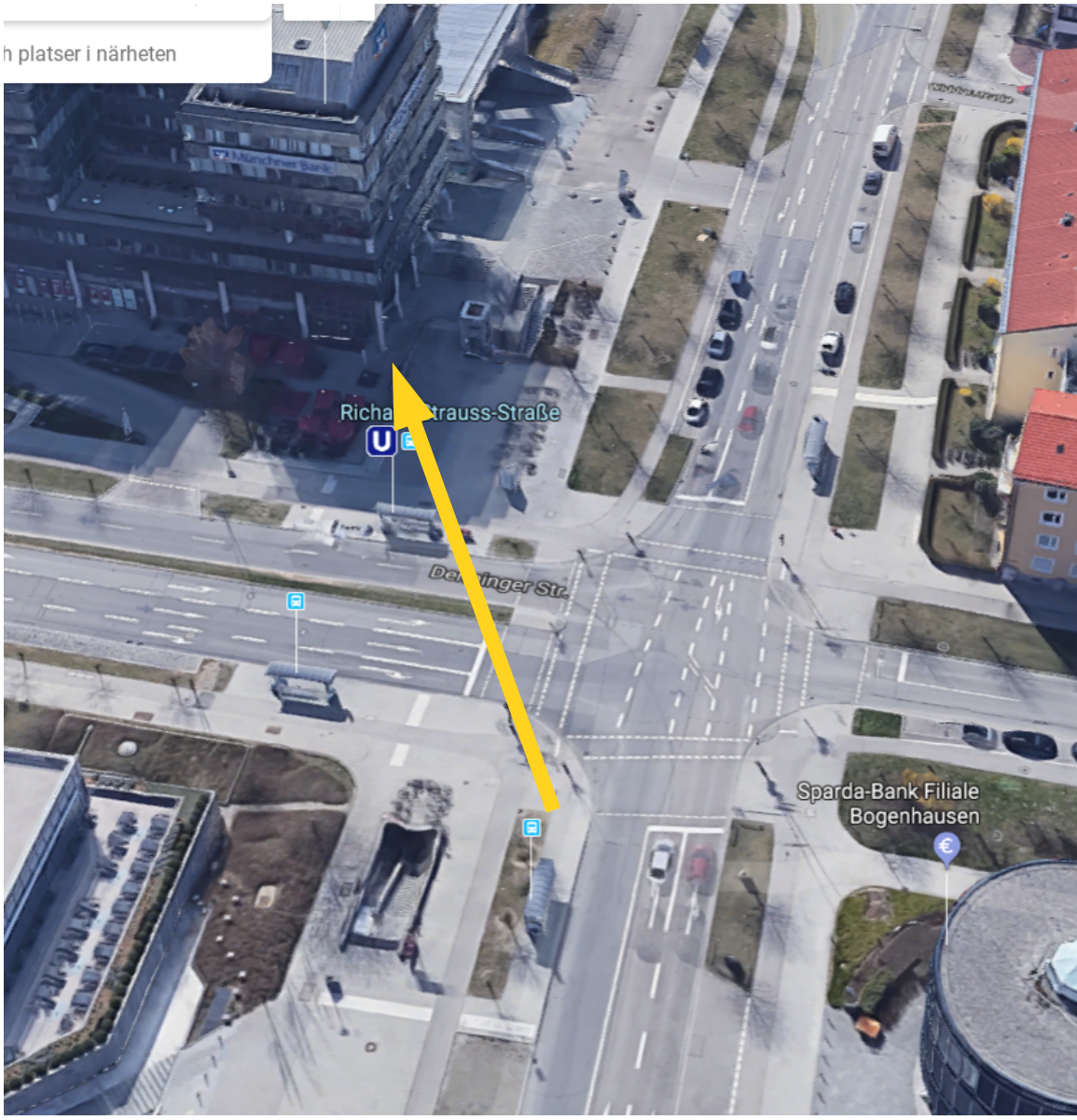
Dinner today 19:00

- U6+Bus 59
  - Dietlindenstr.
  - Bus 59 Giesing
  - Richard-Strauss-str.
- U6+U4
  - Odeonsplatz
  - U4 Arabellapark
  - Richard-Strauss-str.



Dinner today 19:00

- U6+Bus 59
  - Dietlindenstr.
  - Bus 59 Giesing
  - Richard-Strauss-st
- U6+U4
  - Odeonsplatz
  - U4 Arabellapark
  - Richard-Strauss-st



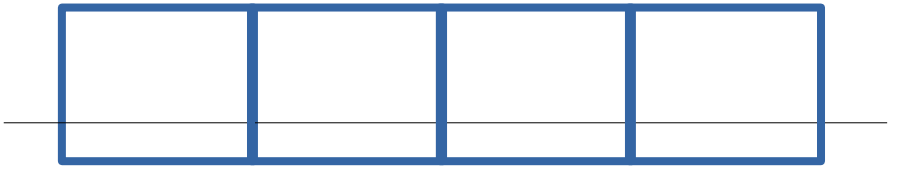
First block

- Overview & Introduction
  - Measurement process
  - Background & source regions
  - Linear algebra approximation
  - Likelihood & statistics
- after: background
- after: fitting

What Counts

Single spectral bin

- Bernoulli coin flip
  - k=0 (p)
  - k=1 (1-p)

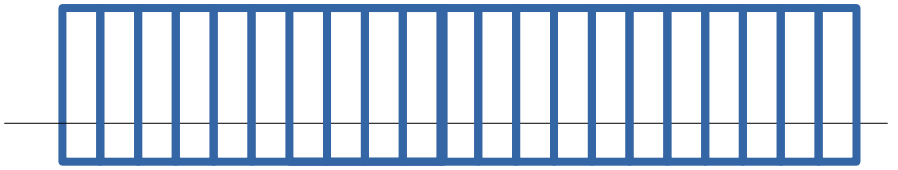


- Binomial
  - n tries, first k successful
  - $P=p^k (1-p)^{(n-k)}$

$$P(X = k) = \binom{n}{k} (p)^k (1-p)^{n-k}$$

- Poisson
  - $n \to \inf$ but $pn=\lambda$

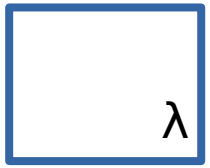
rule of thumb: if n>20 and p<0.05
n>100 and np<10



$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

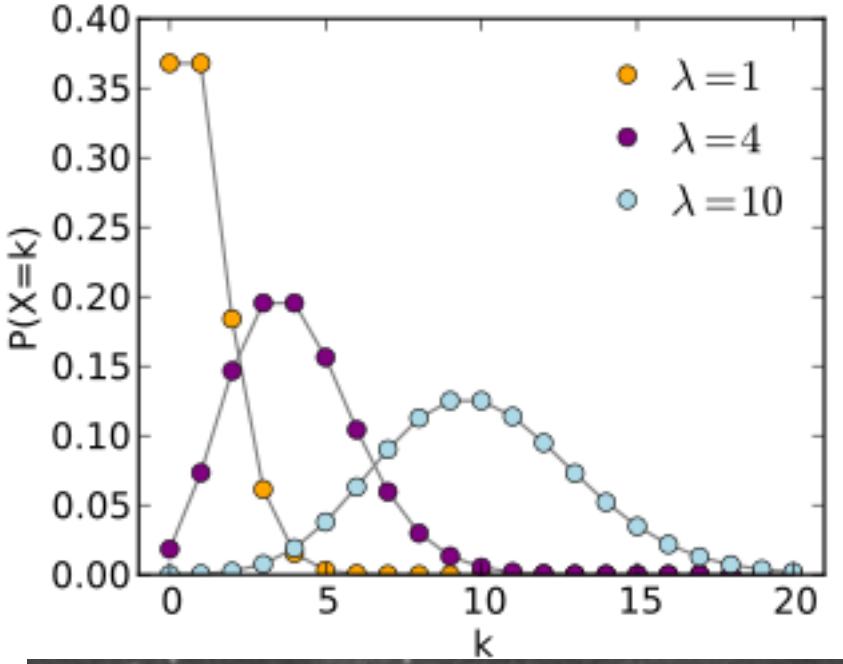
Single spectral bin

- Poisson



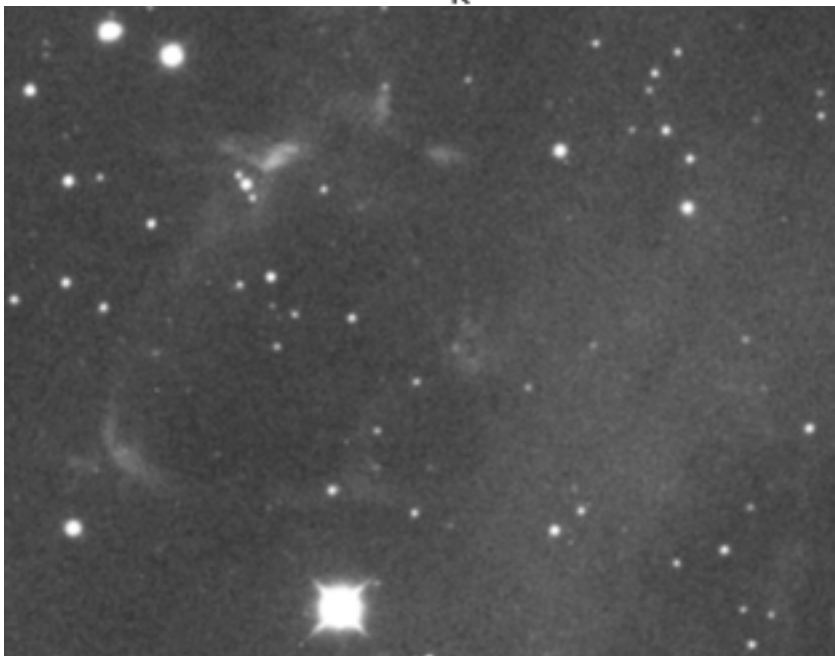
  - k: integer
  - λ: real (mean&variance)
  - Asymmetric
  - Integer
  - Positive

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$



- Scaling — shape changes
- Addition — (Poisson distribution) Variability!
- Subtraction — (Skellam distribution)

Samples
Electronics (shot noise)
Photon counting (Poisson noise)



Single spectral bin

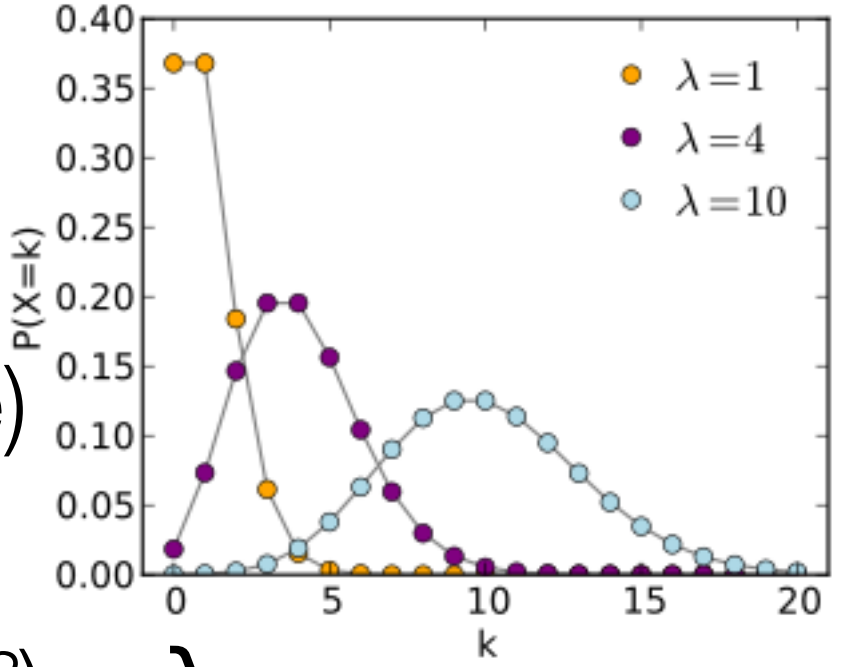
- Poisson

  λ

  $$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

  – k: integer

  – **λ**: real (mean&variance)



- Gaussian

  – Mean (μ) & variance (σ^2) = **λ**

  – Mean (μ) & variance (σ^2) = k

  – real, can be negative

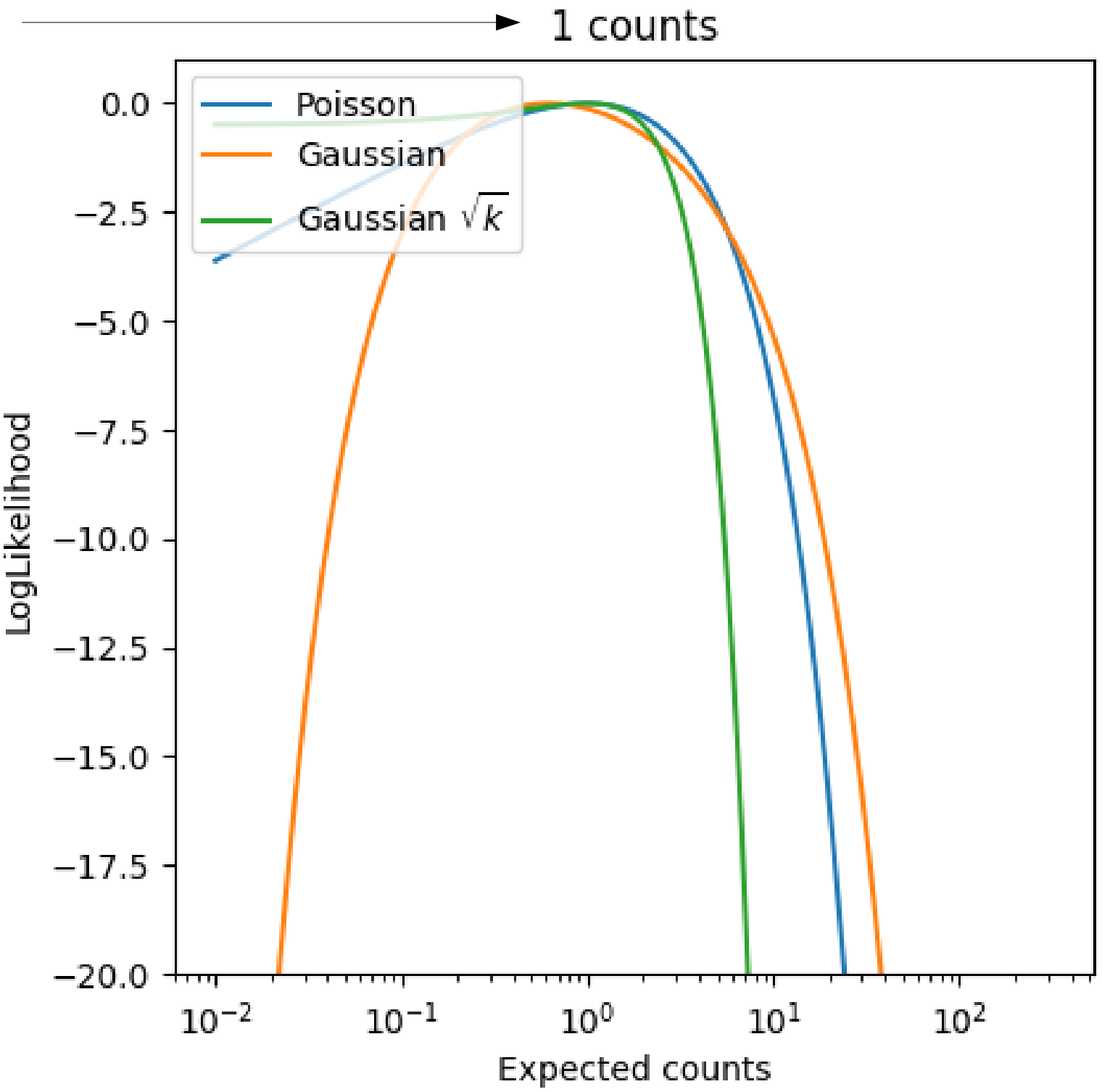
Known data

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



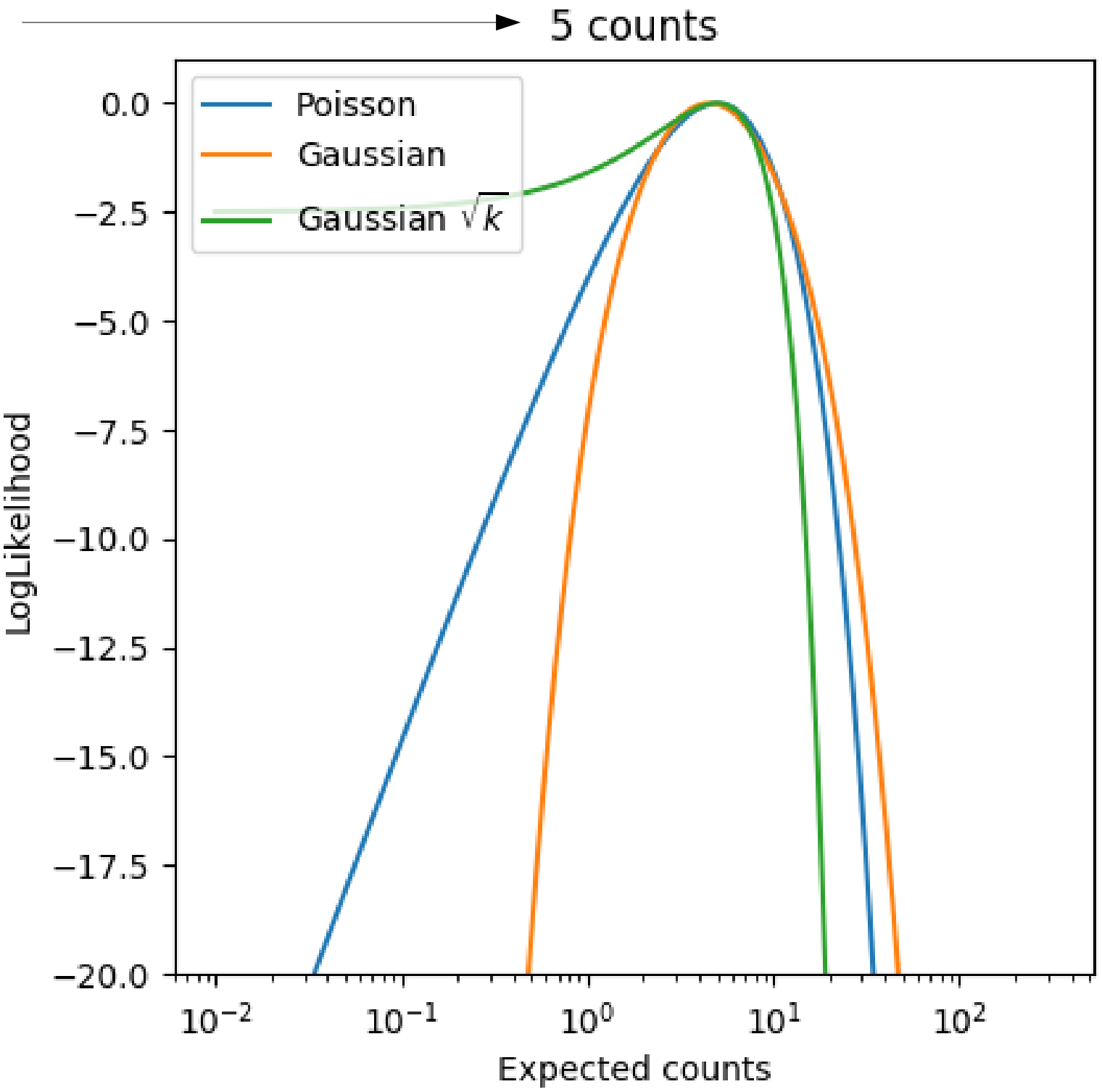
Known data

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

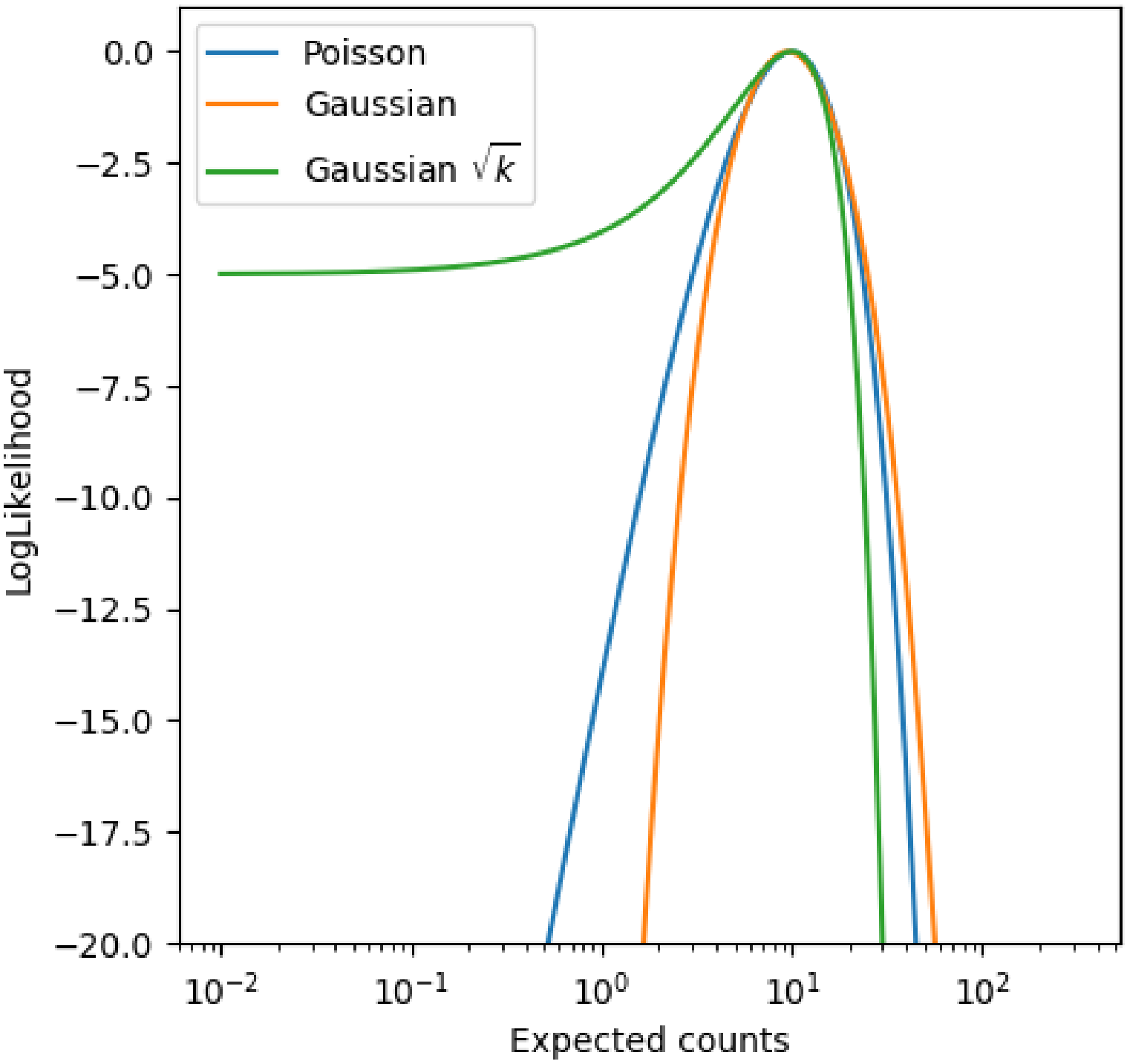
10 counts

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

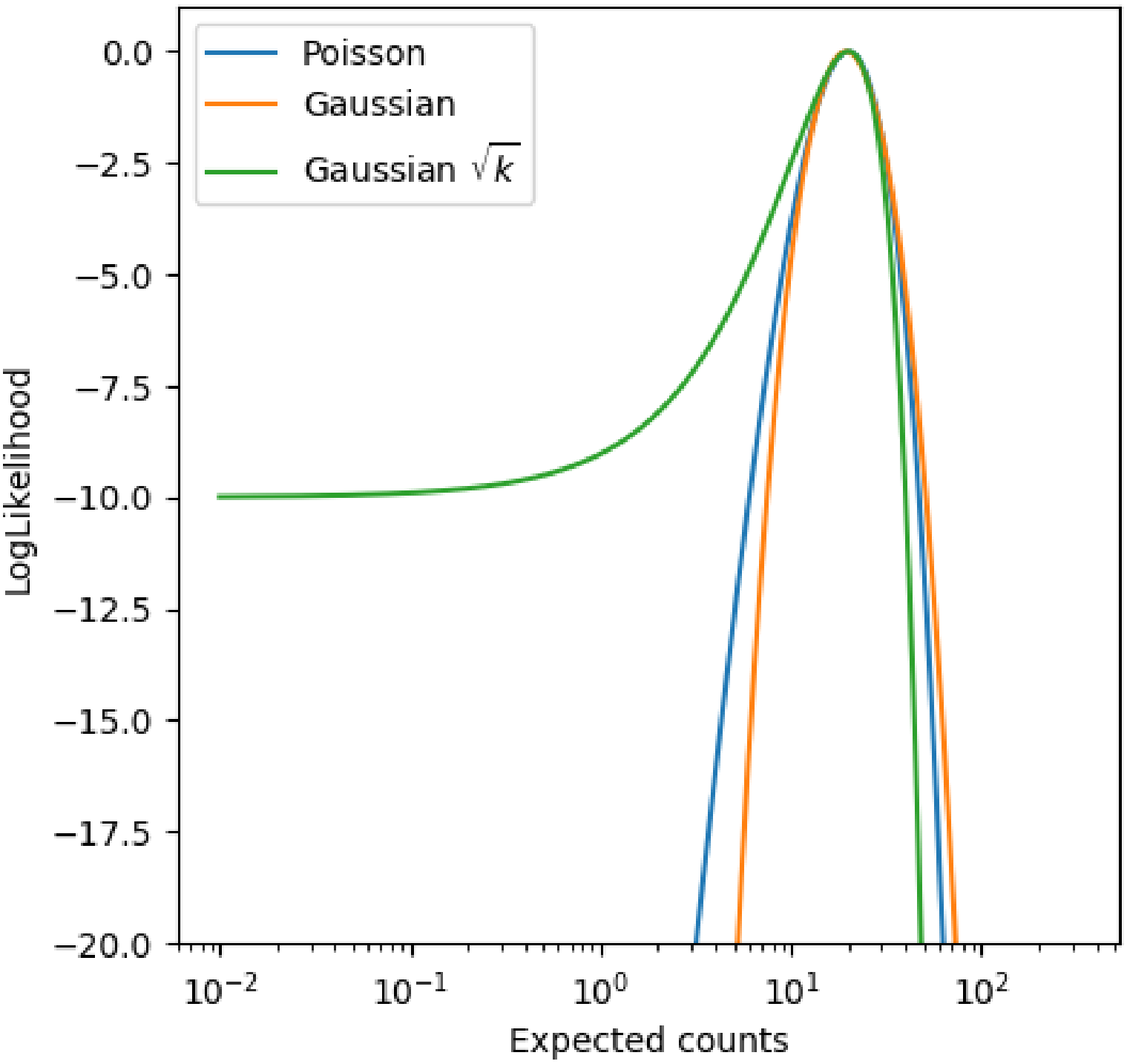
20 counts

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

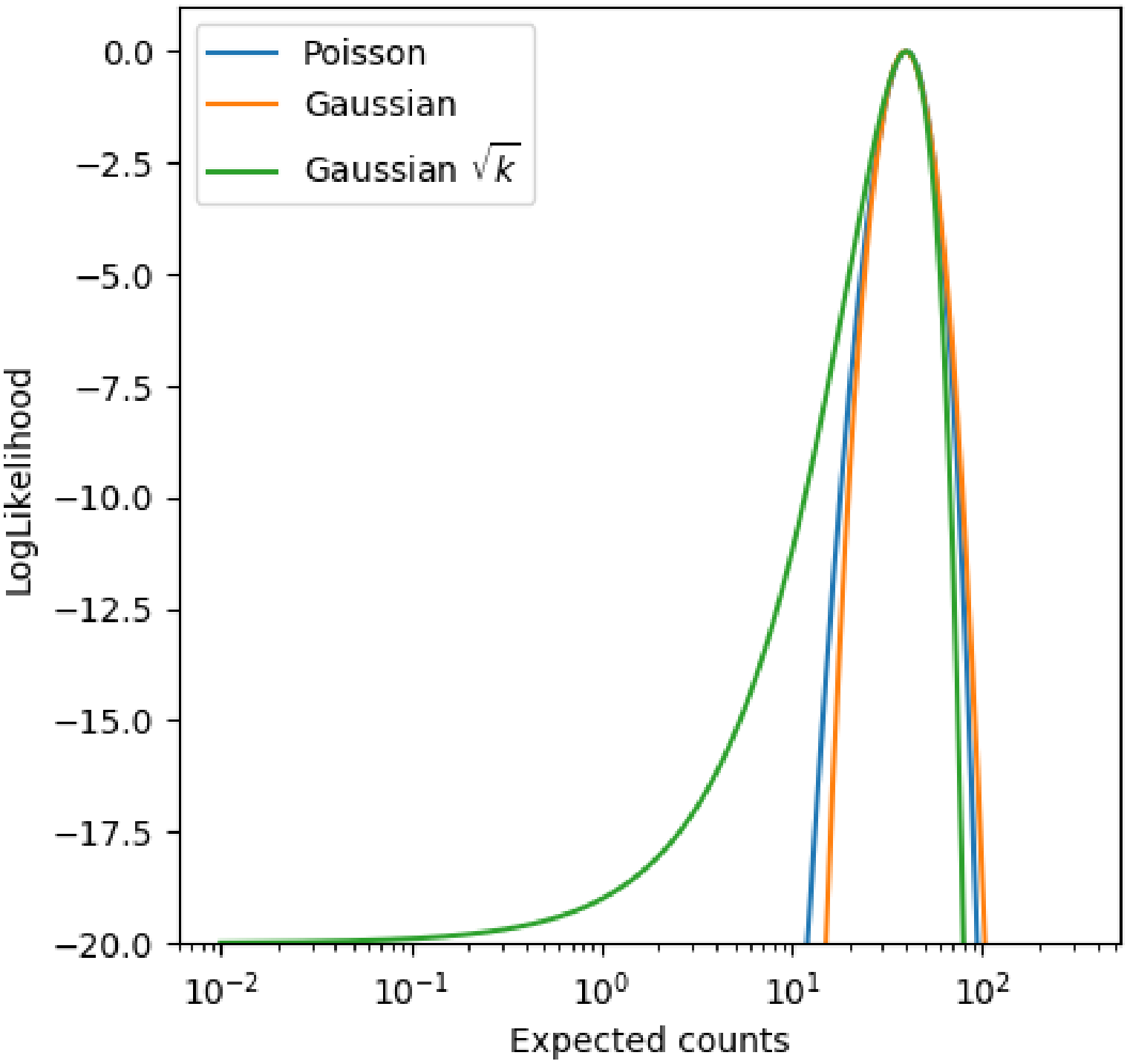
40 counts

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Known data

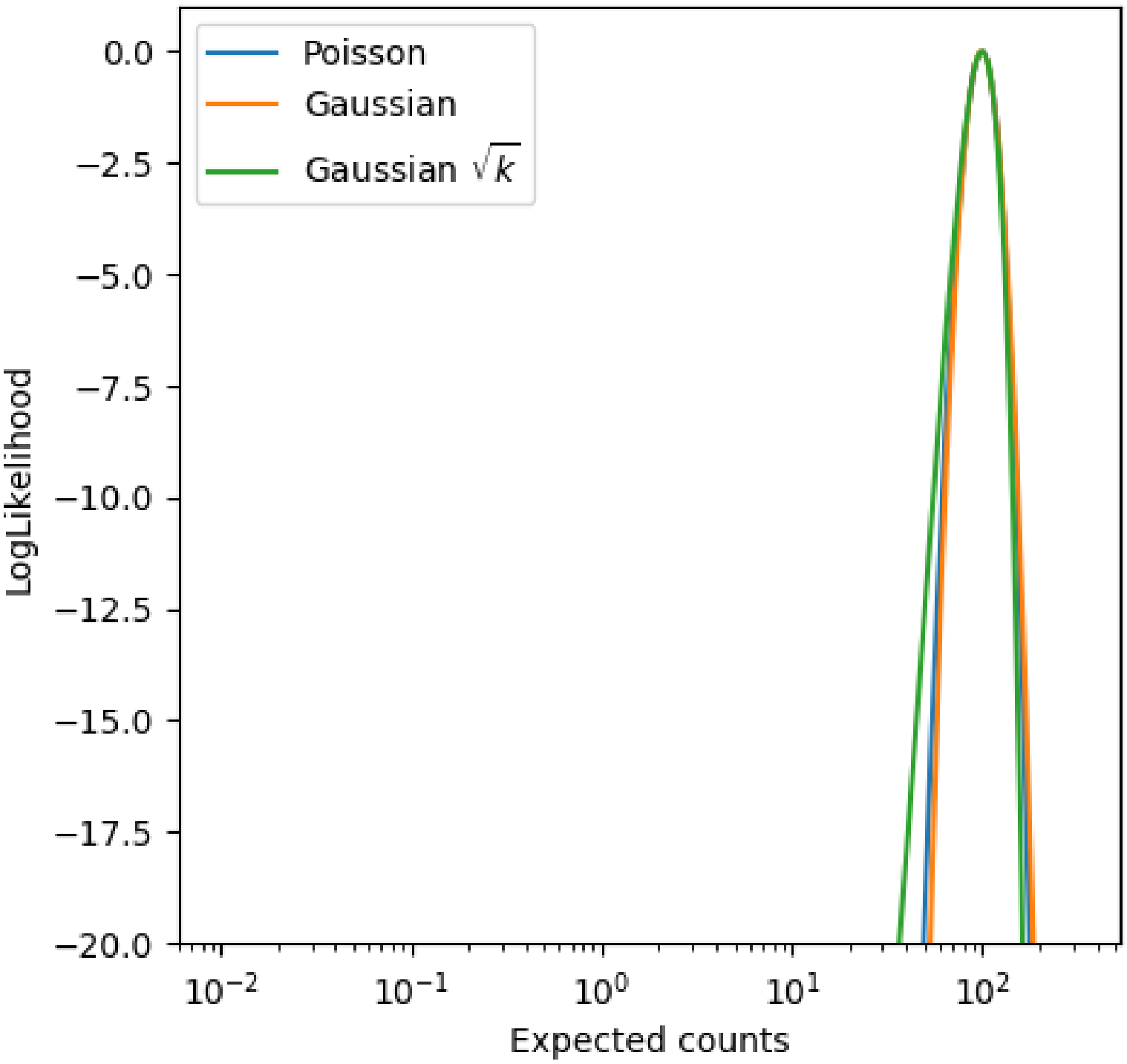
100 counts

$$P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Unknown rate

Likelihood

Probability (frequency) to produce exactly this data



Approximation quality

- Tails have different slopes
  - Gauss high-end more permissive
  - Poisson low-end more permissive
- Right way: Poisson
- Historically: Gauss faster to evaluate

“Statistics”

- Poisson
  - Likelihood $\mathcal{L}(k|\lambda) = e^{-\lambda}\lambda^k/k!$
    -2*log → $-2\log\mathcal{L}(k|\lambda) = 2\lambda - 2k\log\lambda + C$

- Gaussian
  - Likelihood $\mathcal{L}(k|\mu,\sigma) = \exp[-((x-\mu)/\sigma)^2/2]/\sqrt{2\pi\sigma^2}$
    -2*log → $-2\log\mathcal{L}(x|\mu,\sigma) = ((x-\mu)/\sigma)^2 + C$

“Statistics”

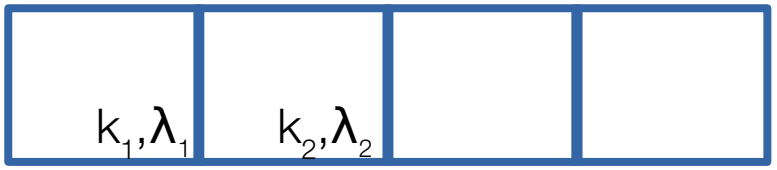
- Poisson
  - Likelihood $\mathcal{L}(k|\lambda) = e^{-\lambda}\lambda^k/k!$
  - -2*log → $-2\log\mathcal{L}(k|\lambda) = \underbrace{2\lambda - 2k\log\lambda + C}_{\textbf{CStat, Cash}}$

Cash (1979)

- Gaussian
  - Likelihood $\mathcal{L}(k|\mu,\sigma) = \exp[-((x-\mu)/\sigma)^2/2]/\sqrt{2\pi\sigma^2}$
  - -2*log → $-2\log\mathcal{L}(x|\mu,\sigma) = \underbrace{((x-\mu)/\sigma)^2}_{\textbf{Chi}^2} + C$

Does not mean they follow a chi² distribution!

Multiple bins



- Poisson

$$\mathcal{L}(k_1, k_2|\lambda_1, \lambda_2) = e^{-\lambda_1}\lambda_1^k{}_1 e^{-\lambda_2}\lambda_2^k{}_2$$

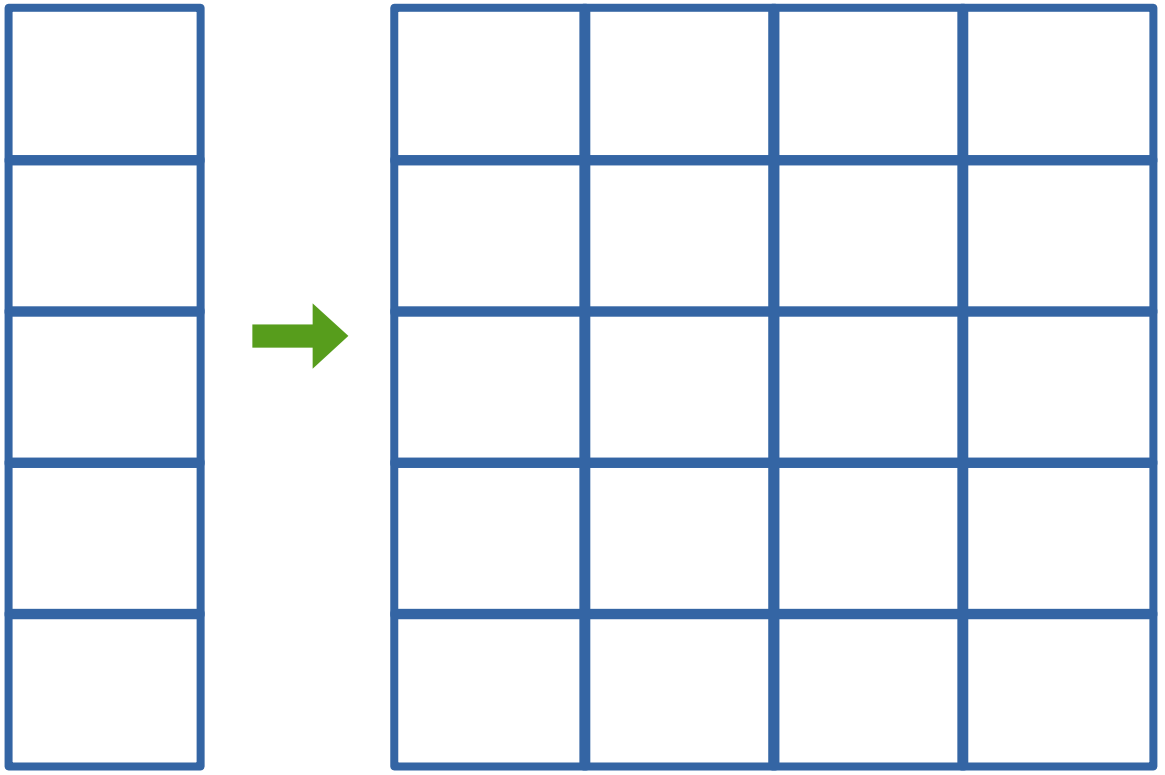
$$2\lambda_1 - 2k_1 \log \lambda_1 + 2\lambda_2 - 2k_2 \log \lambda_2 + C$$

- Gaussian

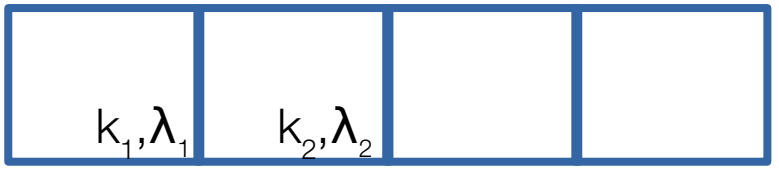
$$\mathcal{L}(x_1, x_2|\mu_1, \sigma_1, \mu_2, \sigma_2) = \exp[-((x_1 - \mu_1)/\sigma_1)^2] \exp[-((x_2 - \mu_2)/\sigma_2)^2]$$

$$((x_1 - \mu_1)/\sigma_1)^2 + ((x_2 - \mu_2)/\sigma_2)^2$$

Multiple bins



Flux



Counts

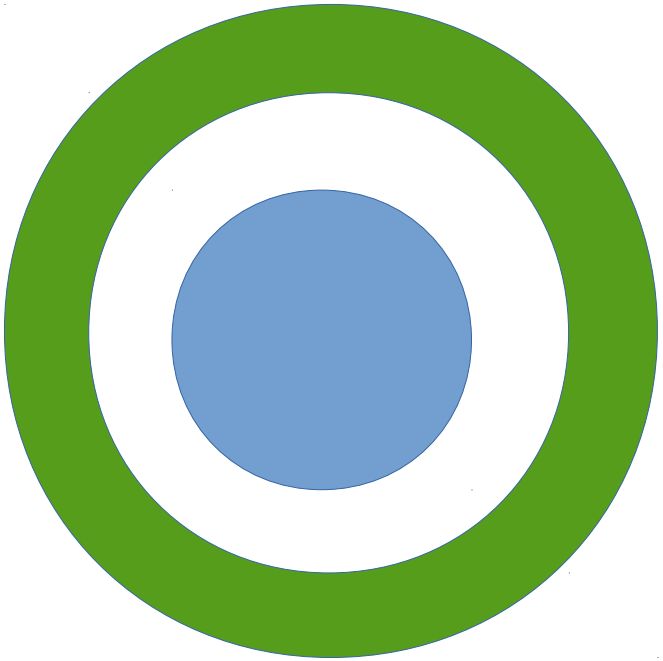
$$\overrightarrow{\lambda} = \overrightarrow{F} \cdot \underline{R}$$

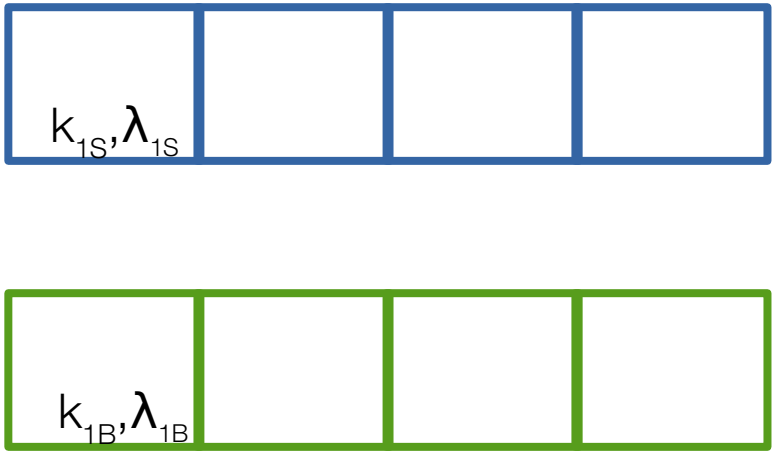
$$C = 2\overrightarrow{\lambda} \cdot \overrightarrow{\lambda} - 2\overrightarrow{k} \cdot \log \overrightarrow{\lambda}$$

Remember:
λ=number / cm² / s / keV * dE * dt * dA
k=number

Backgrounds

Backgrounds



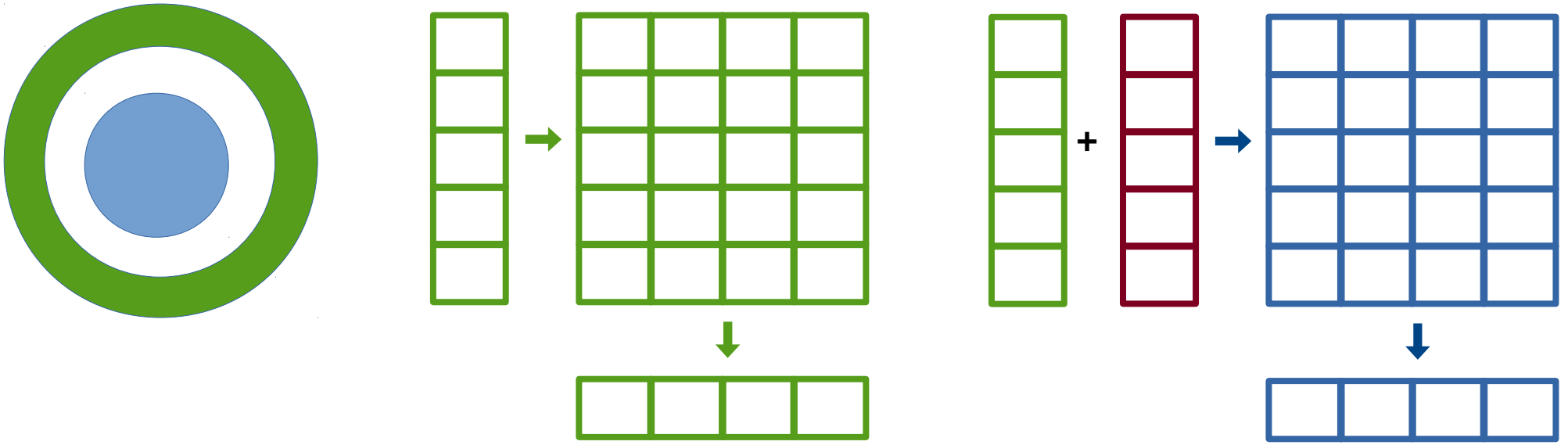


Assume time, location-independence

$k_{src},\lambda_{src},t_{src},A_{src}$

$k_{bkg},\lambda_{bkg},t_{bkg},A_{bkg}$

Background + Source



$$\vec{\lambda}_{\mathrm{src}} = \vec{F}_{\mathrm{src}} \cdot \underline{R}_{\mathrm{src}} + \vec{F}_{\mathrm{bkg}} \cdot \underline{R}_{\mathrm{src}}$$

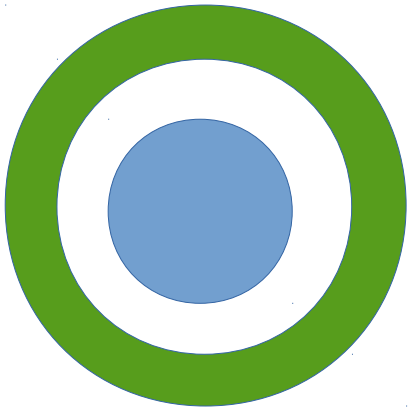
$$\vec{\lambda}_{\mathrm{bkg}} = \vec{F}_{\mathrm{bkg}} \cdot \underline{R}_{\mathrm{bkg}}$$

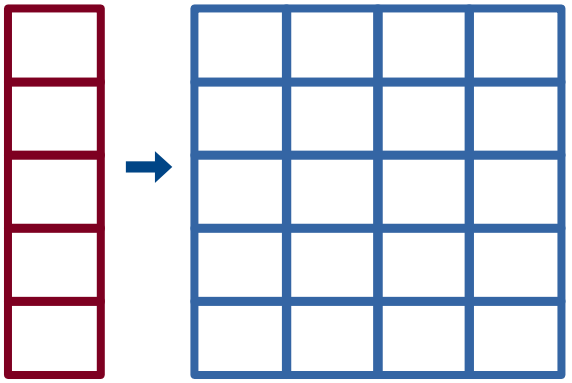
$$C = \quad 2\vec{\lambda}_{\mathrm{src}} \cdot \vec{\lambda}_{\mathrm{src}} - 2\vec{k}_{\mathrm{src}} \cdot \log \vec{\lambda}_{\mathrm{src}} + 2\vec{\lambda}_{\mathrm{bkg}} \cdot \vec{\lambda}_{\mathrm{bkg}} - 2\vec{k} \cdot \log \vec{\lambda}_{\mathrm{bkg}}$$

Assumptions:
- area energy-independent
- rate constant with area, time, location

Remember:
λ=number / cm^2 / s / keV * dE * dt * dA

Background + Source





$$\times \frac{A_{\mathrm{src}}}{A_{\mathrm{bkg}}} \frac{t_{\mathrm{src}}}{t_{\mathrm{bkg}}}$$

$$\overrightarrow{\lambda}_{\mathrm{src}} = \overrightarrow{F}_{\mathrm{src}} \cdot \underline{R}_{\mathrm{src}} + \overrightarrow{F}_{\mathrm{bkg}} \cdot \underline{R}_{\mathrm{src}}$$

$$\overrightarrow{\lambda}_{\mathrm{bkg}} = \overrightarrow{F}_{\mathrm{bkg}} \cdot \underline{R}_{\mathrm{bkg}}$$

$$C = 2\overrightarrow{\lambda}_{\mathrm{src}} \cdot \overrightarrow{\lambda}_{\mathrm{src}} - 2\overrightarrow{k}_{\mathrm{src}} \cdot \log \overrightarrow{\lambda}_{\mathrm{src}} + 2\overrightarrow{\lambda}_{\mathrm{bkg}} \cdot \overrightarrow{\lambda}_{\mathrm{bkg}} - 2\overrightarrow{k} \cdot \log \overrightarrow{\lambda}_{\mathrm{bkg}}$$

Assumptions:
- area energy-independent
- rate constant with area, time, location

Remember:
λ=number / cm² / s / keV * dE * dt * dA

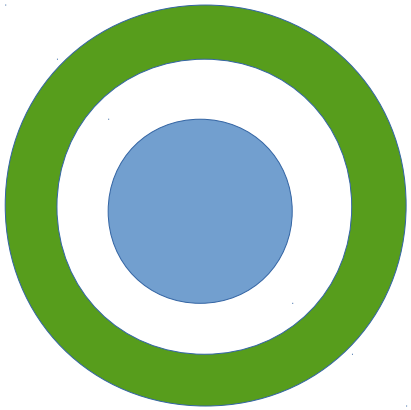
Background + Source

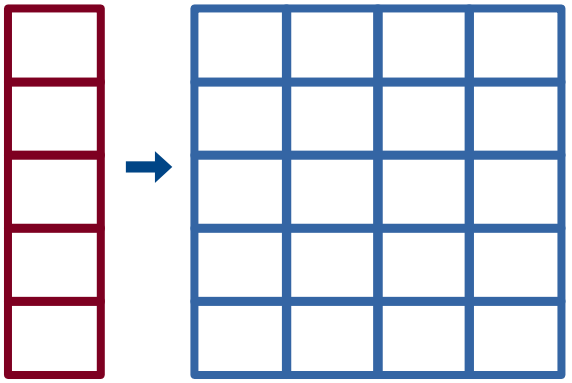
- src+bkg Gauss → Gauss (subtractable, flats/darks)
- src+bkg Poisson → Poisson
  - High counts (>100) in every single src and bkg bin → Gauss + Subtract with bkg variance propagation
  - Subtract & model with Skellam distribution
  - Do the right thing and model both as Poisson
    -
    -
    -

Background + Source

- src+bkg Gauss → Gauss (subtractable, flats/darks)
- src+bkg Poisson → Poisson
  - High counts (>100) in every single src and bkg bin → Gauss + Subtract with bkg variance propagation
  - Subtract & model with Skellam distribution
  - Do the right thing and model both as Poisson
    - Poisson estimate of rate in each bin, independently
    - Function approximation of background
      - In counts (empirical model)
      - Physical background flux model
      - Fit simultaneously with source
      - Fit background model first, use best-fit background shape for source fit

Background + Source





$$\times \frac{A_{\mathrm{src}}}{A_{\mathrm{bkg}}} \frac{t_{\mathrm{src}}}{t_{\mathrm{bkg}}}$$

$$\overrightarrow{\lambda}_{\mathrm{src}} = \overrightarrow{F}_{\mathrm{src}} \cdot \underline{R}_{\mathrm{src}} + \overrightarrow{F}_{\mathrm{bkg}} \cdot \underline{R}_{\mathrm{src}}$$

$$\overrightarrow{\lambda}_{\mathrm{bkg}} = \overrightarrow{F}_{\mathrm{bkg}} \cdot \underline{R}_{\mathrm{bkg}}$$

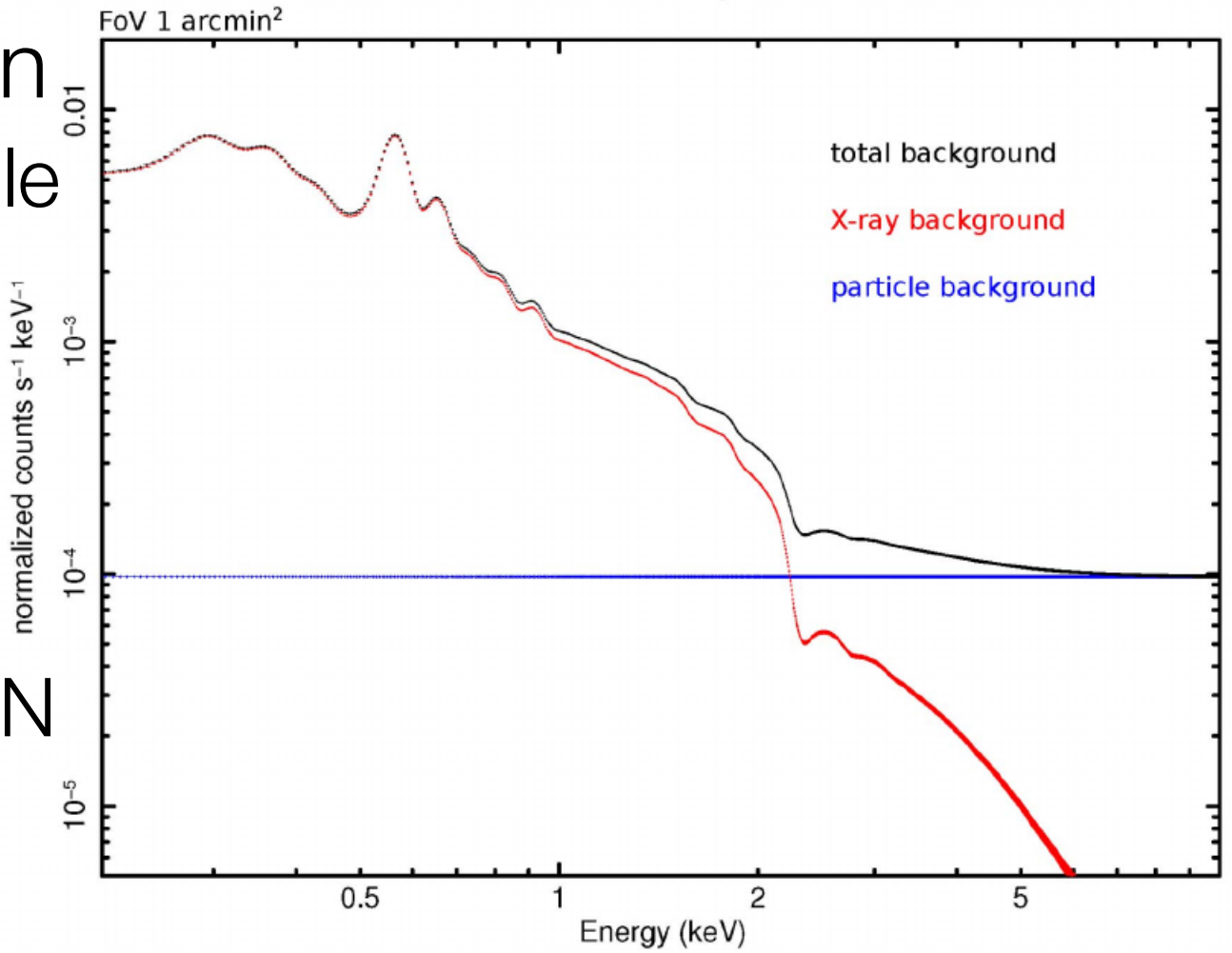
$$C = 2\overrightarrow{\lambda}_{\mathrm{src}} \cdot \overrightarrow{\lambda}_{\mathrm{src}} - 2\overrightarrow{k}_{\mathrm{src}} \cdot \log \overrightarrow{\lambda}_{\mathrm{src}} + 2\overrightarrow{\lambda}_{\mathrm{bkg}} \cdot \overrightarrow{\lambda}_{\mathrm{bkg}} - 2\overrightarrow{k} \cdot \log \overrightarrow{\lambda}_{\mathrm{bkg}}$$

Assumptions:
- area energy-independent
- rate constant with area, time, location

Remember:
λ=number / cm² / s / keV * dE * dt * dA

eROSITA background

- Diffuse emission
  - Local hot bubble
  - Galactic disk
  - Galactic halo
- Cosmic background
  - Unresolved AGN
- High-energy particle background

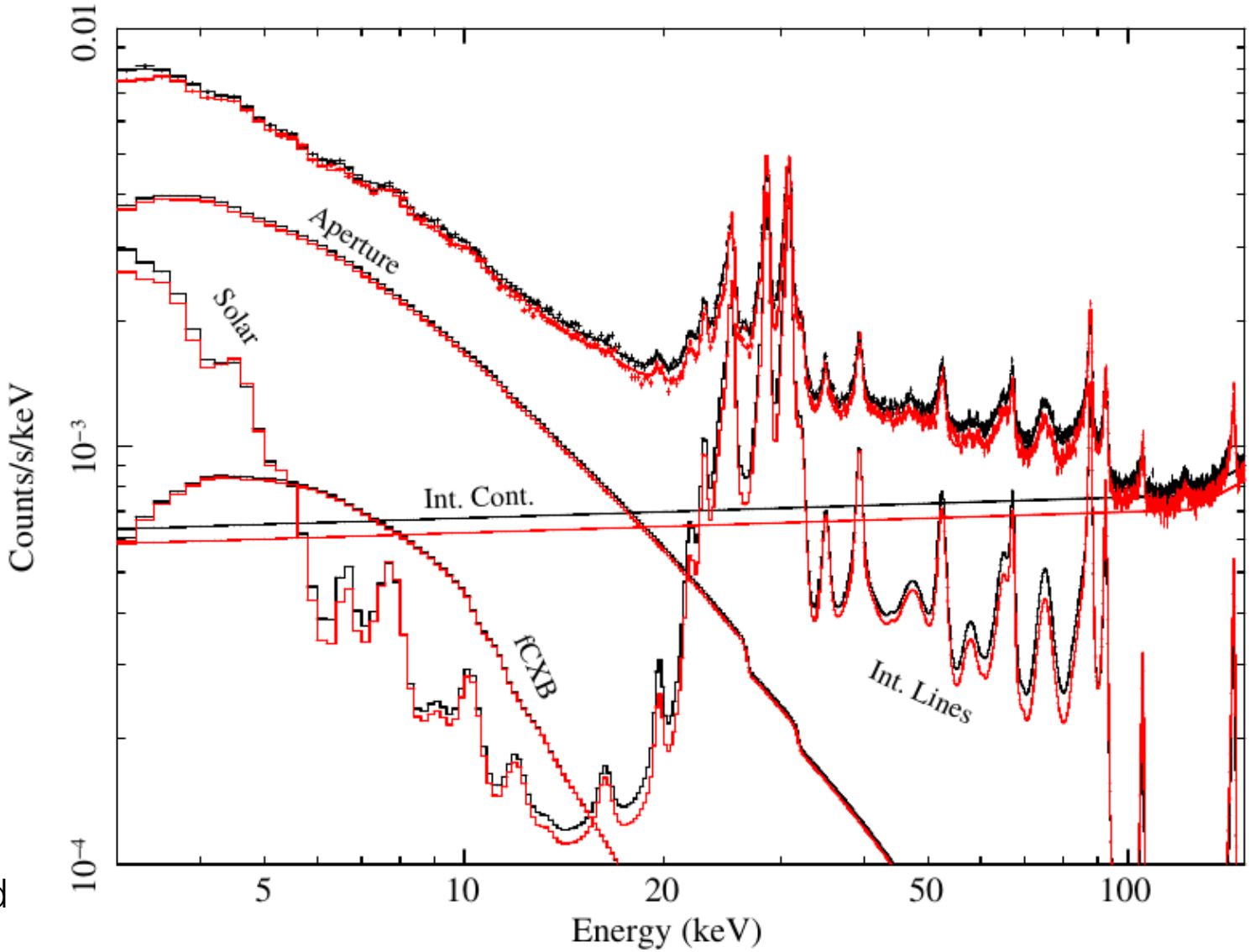


https://wiki.mpe.mpg.de/eRosita/ScienceRelatedStuff/Background

Semi-physical background models

Maximize poisson likelihood at all bins
→shape

Particle background
Cosmic background
Instrumental background
…
Location & time-dependent



NuSTAR (Wik+14)

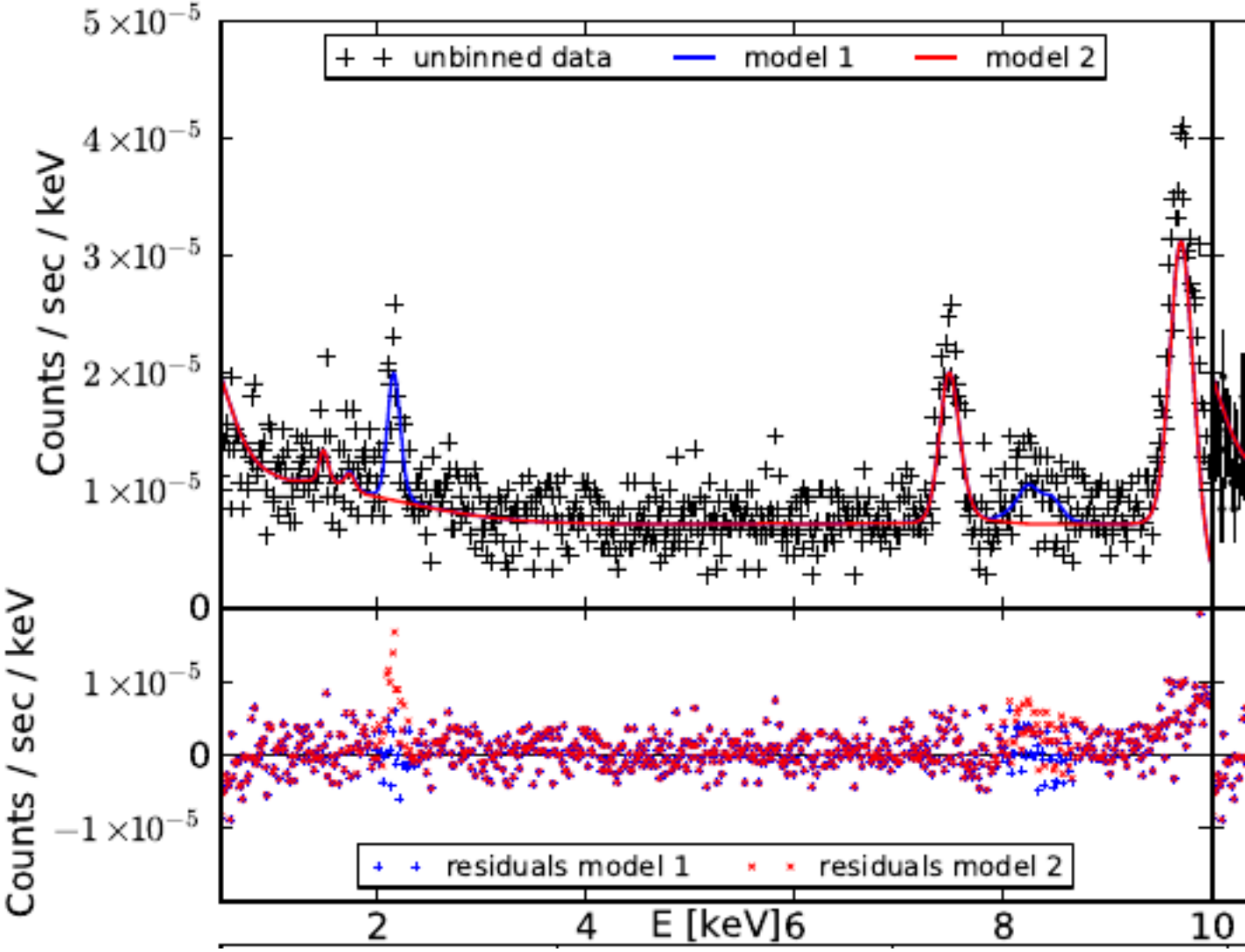
→ especially important for extended source

Empirical background models

Maximize poisson likelihood at all bins →shape

Pros:
- Can contain physical knowledge & smoothness
- Small uncertainties
- 0 bin counts ok

Cons:
- Need to specify model
- Fit can be poor



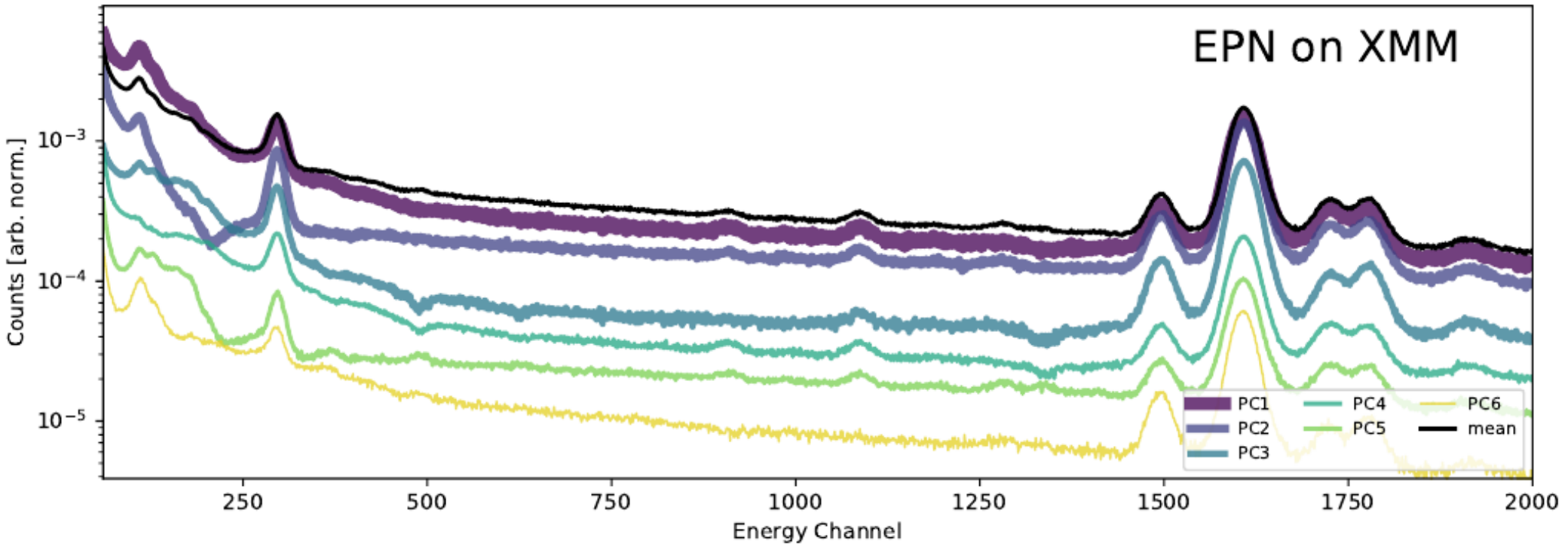
Chandra

(XMM, Chandra, Swift models in-house)

Empirical background models

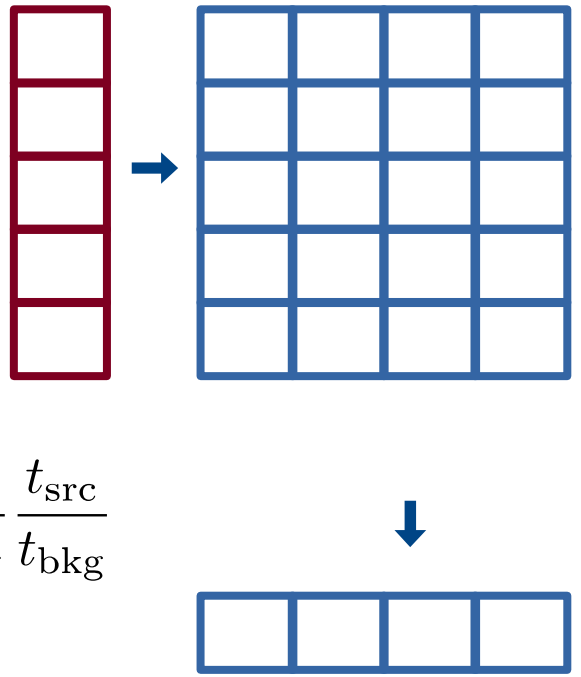
Automated shape finding
Simmonds, Buchner et al. (2017)

XMM/PN,MOS, Chandra/ACIS, NuSTAR,
Suzaku, RXTE, Swift/XRT



Background: Individual bins

$$\times \frac{A_{\rm src}}{A_{\rm bkg}} \frac{t_{\rm src}}{t_{\rm bkg}}$$



Estimate most likely background rate in each bin

Add scaled to source region counts

(wstat,
Xspec default if set to cstat with no background model)
pgstat

Pros:
- no model specification needed

Cons:
- no continuity
- unnecessarily large uncertainties
- need >0 counts per bin

Inference with likelihoods

-0.5 Cstat, -0.5 chi²

$$\mathcal{L}(\overrightarrow{k}|\theta_1, \theta_2, ..., \theta_d, M, R, B, ...)$$

Higher L: model under these parameters often makes this data
Lower L: less frequently

→ Frequency of data $P(D|\theta)$

Likelihood <u>function</u> at D, at parameter values (not a density)

Inference desiderata

- Parameter ranges allowed or probable (L, T, …, physical parameters)

$P(\theta|D)d\theta$ Probability density

In infinitely small region: zero probability

$P(\theta|D)d\theta$

Density Volume

Probability mass

Find regions with high probability mass

$P(D|\theta)$

Parameter space exploration

Conditional probabilities

- Bayes theorem
- P(A|B) != P(B|A)
- Normalisation
- Parameter inference
- Model inference
- Interpretation

Conditional probabilities

$$A \frown B = B \frown A$$

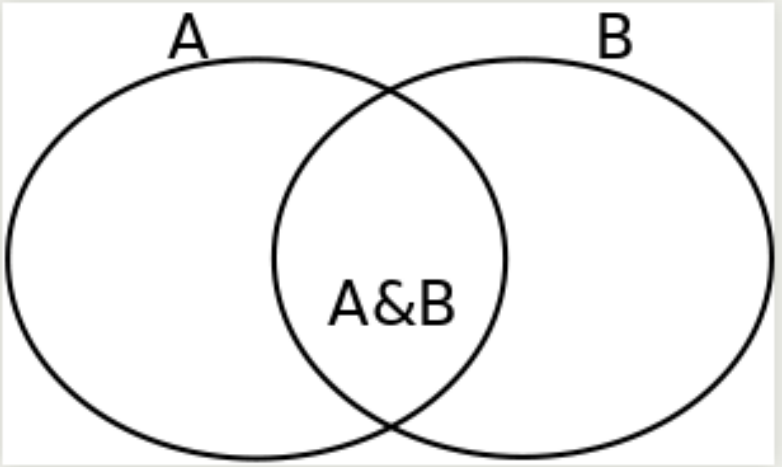
$$p(A \frown B) = p(B \frown A)$$

$$p(A|B)p(B) = p(B|A)p(A)$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

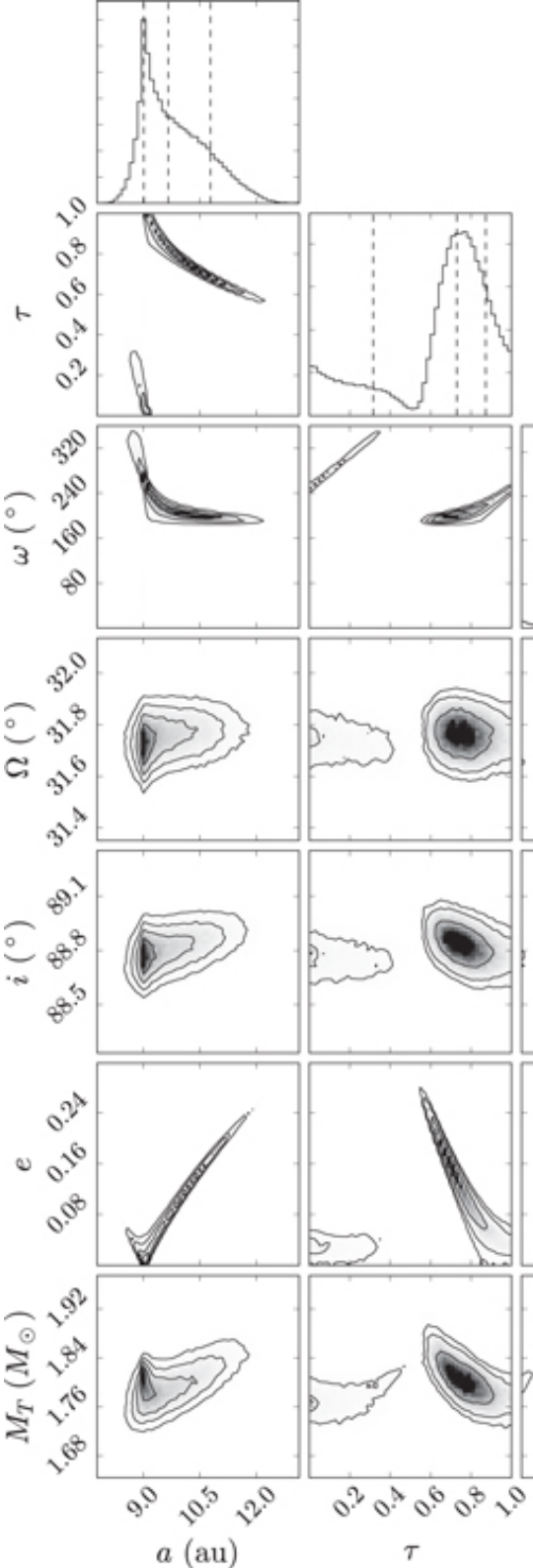
Bayes theorem



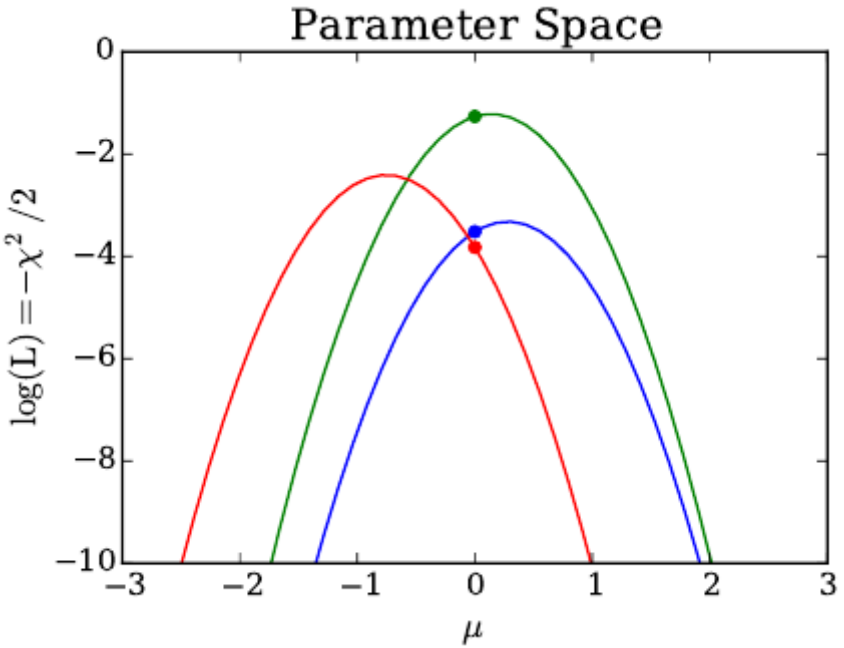
Parameter space exploration

Parameter space exploration

- Local optimization
  - LM, simplex, … (many)
  - Monte carlo optimization
- Local sampling: MCMC
  - Tempering
  - Limitations
- Global optimization
  - Genetic algorithms (DE)
- Global sampling
  - Nested sampling



Best fit parameters



If many data are created under θ
logL interval -1 below best fit
Contains true value 68% of realisations

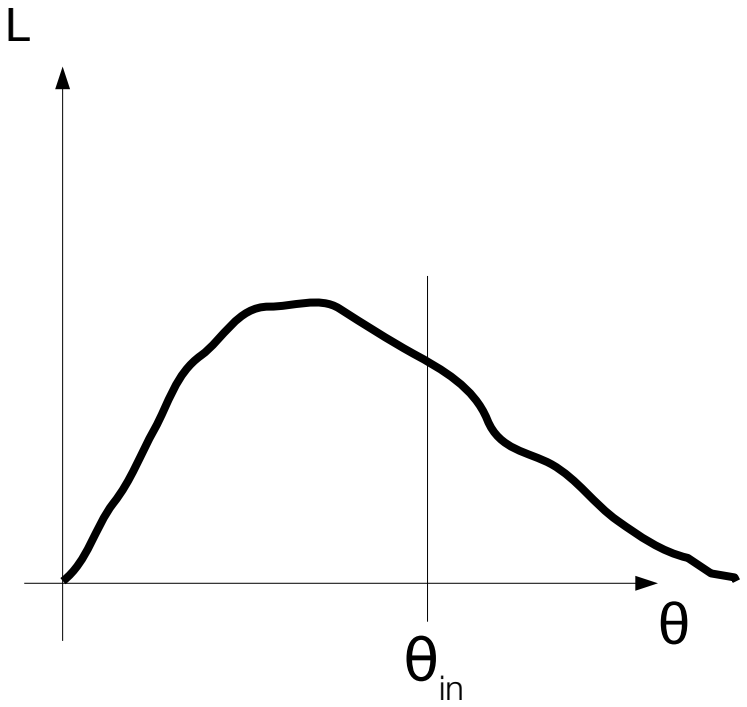
- If away from boundary
- If model is linear
- If ndata → high (symmetric, single gauss)
- If θ is true parameter

  → then

Confidence interval

What was the question again?
Are conditions fulfilled?
What do unequal “errors” mean?

Best fit parameters



If conditions are not met

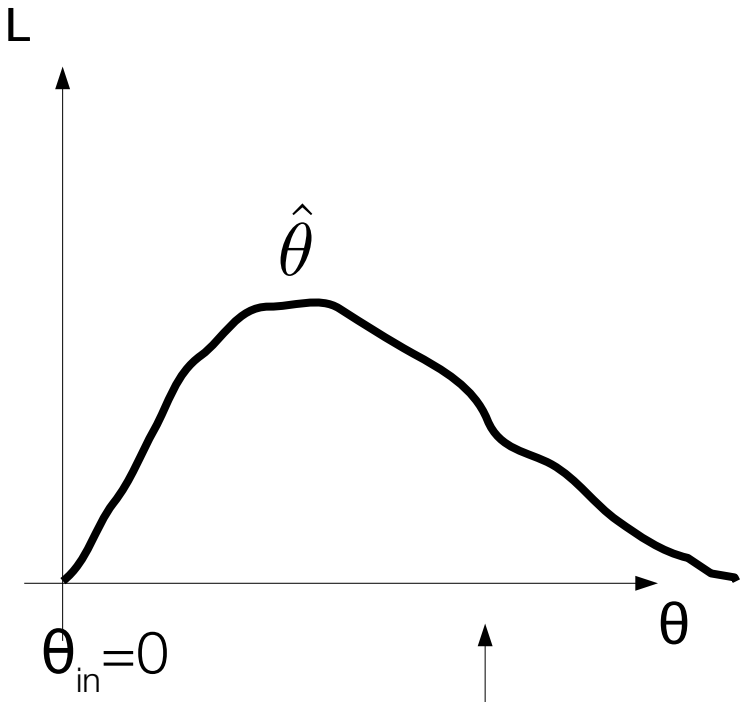
(always)

- If away from boundary
- If model is linear
- If ndata → high (symmetric, single gauss)
- If θ is true parameter

→ Monte Carlos simulations (parametric bootstrap)

Calibrate a Confidence interval

Detection

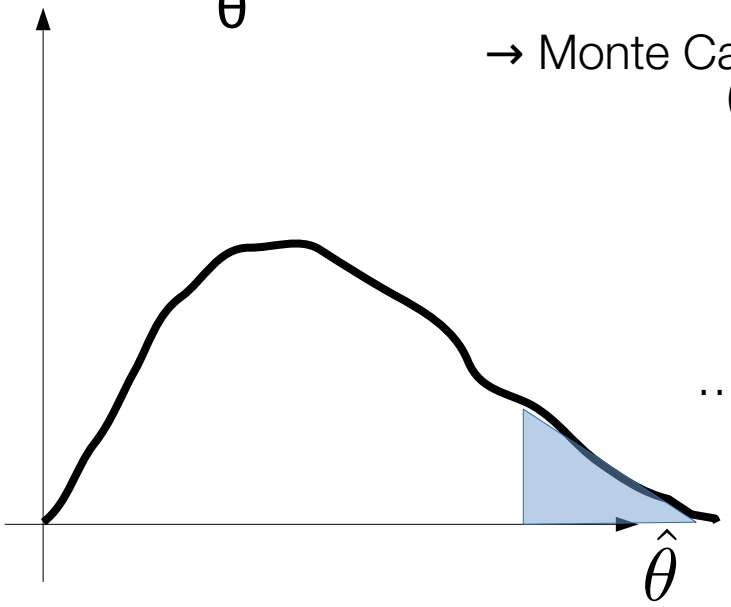


If conditions are not met

(always)

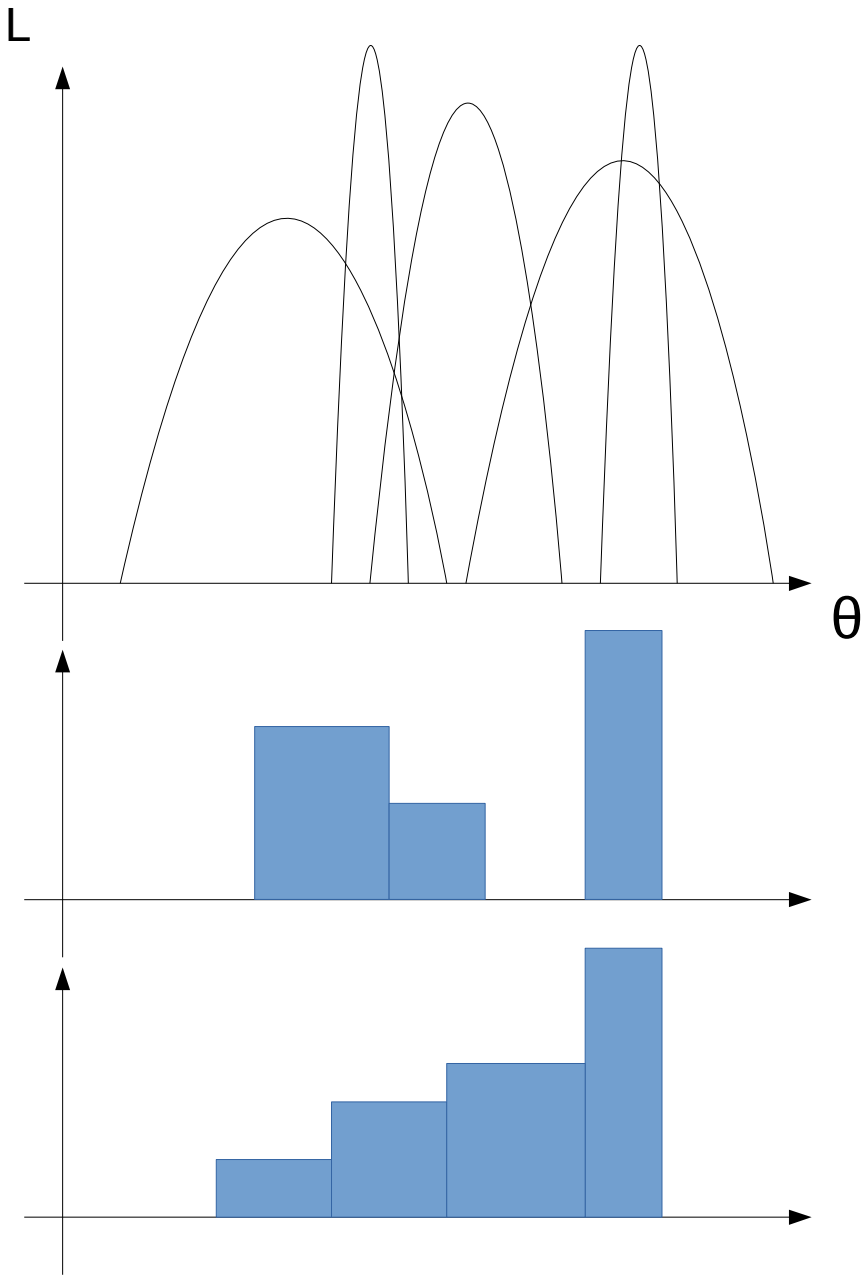
- If away from boundary
- If model is linear
- If ndata → high (symmetric, single gauss)
- If θ is true parameter

→ Monte Carlos simulations
(parametric bootstrap)



… p-values

Best fit distributions



Convolution of

True parameter distribution +
Measurement error & analysis method

Confidence intervals

Histogram of best fits

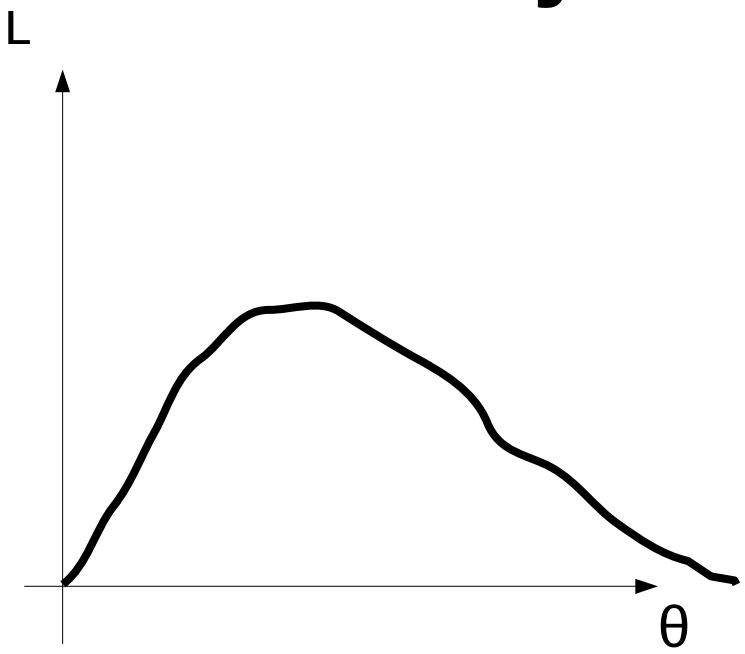
Meaning?
Upper limits?

Cumulative distribution

Clean solution:
Model population distribution (HBM) Buchner+17a

Sampling

Bayesian posterior



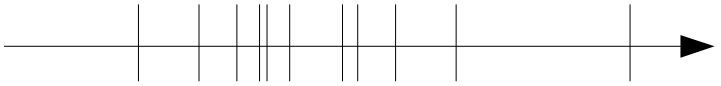
$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

$$P(\theta|D)d\theta$$

Density Volume

Probability mass

Find regions with high probability mass



Idea: Sample parameter solutions proportionally to their probability

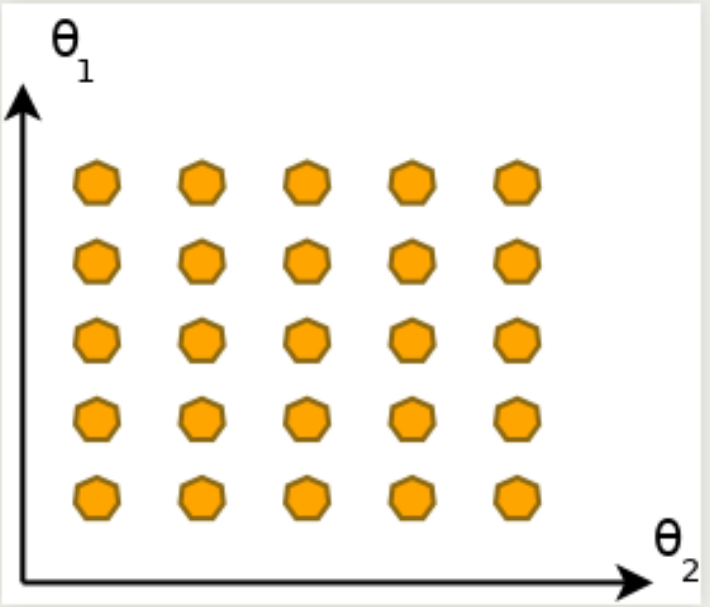
For example with a grid

Posterior grid

- evaluate *likelihood* at every point
  - how prone is the process to produce the observed data
- Compute relative importance:

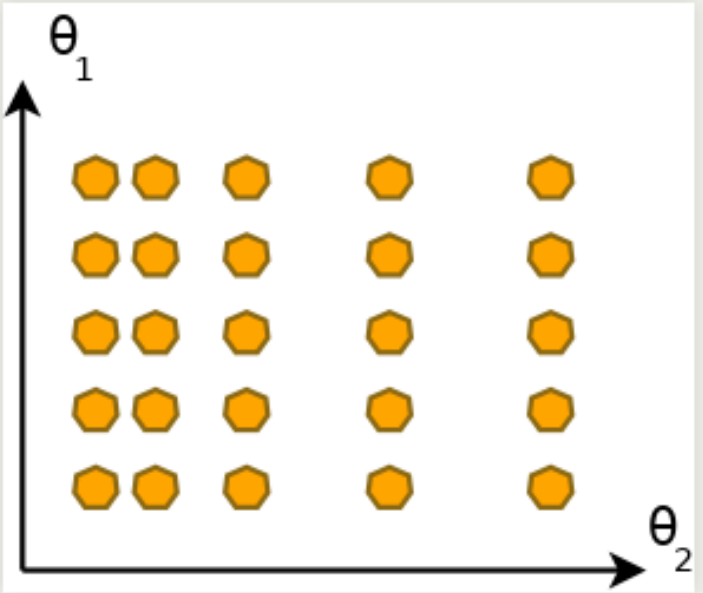
$$\mathcal{L}/\overline{\mathcal{L}}$$

- Grab those that make up 90% of $\sum \mathcal{L}$
- $Z = \overline{\mathcal{L}}$ "evidence" is average likelihood

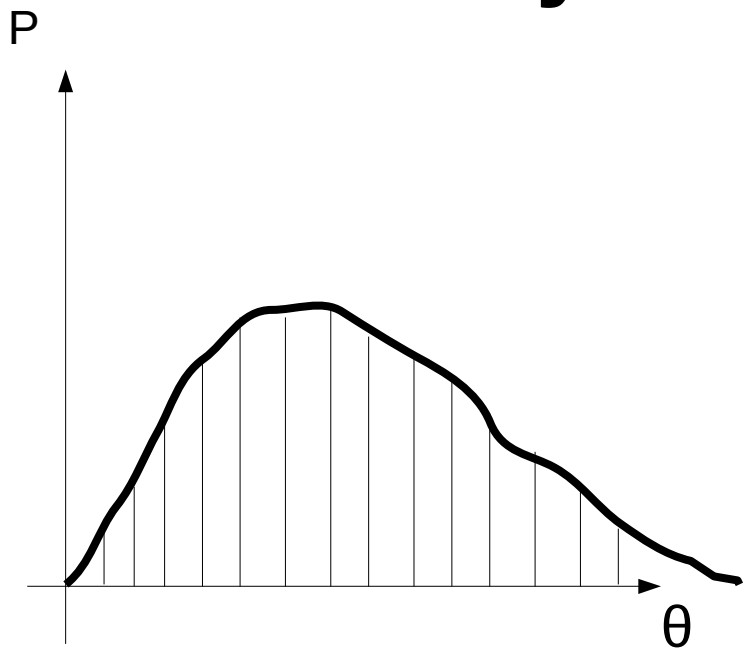


Posterior grid

- Result is dependent on placement
- Equal spacing in θ_1 or in $\log \theta_1$.
- Choice of spacing is called "prior"
- coin = investment in computing there, put coins where it is worthwhile



Bayesian posterior



parameter solutions weighted by their probability

Credible intervals

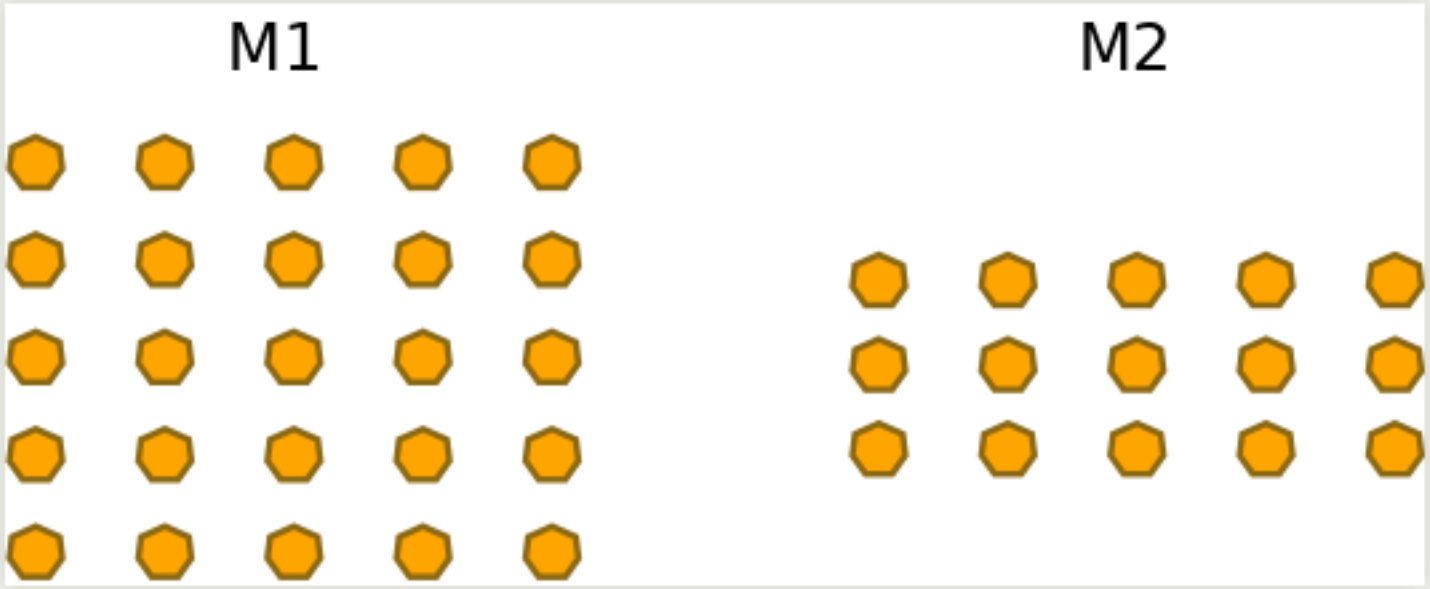
Definitions:

Density → cumulative → quantiles

Highest Density Intervals

Borders (upper limits)

Two models



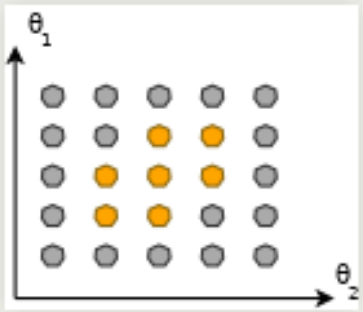
- Compare two parameter spaces by

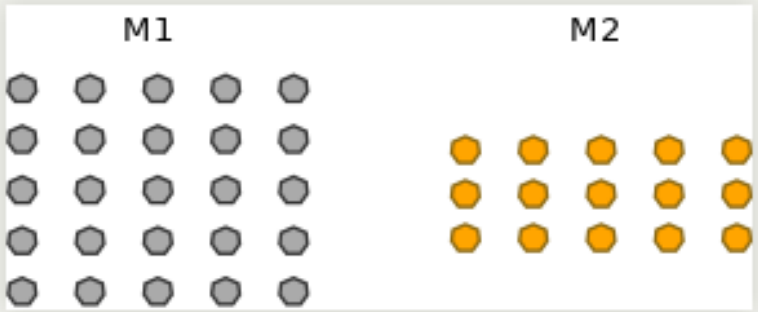
$$\sum \mathcal{L}\big|_{M1} / \sum \mathcal{L}\big|_{M2}$$

- How many coins to put in M1, M2?
- model prior

Parameter Estimation vs. Model Comparison

- Remove coins contributing less than 10%.
- Under Bayesian inference, same problem:
  - comparing bags of hypotheses





- prior is measure, rule of averaging, deformation of space to "natural variables", investment in/weighting of sub-regions
- most common priors: uniform, log-uniform.
- model priors are relative size of spaces

Curse of dimensionality

- k^d grid → infeasible
- Sample θ

  θ_1 θ_2 θ_3 ....

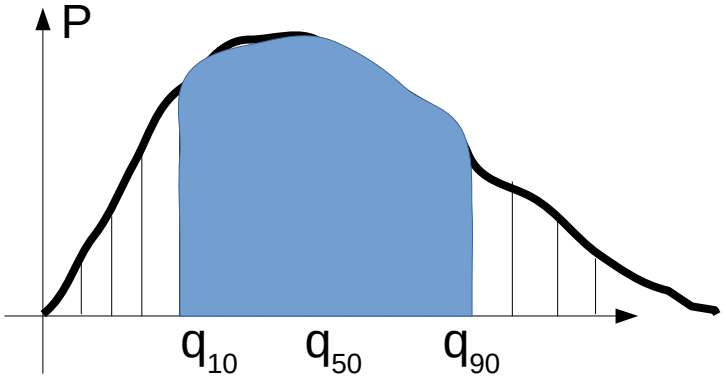
  w_1 w_2 w_3 ....

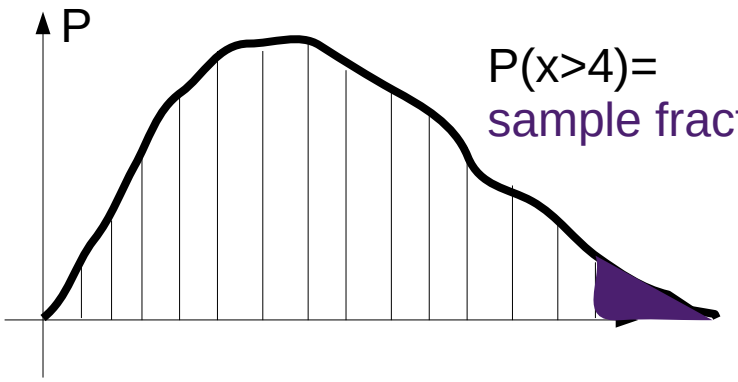
  (Posterior chains)
- Techniques:
  - Importance sampling
  - MCMC
  - Nested sampling

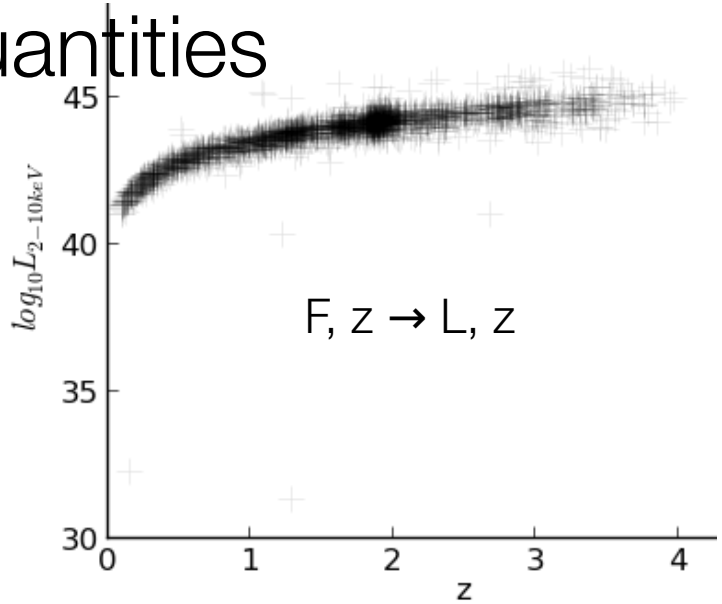
Using posterior chains

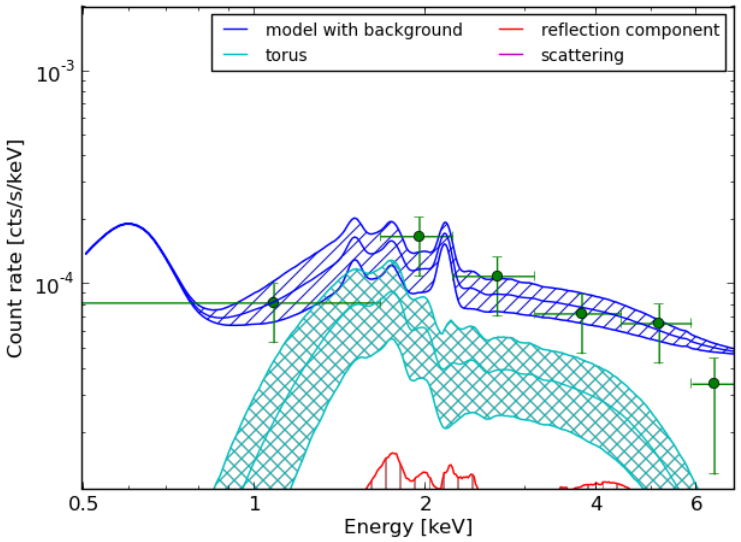
- Posterior chain

  $\theta_1\ \theta_2\theta_3$ ....
- Find regions with high prob
- Compute prob. of regions
- Posterior predictions
- Derived quantities

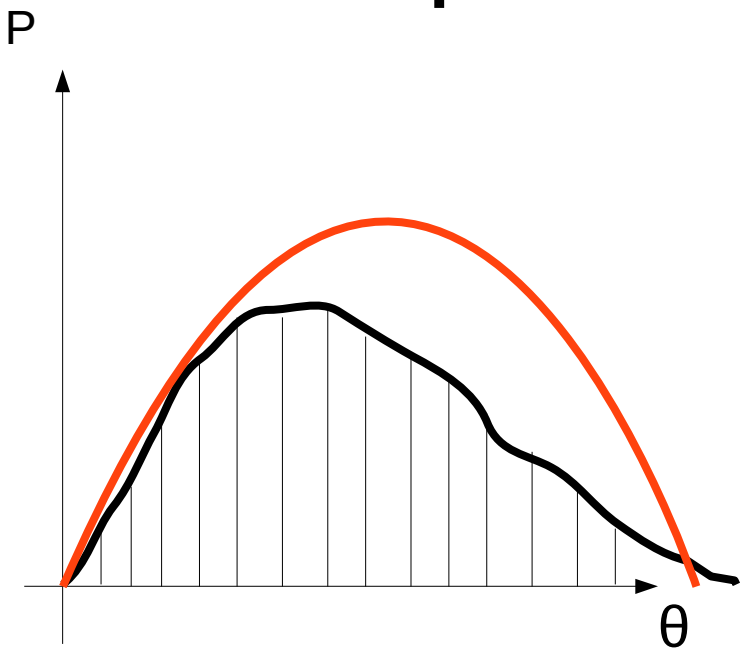








Importance sampling



Draw from proposal distribution Q

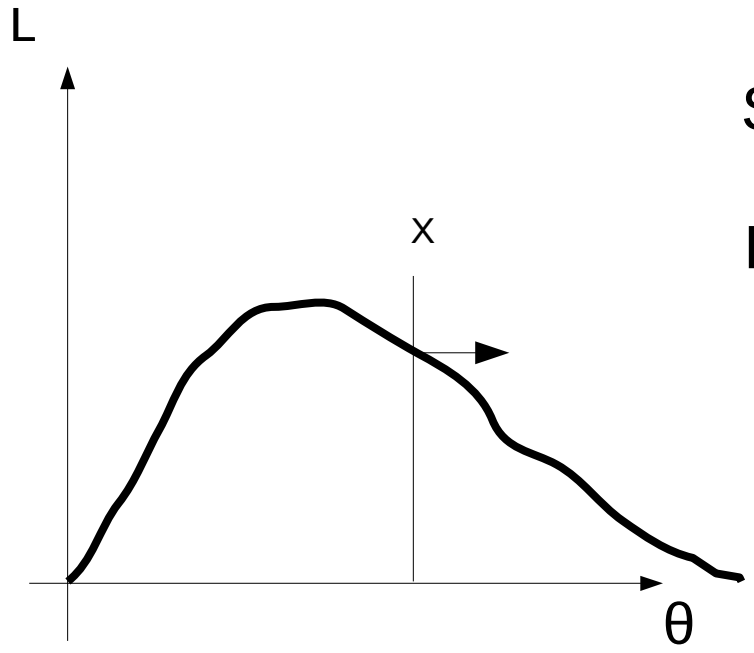
Weigh by $Q(\theta)/P(\theta|D)$

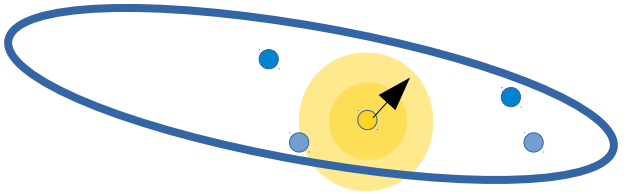
→ weighted θ chain

Advantages:
- Efficient in low-d
- Parallelisable
- Can integrate parameter space

Disadvantages
- Need to find good proposal (VB)
- Poor scaling to 10-20d
- Poor performance if proposal is bad (variance indicator)

Markov Chain Monte Carlo

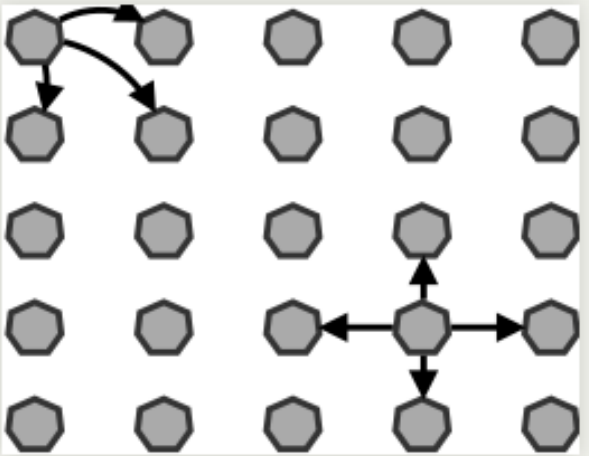




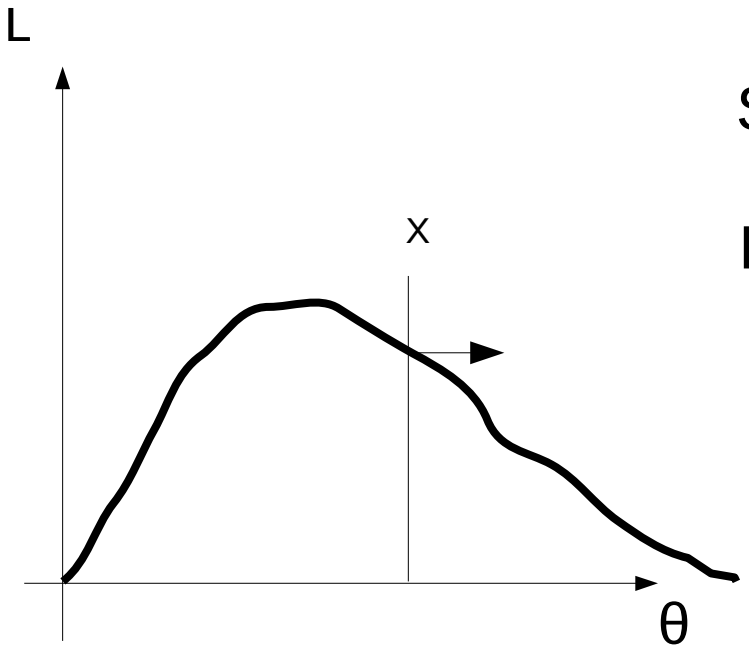
```
Starting point θ

Loop forever:
   θ' = Normal(θ, sigma_p)
   if P(θ'|D)/P(θ|D) > U():
      θ = θ'
   add θ to chain
```

- Missing ingredient: transition kernel
- tune to the problems
- Fraction of visits ~ converges to ~ probability of hypothesis
- Where does chain spend 90% of its visits



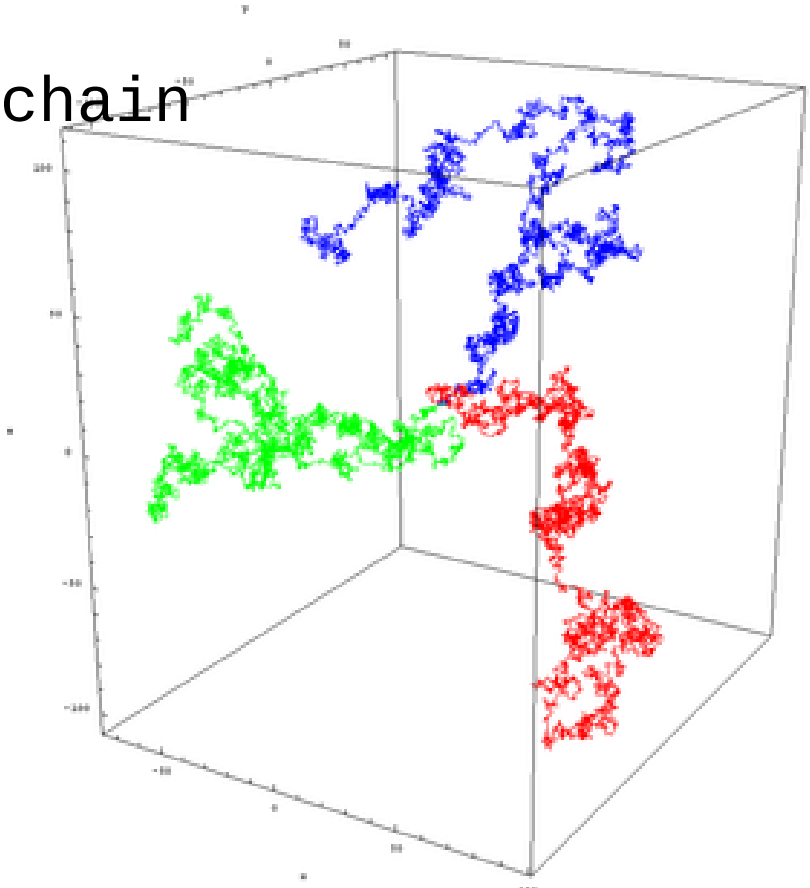
MCMC



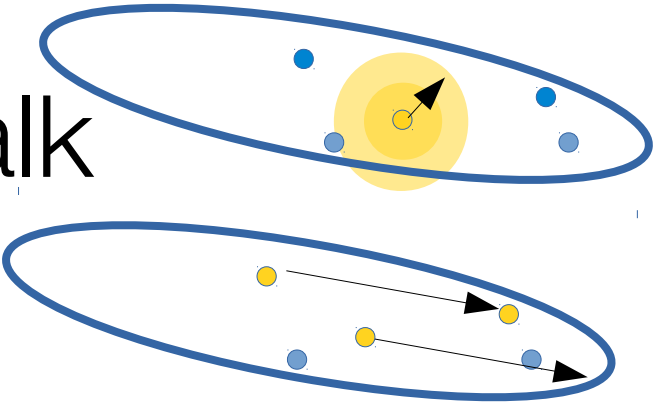

```
Starting point θ

Loop forever:
   θ’ = Normal(θ, sigma_p)
   if P(θ’|D)/P(θ|D) > U():
      θ = θ’
   add θ to chain
```

Emerging behaviour:



MCMC proposals



- Metropolis + Random Walk
- Goodman-Weare (emcee)
- HMC (Hamiltonian Monte Carlo)

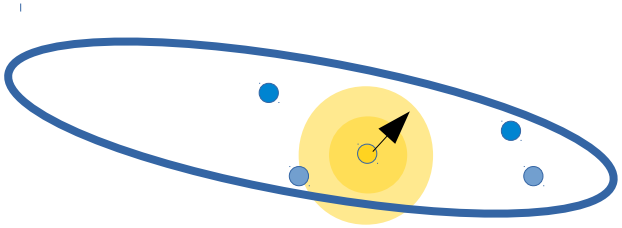
→ animation

https://chi-feng.github.io/mcmc-demo/app.html

Random walk, HMC

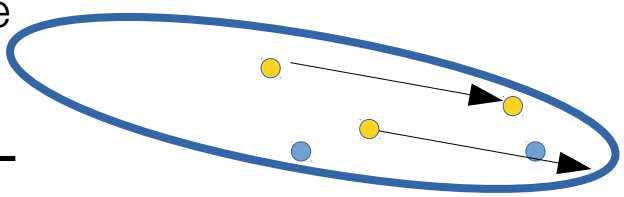
MCMC proposals

- Metropolis Random Walk
  - Adv: simple
  - Disadv: poor mixing



- Affine-invariant ensemble
  - Adv: auto-tuning for gaussian L
  - Disadv: poor mixing in bananas, collapses in high-d (Huijser+15)

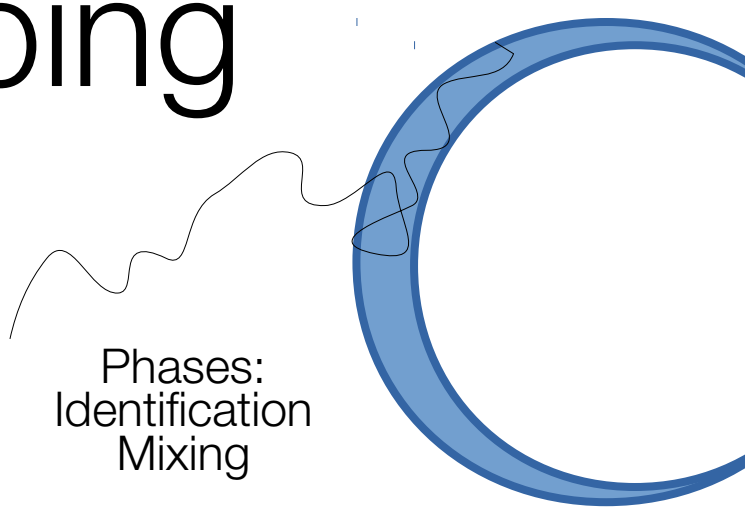
Goodman & Weare (2010)
emcee



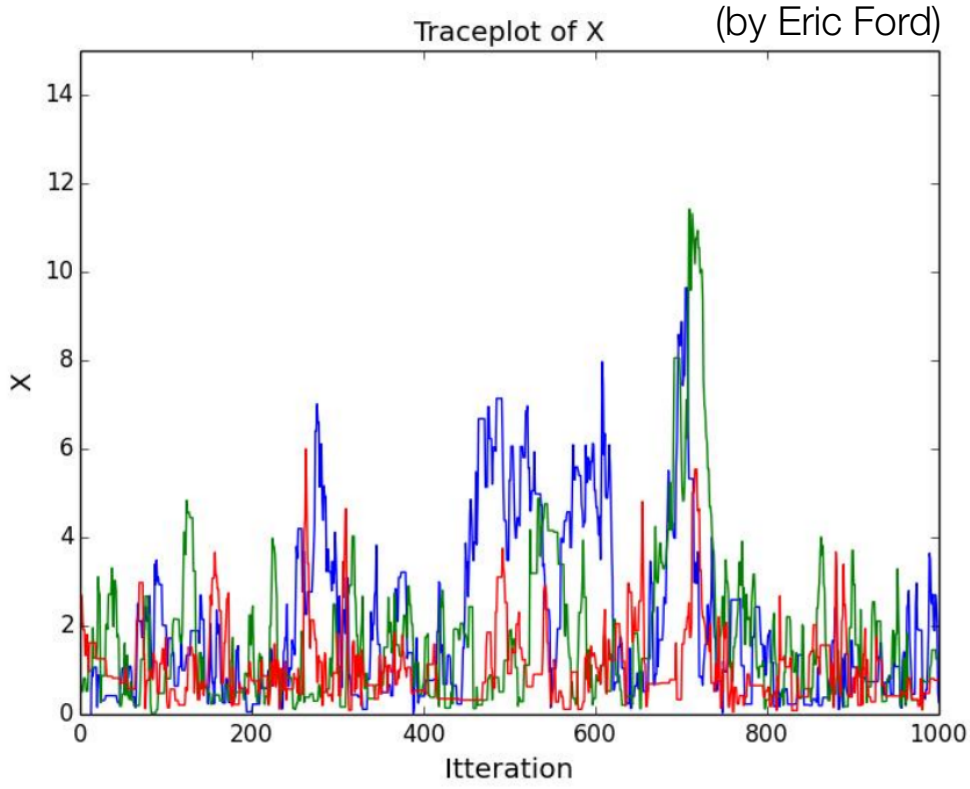
- HMC (Hamiltonian Monte Carlo)
  - Adv: tunes itself to surface
  - Disadv: need gradients of models

MCMC stopping

- MCMC theory: n→inf
- Trace plots
- Autocorrelation length
- Convergence tests
  - Detect if unreliable
  - Gelman-Rubin diagnostic
  - (many more)

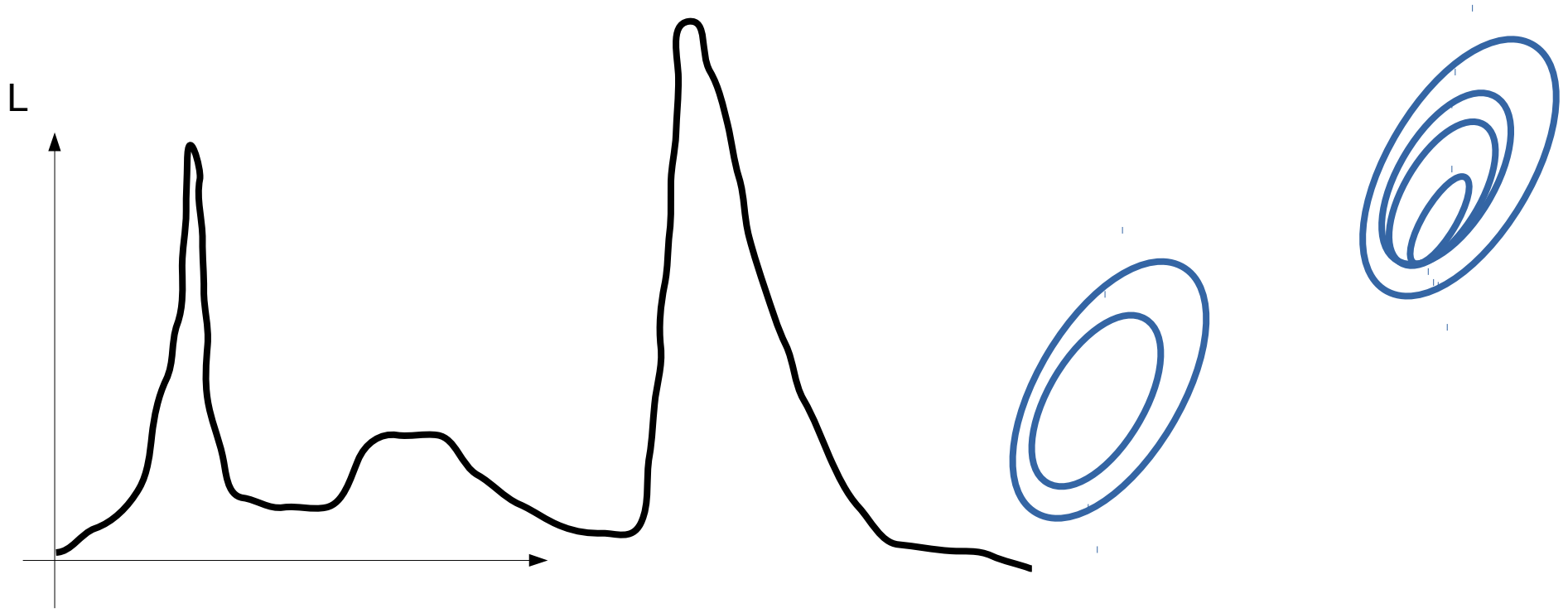


(by Eric Ford)



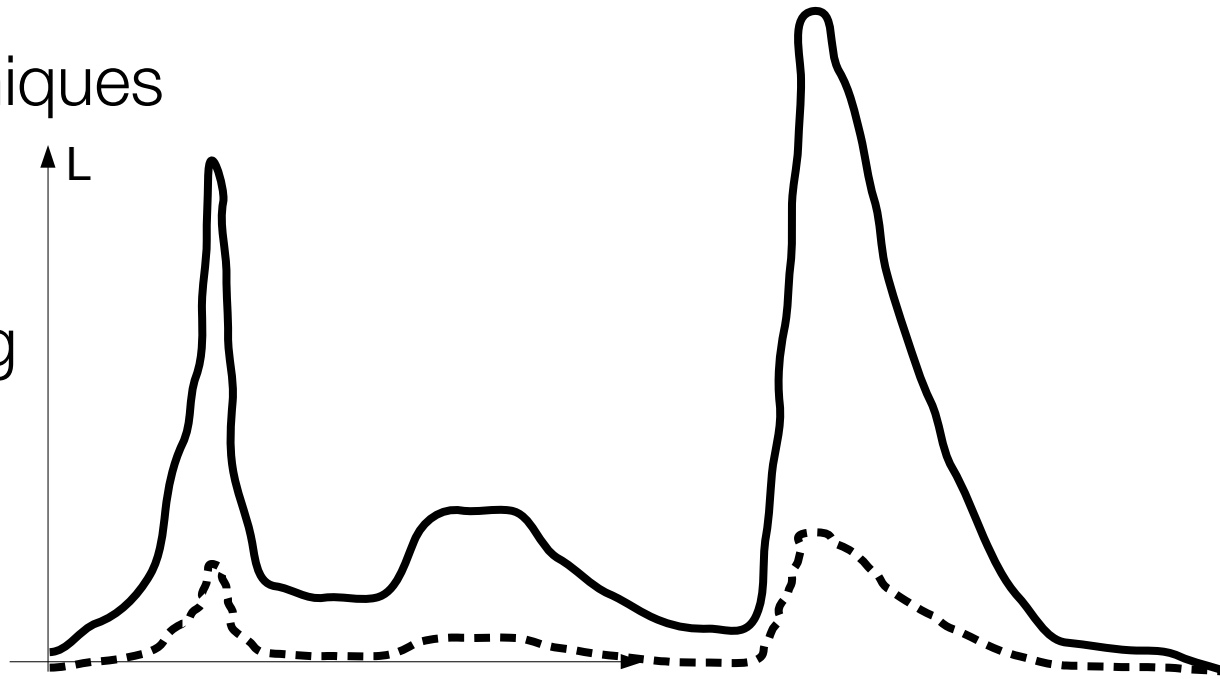
Global optimization

Global maxima



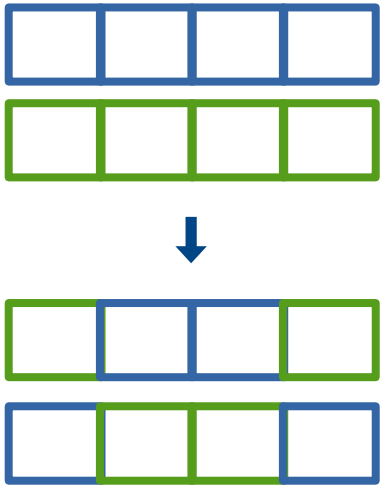
Escaping local maxima: strategies

- Multiple random start positions
  - Augment local techniques
- Make surface easier
  - Tempering/Annealing
- Walker population
  - GW
  - Genetic algorithms (DE)

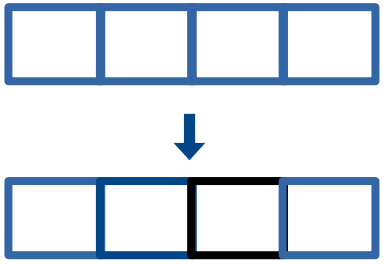


Genetic algorithms

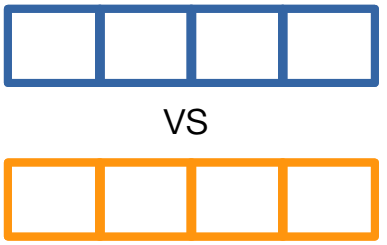
Cross-over



Mutation

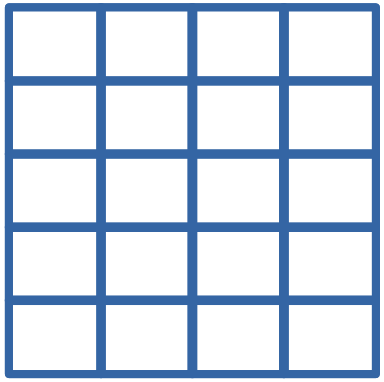


Selection

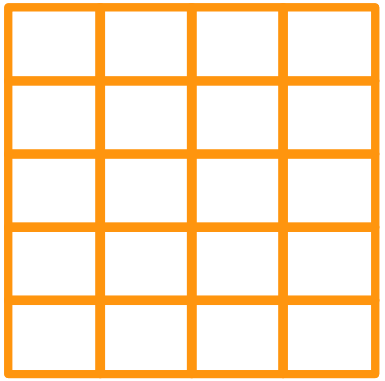


With fitness function (here: L)

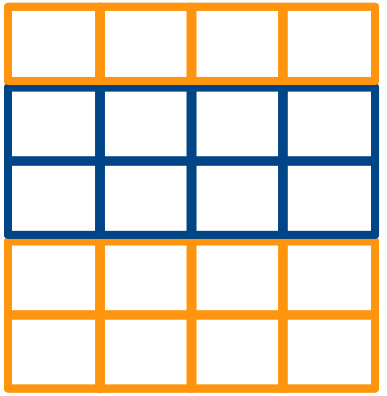
Initial population



Offspring

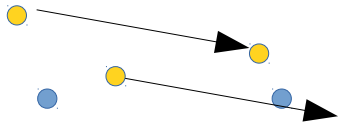


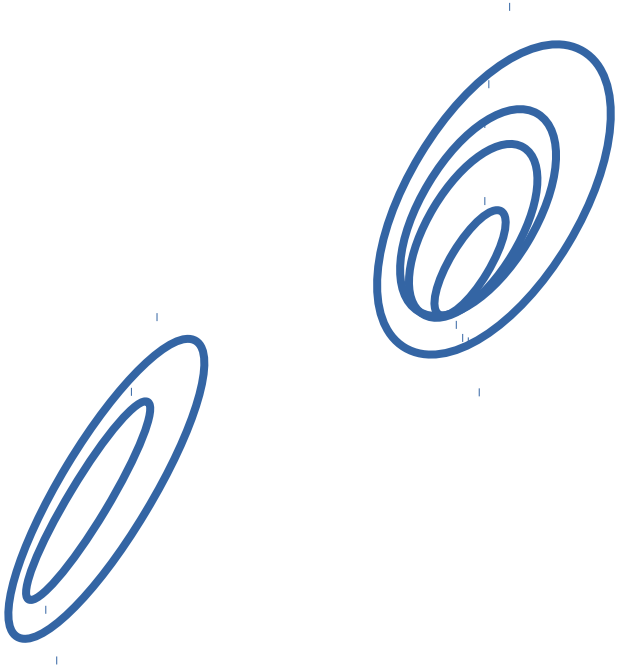
New generation



Genetic algorithms

- Differential evolution





- Zooms into highest regions

`moncar`

Advantages:
- Global
- Robust to degeneracies, auto-tuning proposal
- Works in high-d & non-continuous parameters

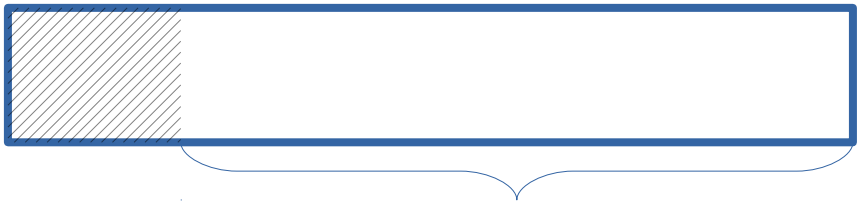
Disadvantages:
- Some tuning parameters
- stopping criterion may not be meaningful
- Does not sample (only best-fit)

Model comparison

Model comparison

Buchner+14

- Empirical models
  - Information content
  - Prediction quality
- Component presence
  - Regions of practical equivalence
- Physical effects
  - Bayesian model comparison
  - Priors often well-justified



https://arxiv.org/abs/1506.02273
Betancourt (2015)

Information criteria

- Akaike information criterion Akaike (1973)
- Is more complex worth storing?

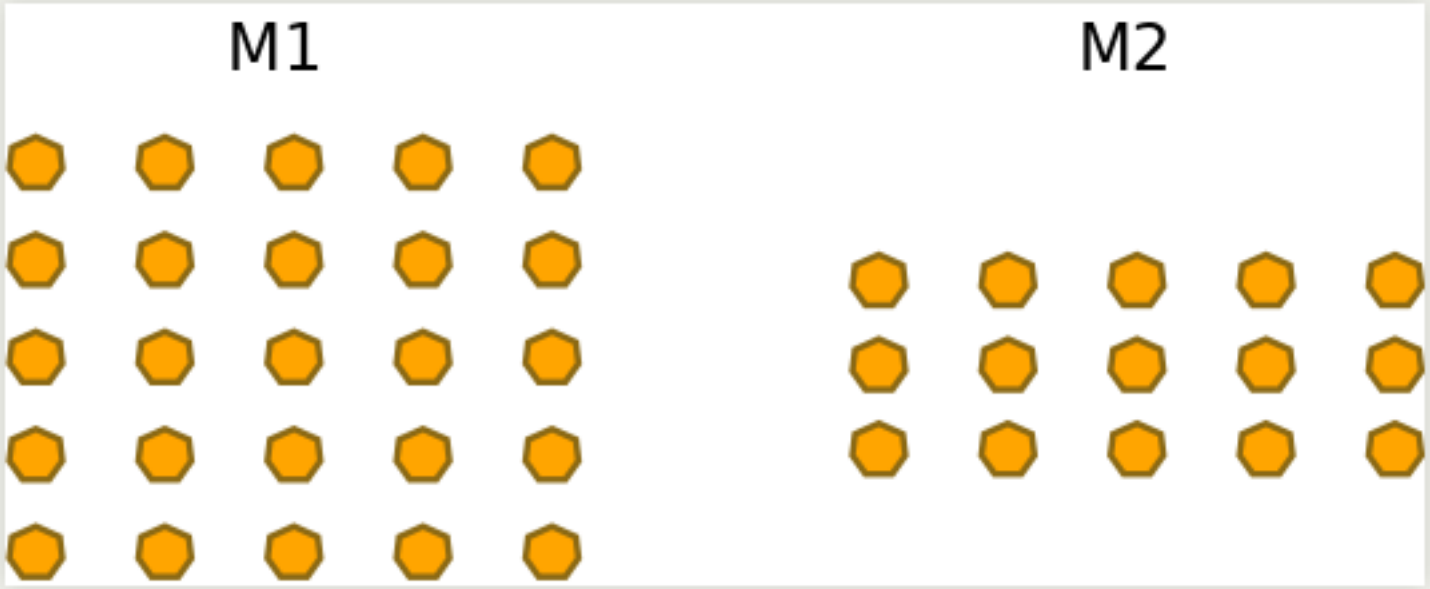
$$AIC = 2 * d - 2 * L_{max}$$

$$AIC = 2 * d + CStat$$

Advantages:
- rooted in information theory
- independent of prior

Disadvantages:
- No uncertainties, thresholds unclear
- …

Two models



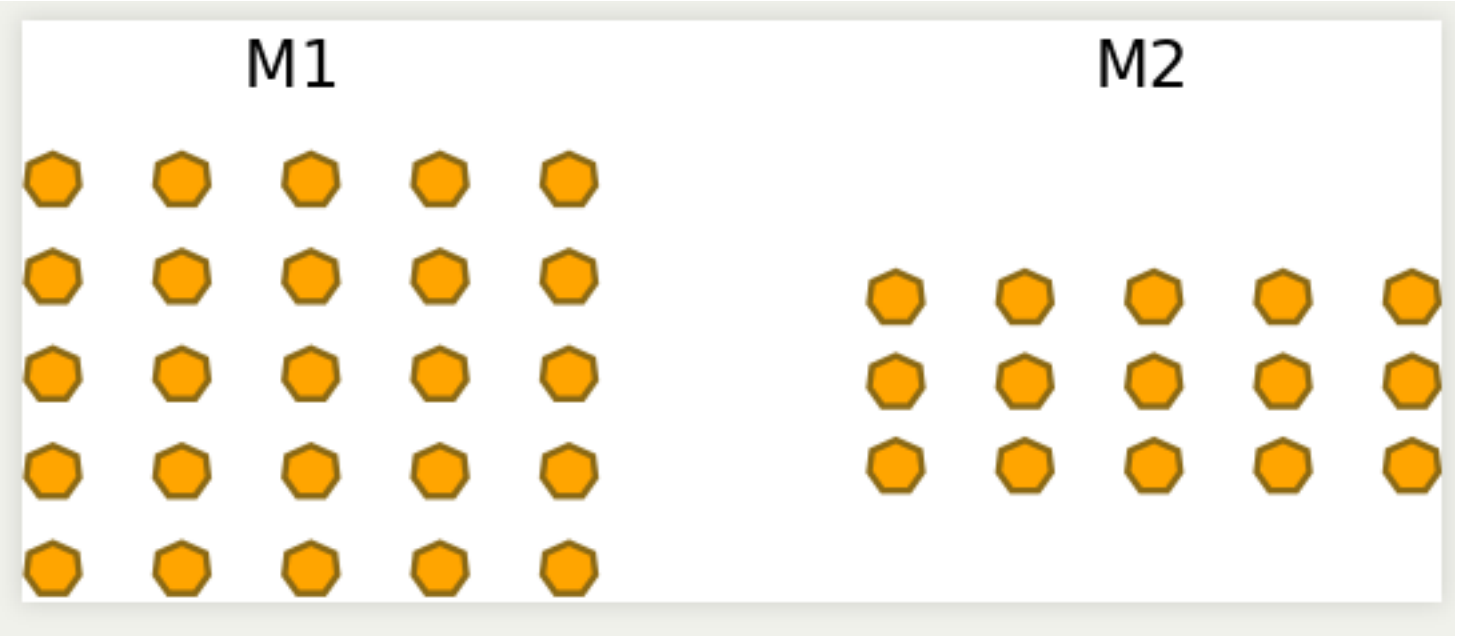
- Compare two parameter spaces by

$$\sum \mathcal{L}\big|_{M1} / \sum \mathcal{L}\big|_{M2}$$

- How many coins to put in M1, M2?
- model prior

Punishing prediction diversity

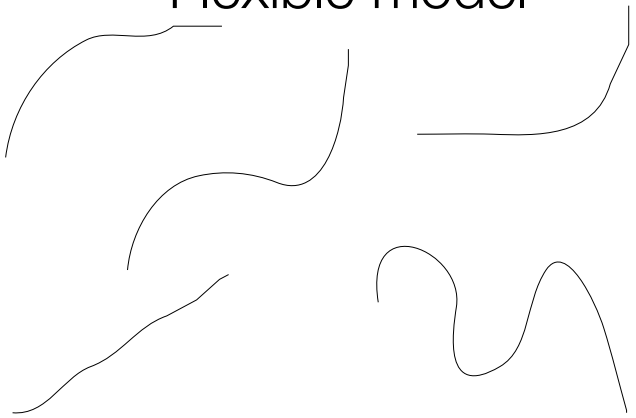
(not number of parameters)

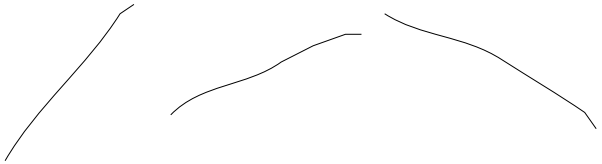


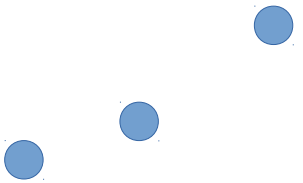
Flexible model

Inflexible model

Data







L high, V tiny

L medium, V medium

What to do with Z

- Z1, Z2

$$\frac{p(M1|D)}{p(M2|D)} = \frac{Z1 \cdot p(M1)}{Z2 \cdot p(M2)}$$

Posterior odds ratio

Bayes factor

Prior odds ratio

What to do with Z

- Z1, Z2

$$\frac{p(M1|D)}{p(M2|D)} = \frac{Z1 \cdot p(M1)}{Z2 \cdot p(M2)}$$

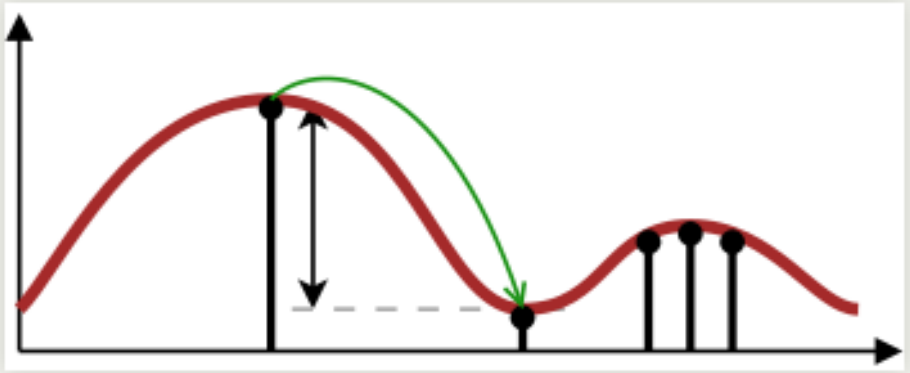
$$\frac{p(M_1|D)}{\sum p(M_i|D)} = \frac{Z_1 \cdot p(M_i)}{\sum_i Z_i \cdot p(M_i)}$$

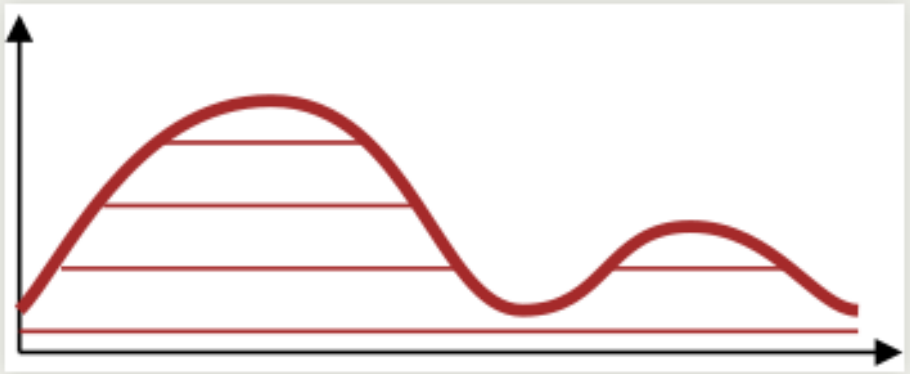
- model priors: leave to reader or motivated by theory
- Discard highly improbable model or marginalise
- Does $\frac{p(M1|D)}{p(M2|D)} = 3/1$ mean M2 is correct in a quarter of the cases?

Global sampling

nested sampling idea

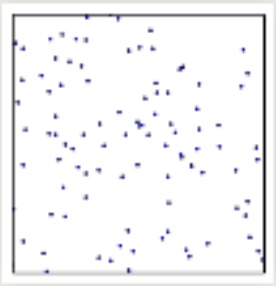
- MCMC: only consider likelihood ratios. Integration by vertical slices
- nested sampling: compute geometric size at various likelihood thresholds
- orthogonal, unique re-ordering of volume by likelihood



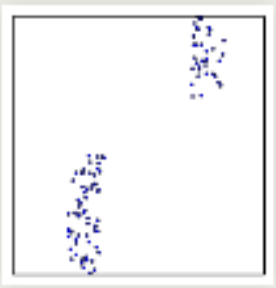


$$\sum \underbrace{\text{Shrinkage} \times \text{Likelihood}}_{\text{Importance of shell}} = Z$$

nested sampling algorithm



- Start with volume 1, draw randomly uniformly 200 points
- remove one, volume shrinks by 1/200.



- draw a new one excluding the removed volume
- Unique ordering of space required: via likelihood

draw a new uniformly random point, with higher likelihood

(the crux of nested sampling)

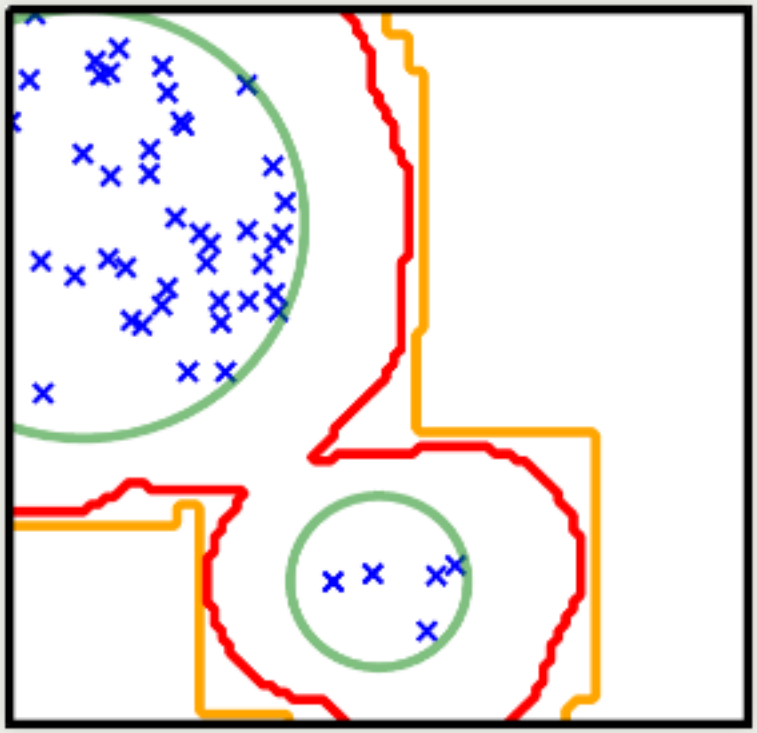
- Scanning up vertically, done at some point
- converges (flat at highest likelihood)

Missing ingredients

- MCMC: Insert tuned transition kernel
- NS: Insert constrained drawing algorithm
  - General solutions: MultiNest, MCMC, HMCMC, Galilean, RadFriends, PolyChord

RadFriends / MultiNest

- **Use existing points** to guess contour
- Expand contour a little bit
- Draw uniformly from contour
- Reject points below likelihood threshold
- RadFriends: Compute distance at which every point has a neighbor. Bootstrap (Leave out) for safety.
- MultiNest clusters and uses ellipses



Animation:

https://johannesbuchner.github.io/mcmc-demo/app.html#RadFriends-NS,standard
(via chi-feng.github.io)

What to do with Z

- Z1, Z2

$$\frac{p(M1|D)}{p(M2|D)} = \frac{Z1 \cdot p(M1)}{Z2 \cdot p(M2)}$$

$$\frac{p(M_1|D)}{\sum p(M_i|D)} = \frac{Z_1 \cdot p(M_i)}{\sum_i Z_i \cdot p(M_i)}$$

- model priors: leave to reader or motivated by theory
- Discard highly improbable model or marginalise
- Does $\frac{p(M1|D)}{p(M2|D)} = 3/1$ mean M2 is correct in a quarter of the cases?

What to do with Z

- Z1, Z2

$$\frac{p(M1|D)}{p(M2|D)} = \frac{Z1 \cdot p(M1)}{Z2 \cdot p(M2)}$$

Posterior odds ratio

Bayes factor

Prior odds ratio

What to do with Z

- Z1, Z2

$$\frac{p(M1|D)}{p(M2|D)} = \frac{Z1 \cdot p(M1)}{Z2 \cdot p(M2)}$$

$$\frac{p(M_1|D)}{\sum p(M_i|D)} = \frac{Z_1 \cdot p(M_i)}{\sum_i Z_i \cdot p(M_i)}$$

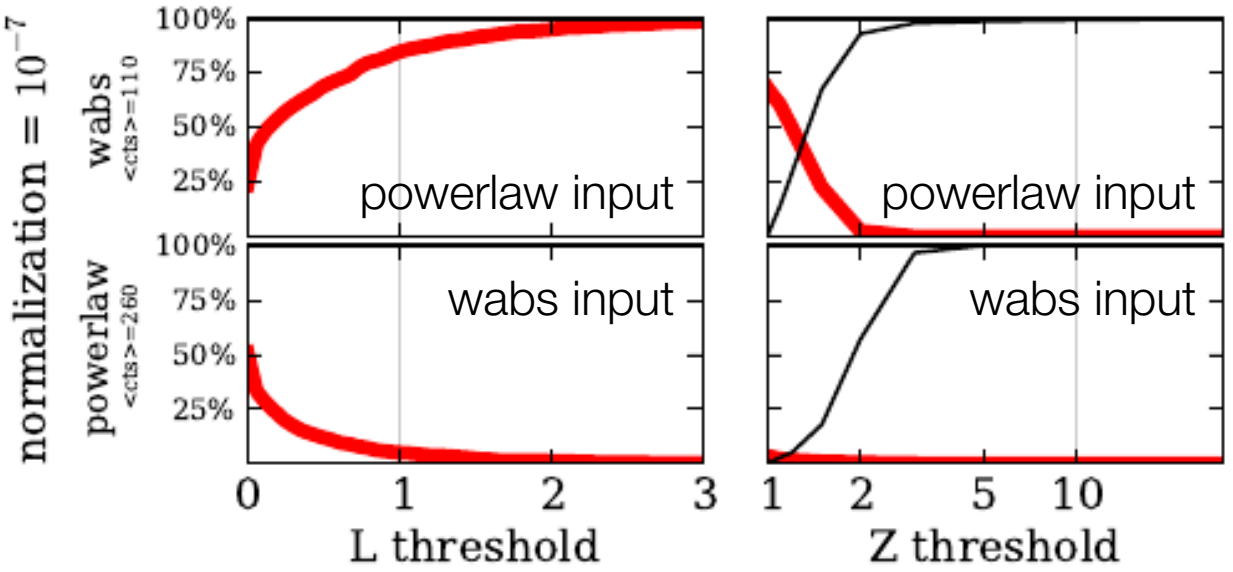
- model priors: leave to reader or motivated by theory
- Discard highly improbable model or marginalise
- Does $\frac{p(M1|D)}{p(M2|D)} = 3/1$ mean M2 is correct in a quarter of the cases?

Calibrating model decisions

- Model probabilities → decisions
- False decision rate (false positives/negatives)
  - Monte Carlo simulations (parametric bootstrap)

Buchner+14

Calibrating model decisions



Buchner+14

False negatives
Non-decisions

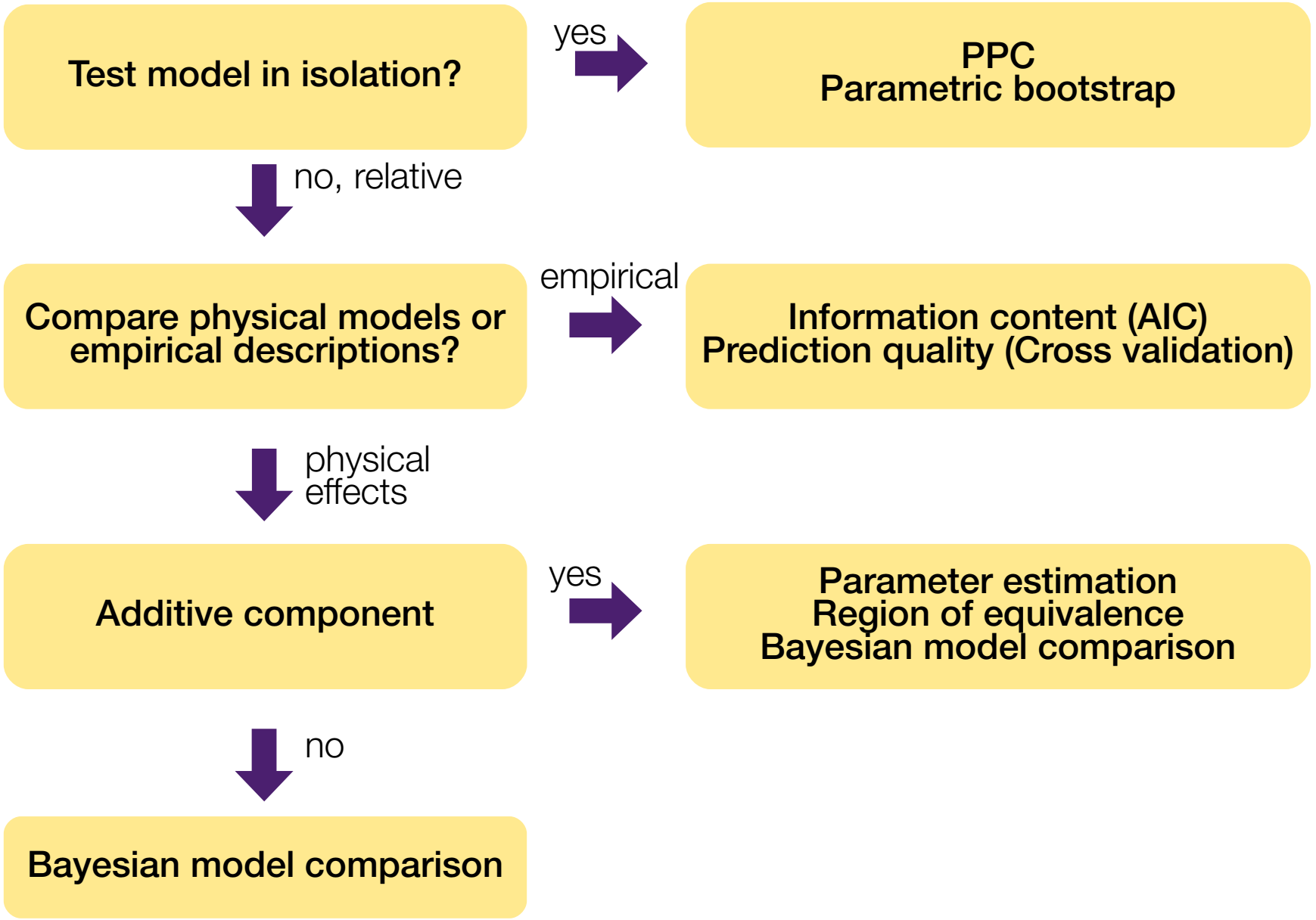
Advantages:
- Get rid of parameter prior dependences
- Have frequentist properties of Bayesian method
- Completely Bayesian treatment + decisions

Disadvantages:
- Can be computationally expensive

Frequentist properties of Bayesian methods

- Make decisions
  - Is parameter greater than C?
  - Is this model “better” than the other?
- Parametric bootstrap
  - Monte Carlo simulation allow arbitrary complexity

Model comparison



Agenda

- Yesterday:
  - Basic statistics, problem setup
  - Parameter estimation methods, credible & confidence intervals
  - Model comparison methods
  - Visualisations
- Today:
  - Extended sources, calibration
  - Stacking information
  - Discussion & Questions
  - Practical pointers & Wrap-up
- Not covered:
  - Tools
  - Pile-up & variability

Spectroscopy
+ lower E
+ higher E
+ imaging
+ time
Outside spectroscopy

Spectra with few counts

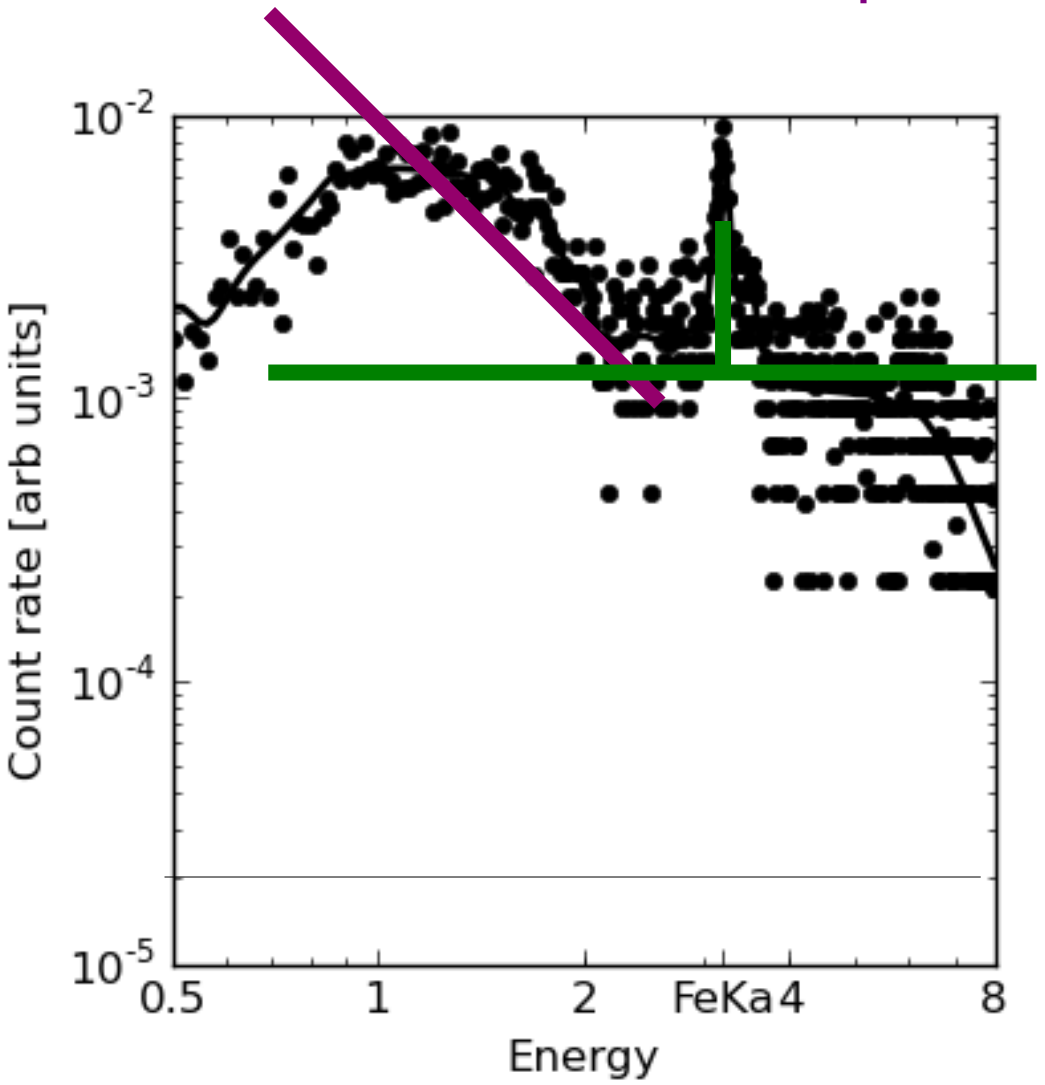
Spectra with few counts

- Are nothing special
- Poisson likelihood + good background handling
- 0 counts

- Think in terms of allowed regions

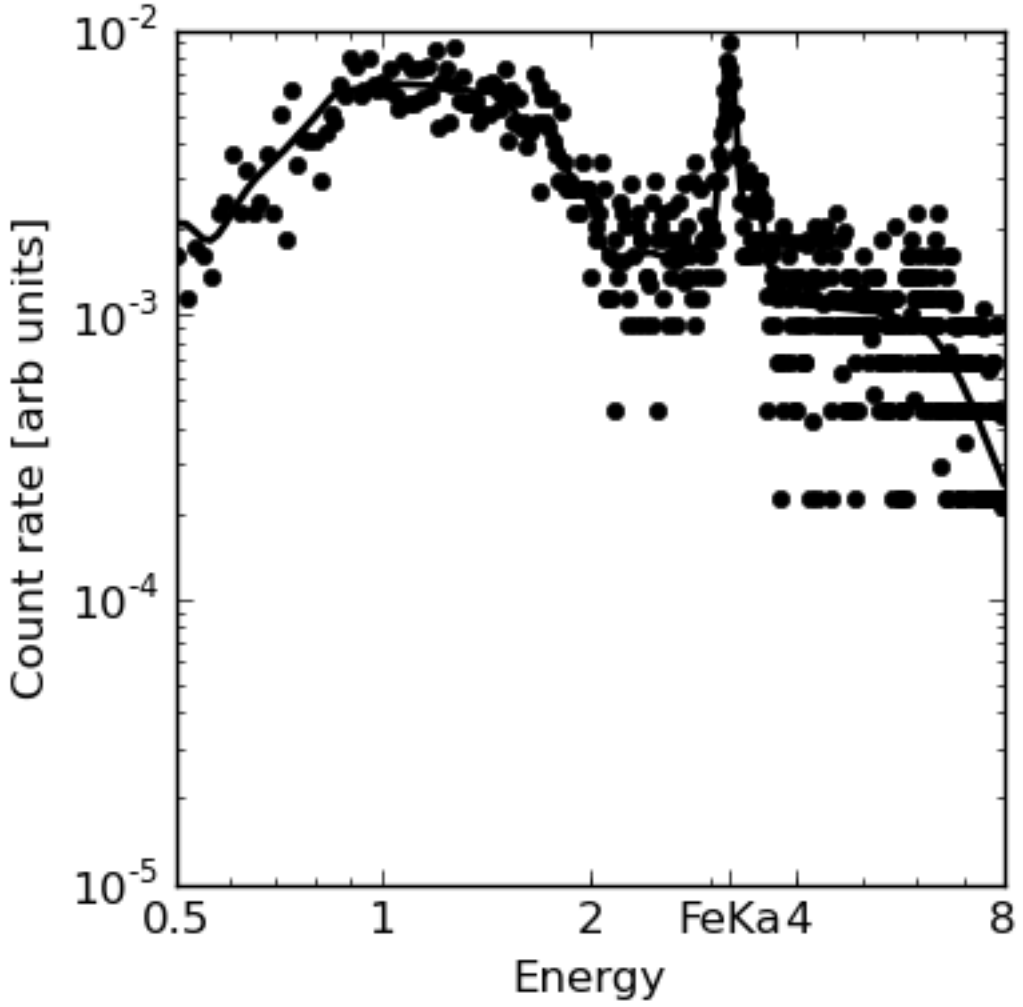
L, N_H from X-ray spectrum

Scattered Powerlaw component

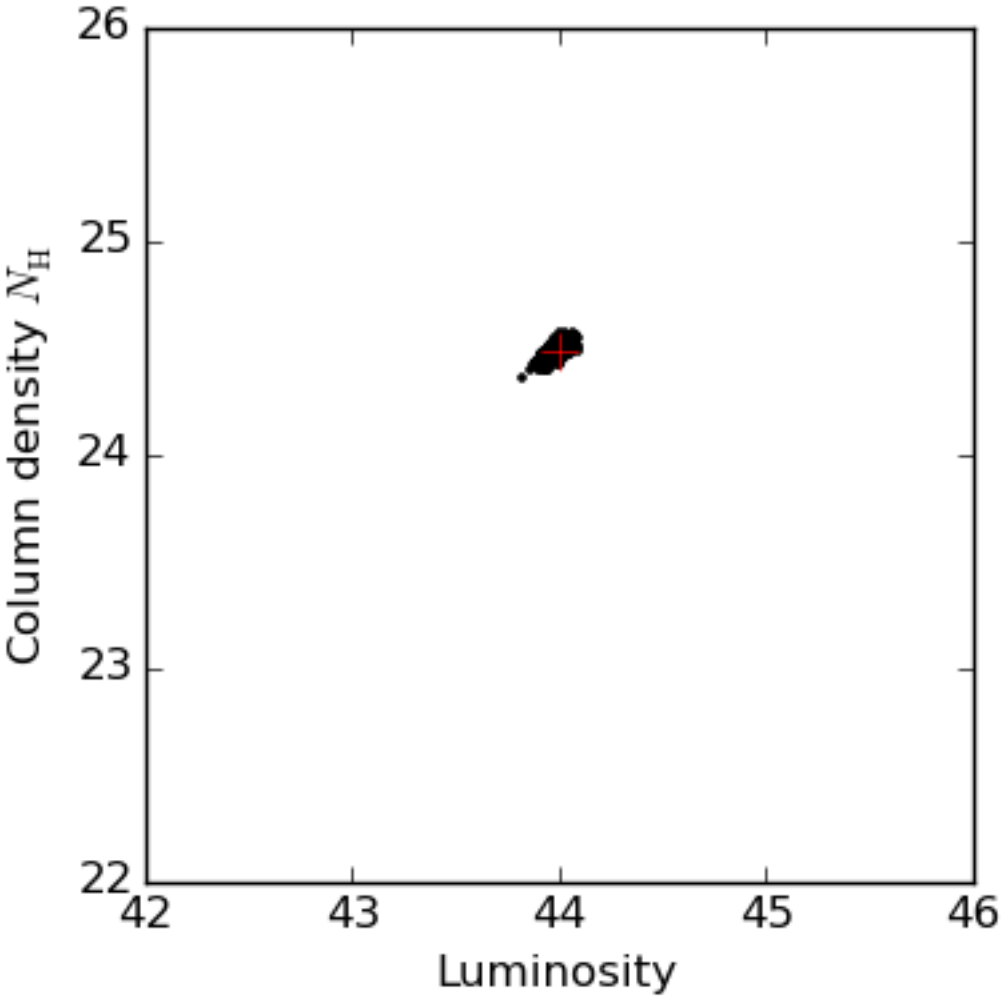


CTK flat spectrum +
FeK line

L, N_H from X-ray spectrum



Probability cloud



Intrinsic parameter distributions

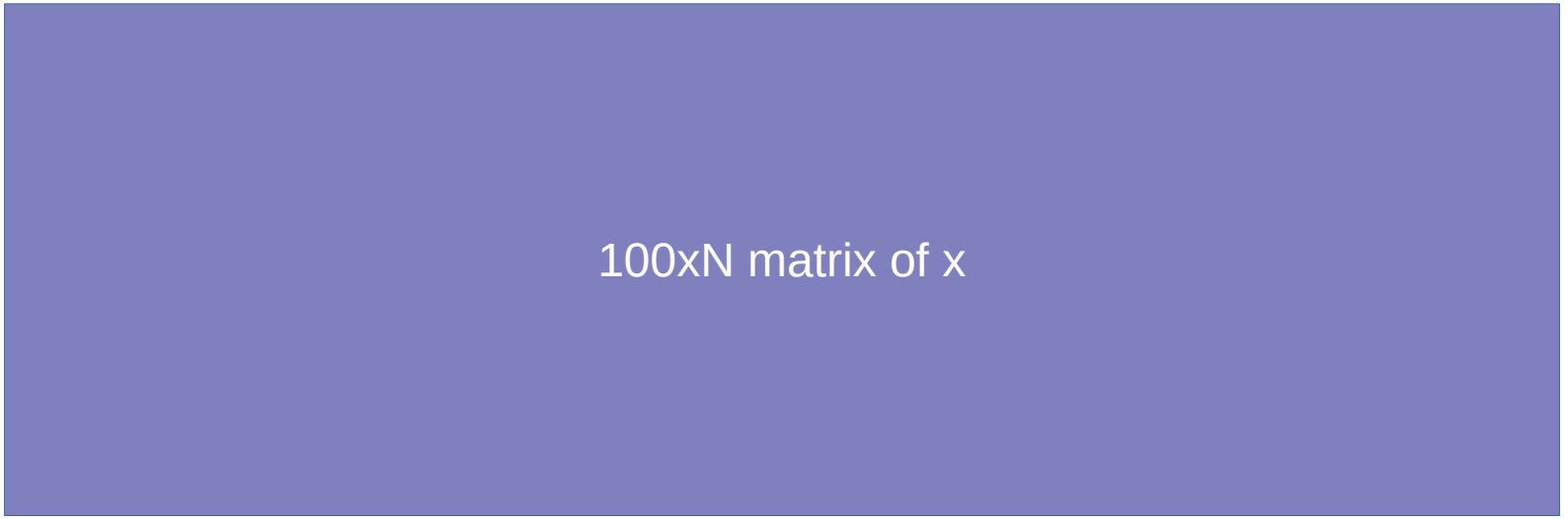
Example

- Measurement gave

  x1 = 4+-1

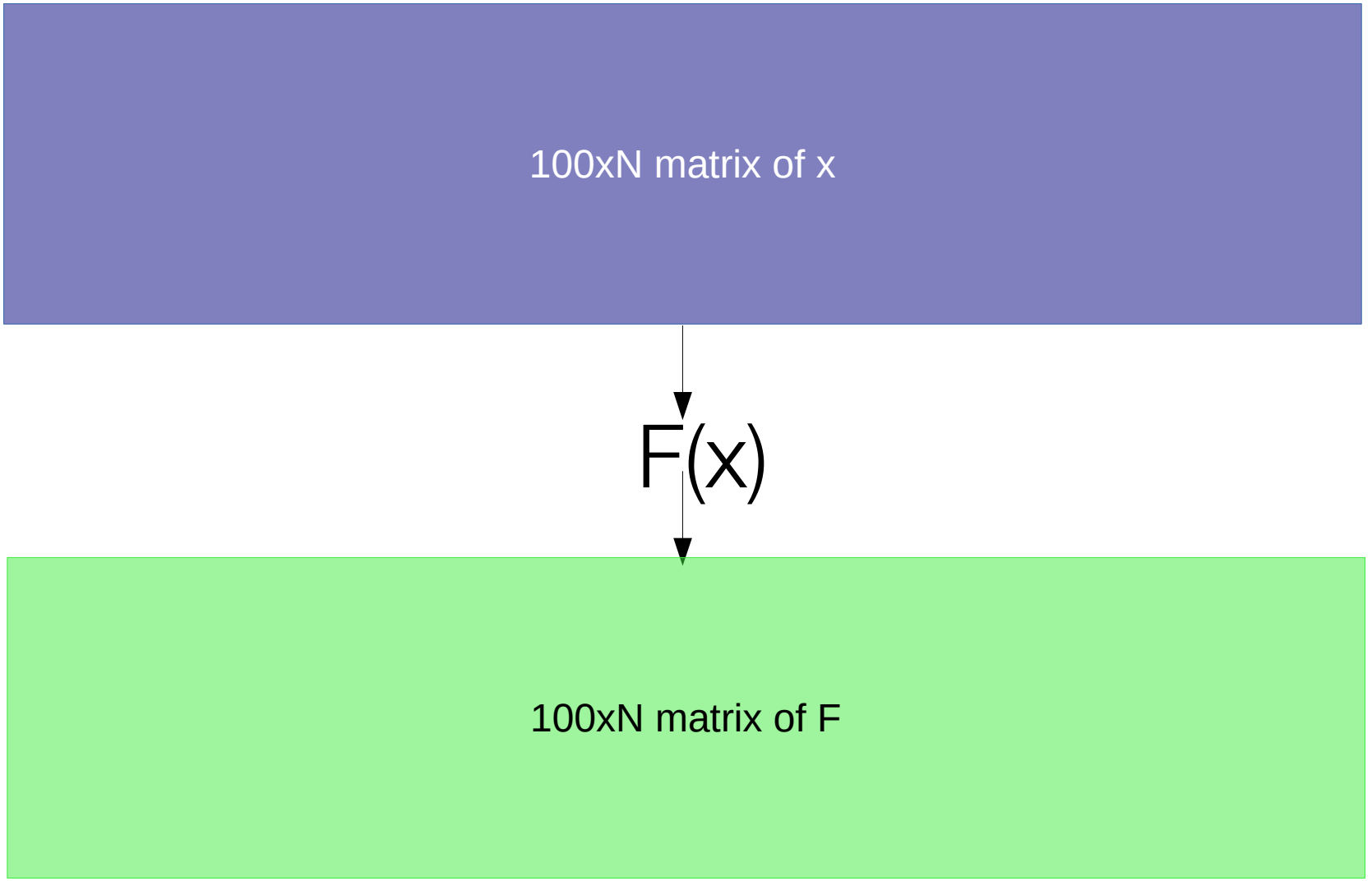
  x2 = 5+-0.1

- Generate 100 samples for each



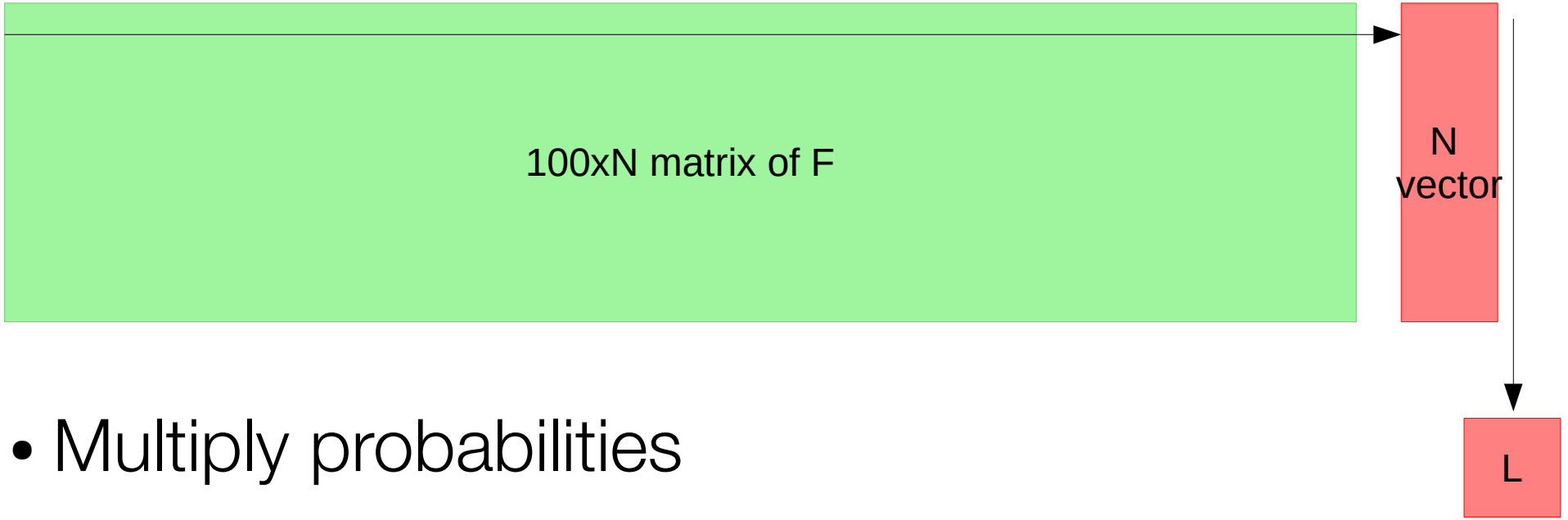
Example

- Evaluate model



Example

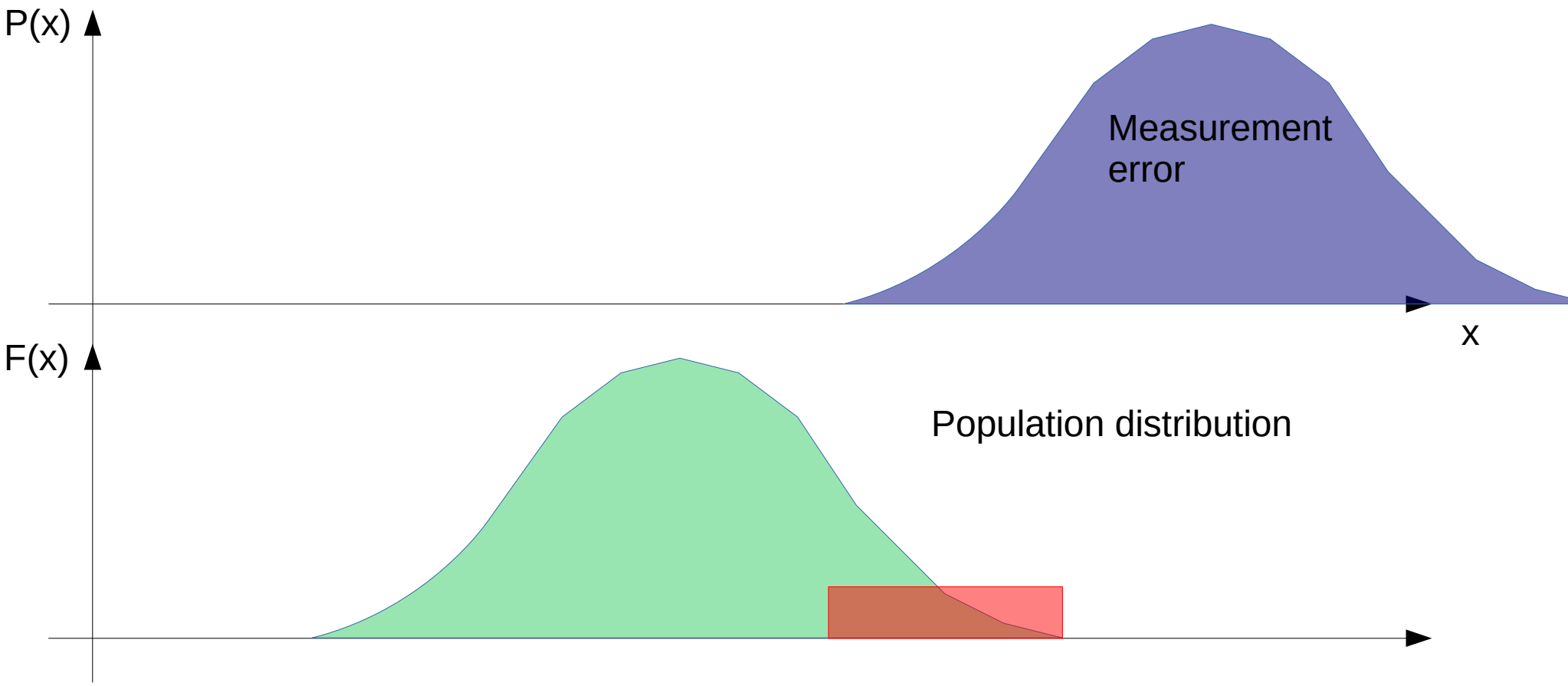
- Sum probabilities



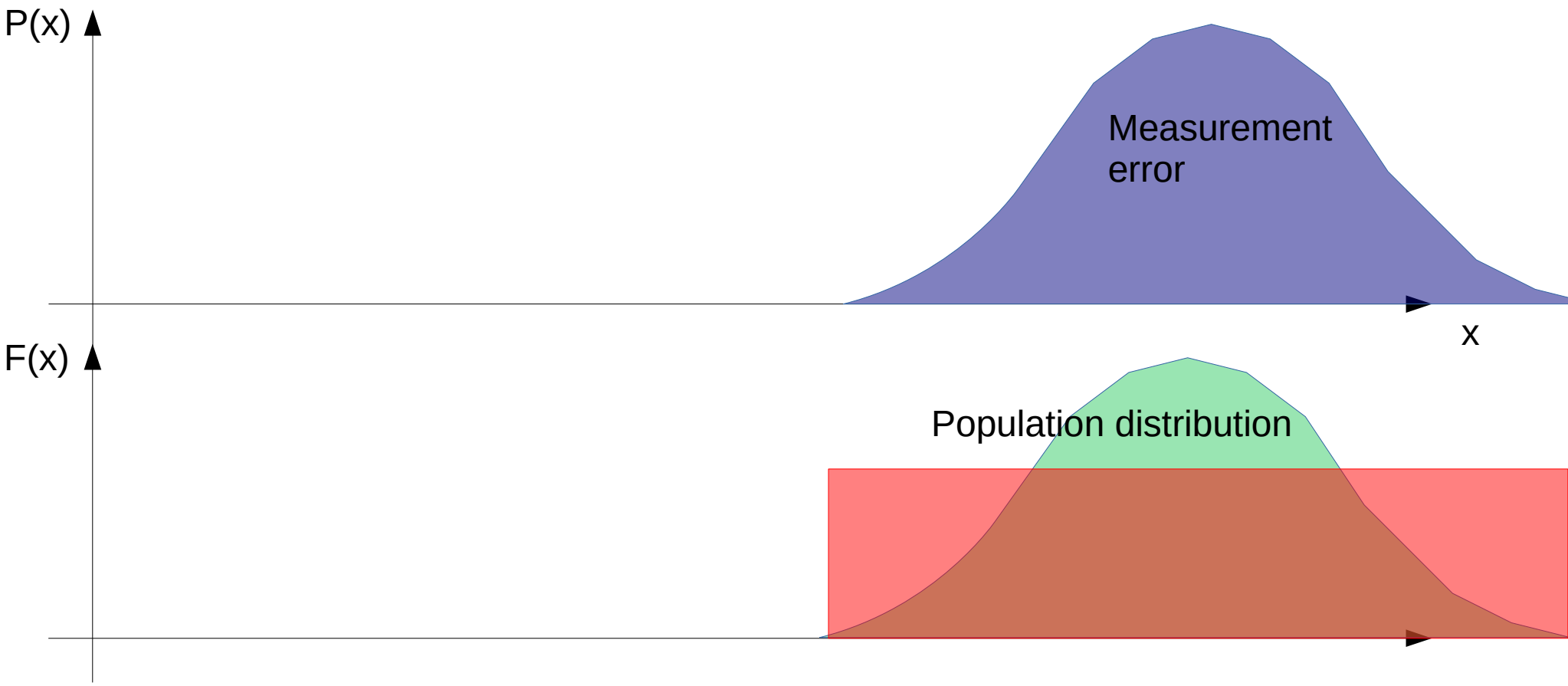
- Multiply probabilities

- Then try out other model parameters

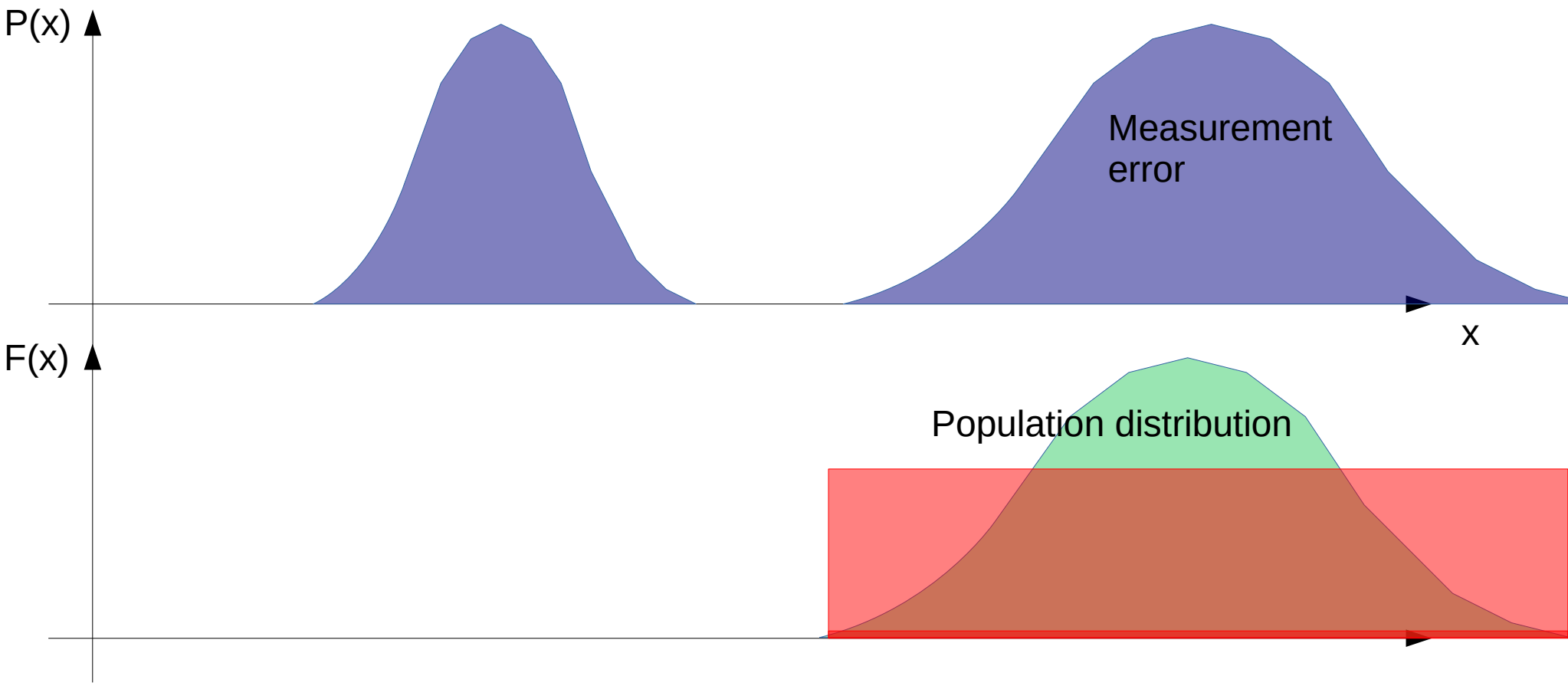
Graphic explanation



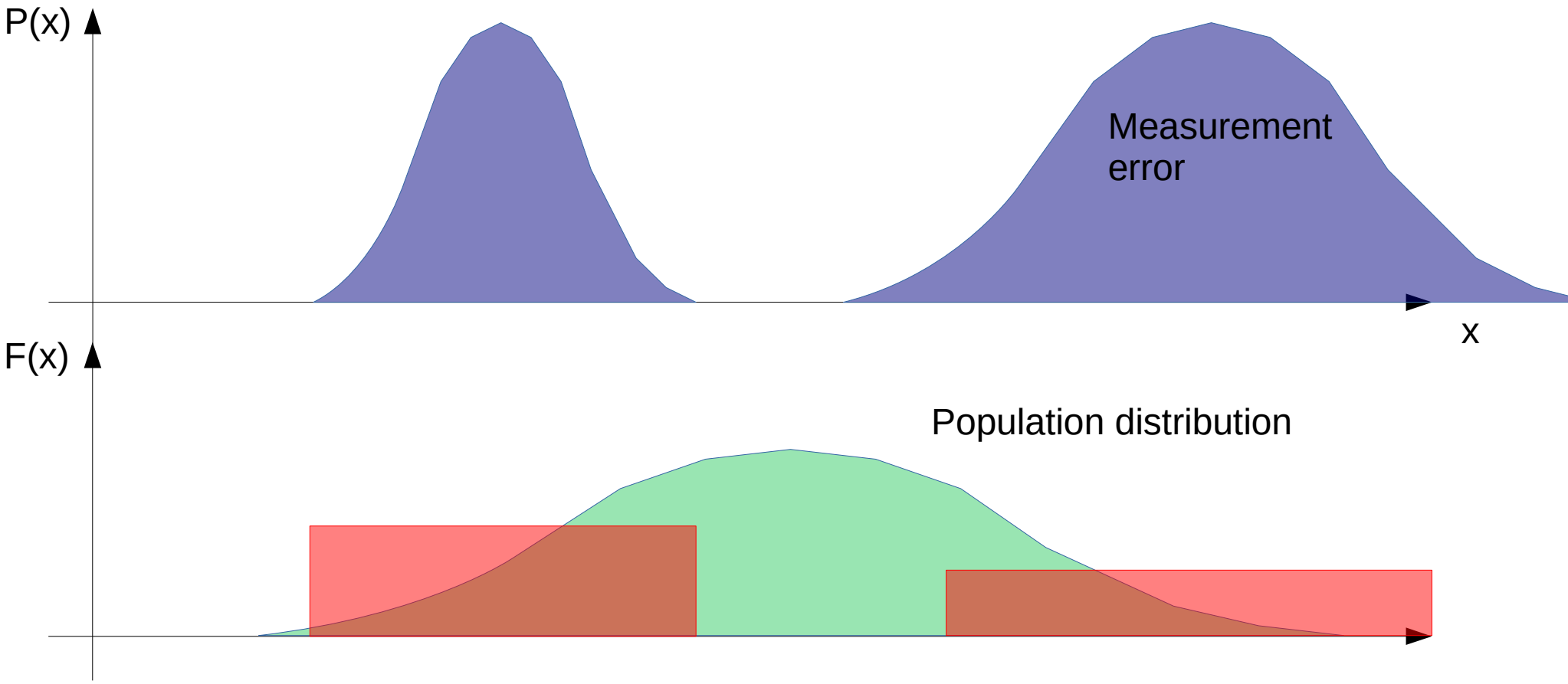
Graphic explanation



Graphic explanation

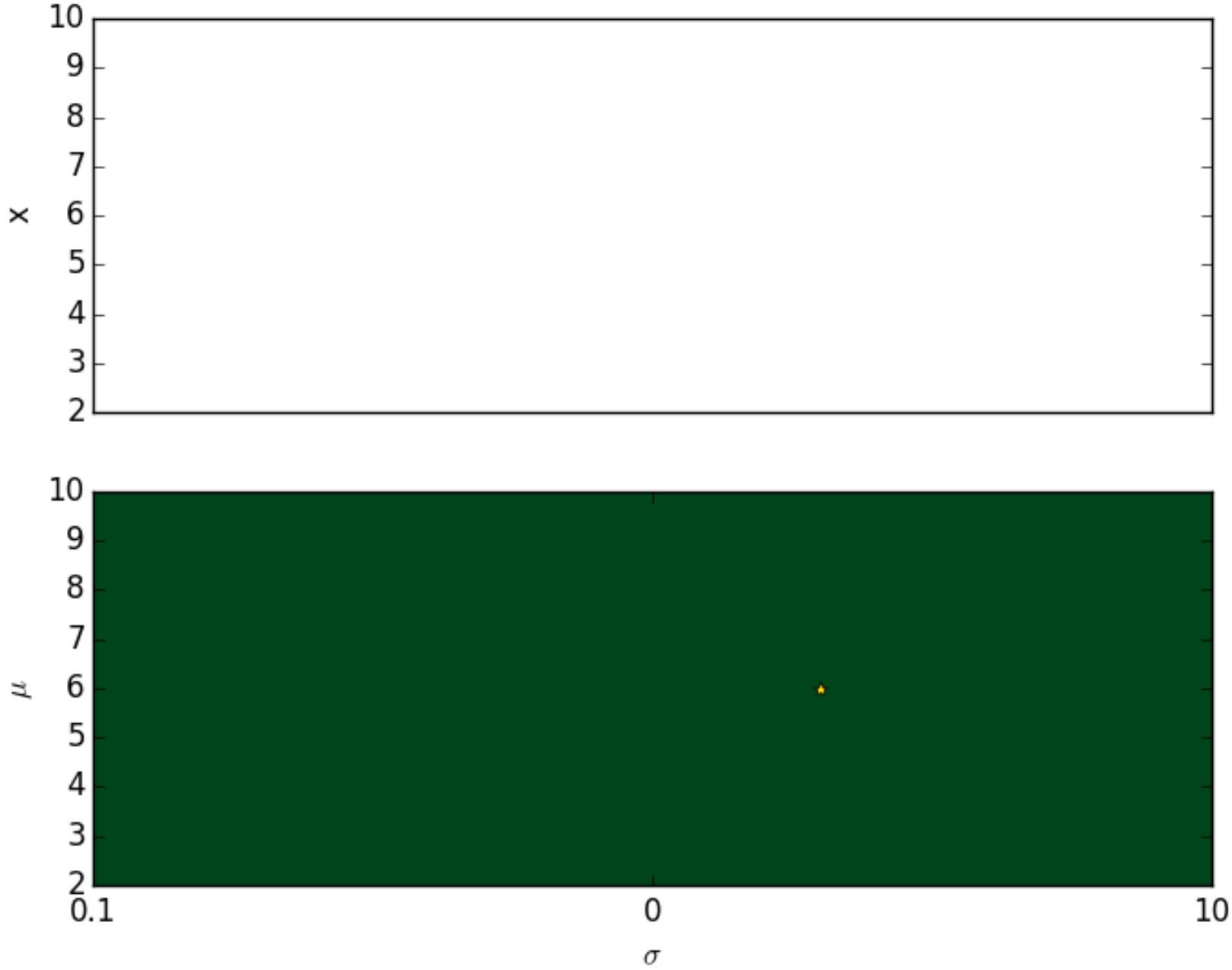


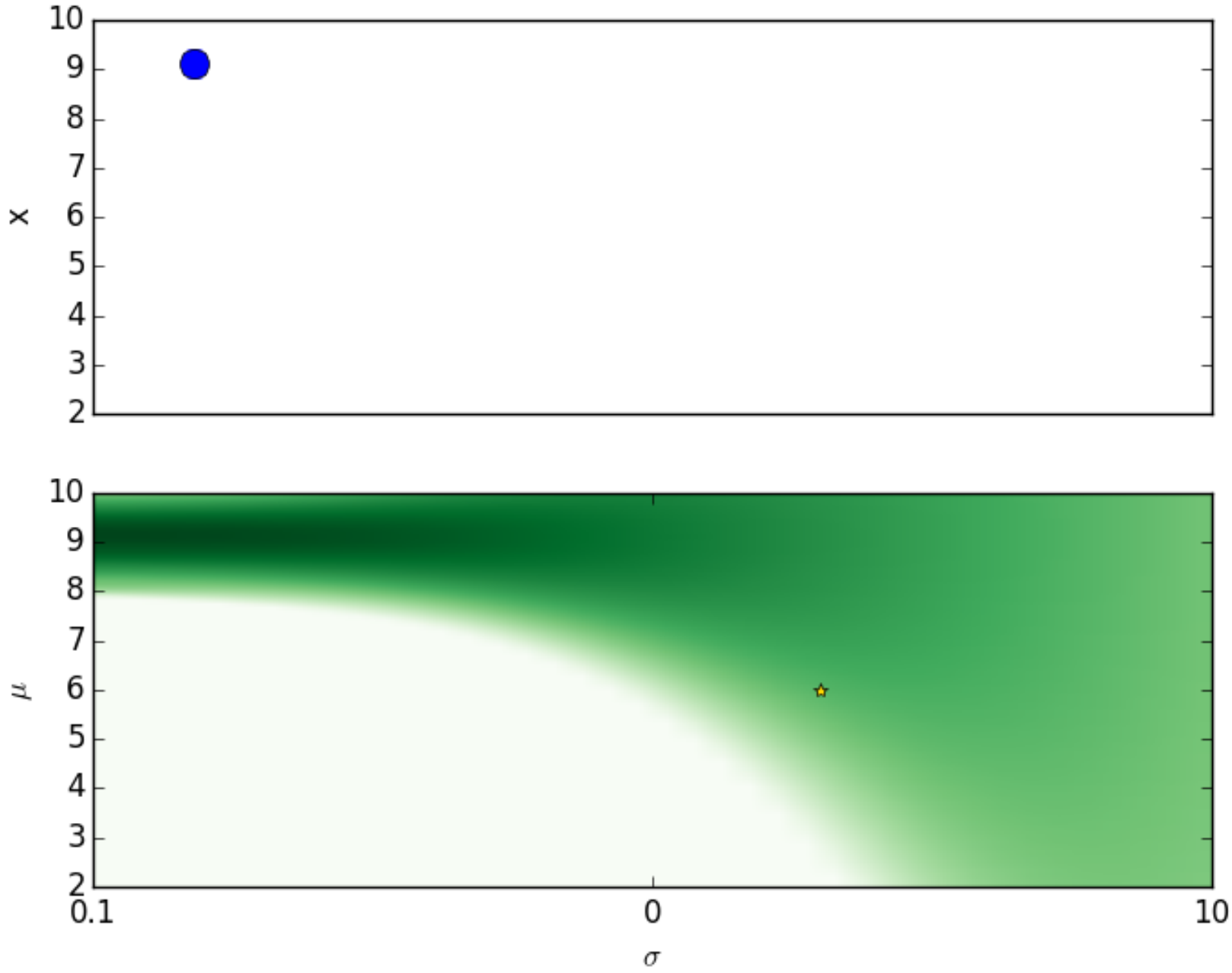
Graphic explanation

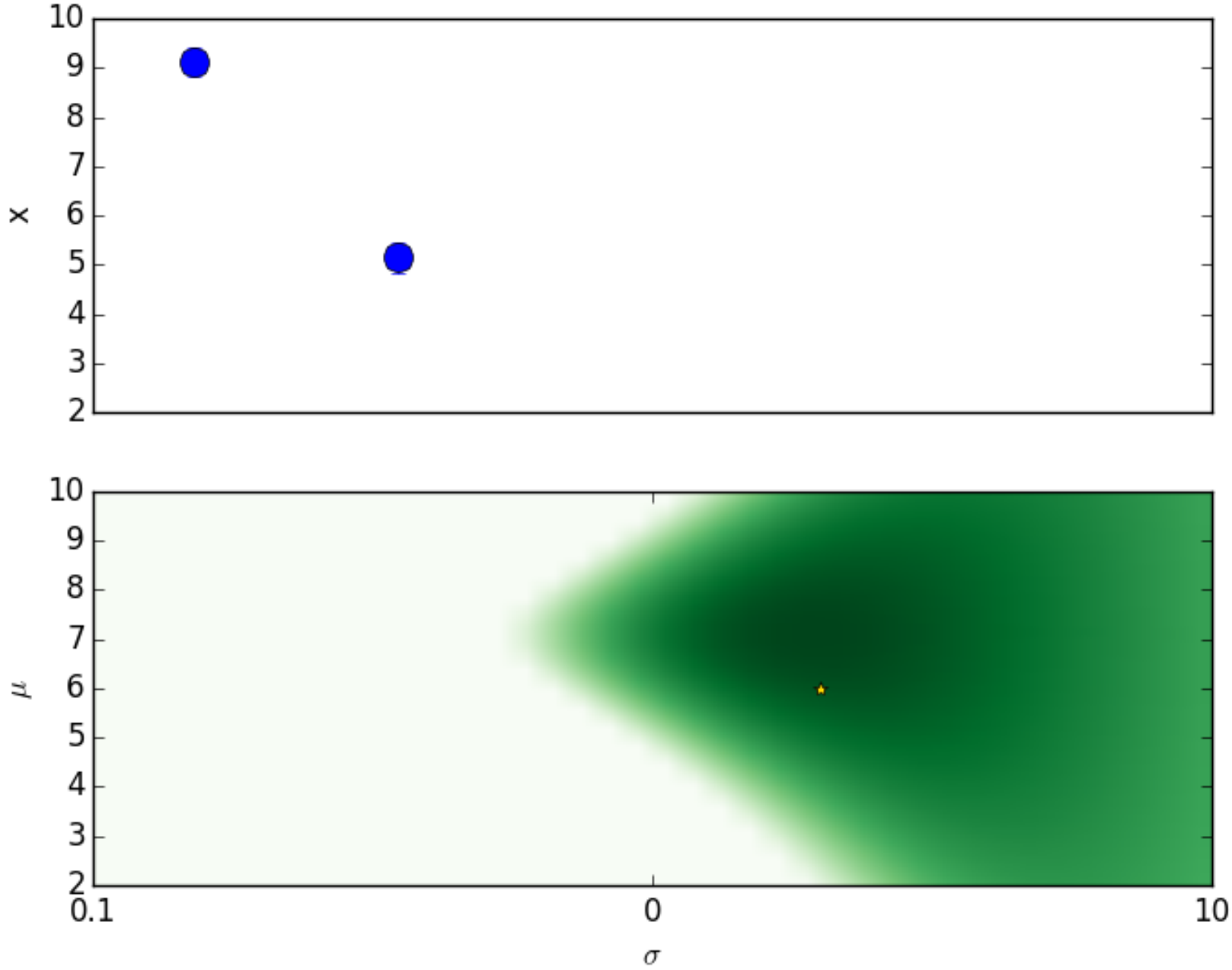


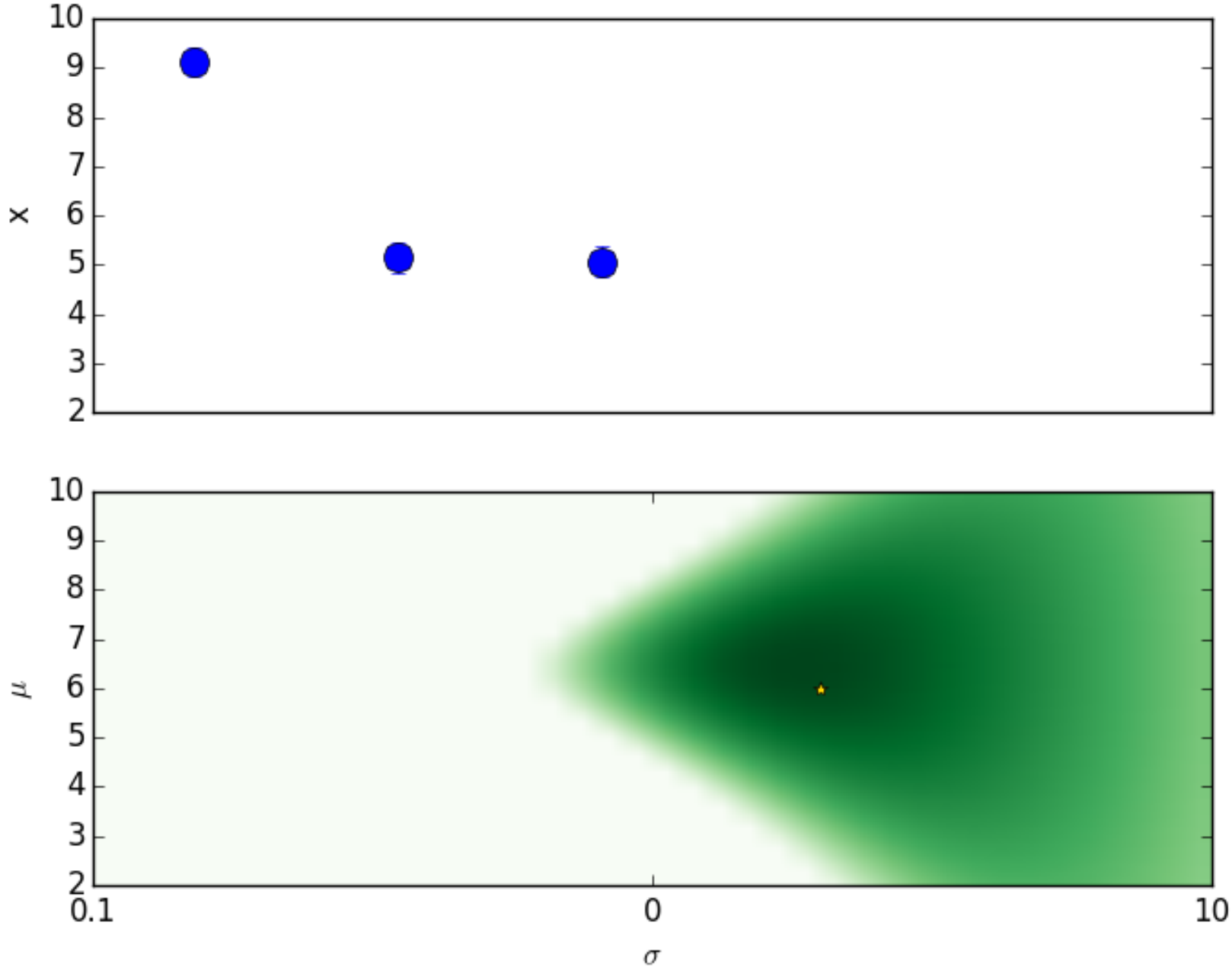
Behaviour

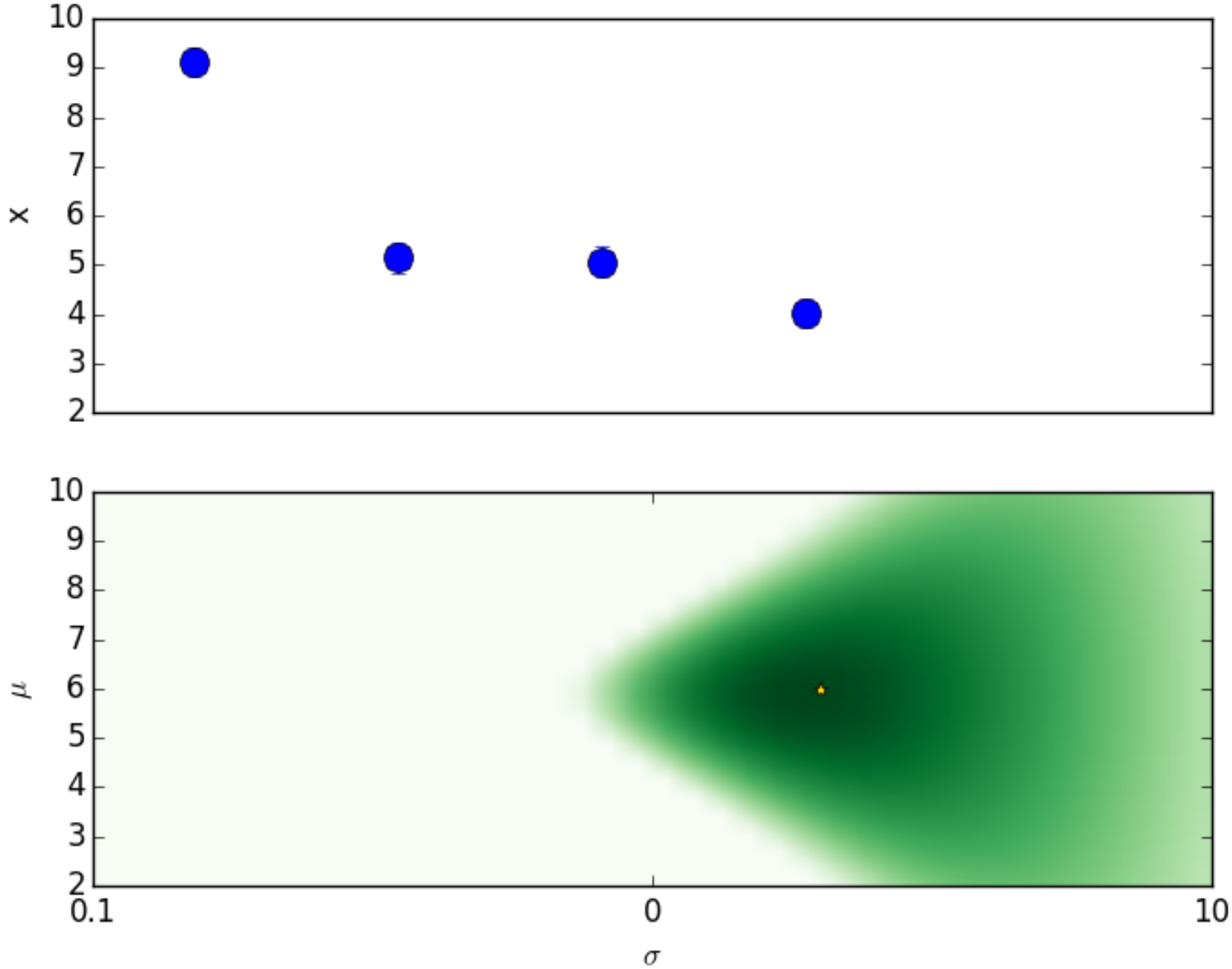
- Generate from 6+-2 with measurement errors

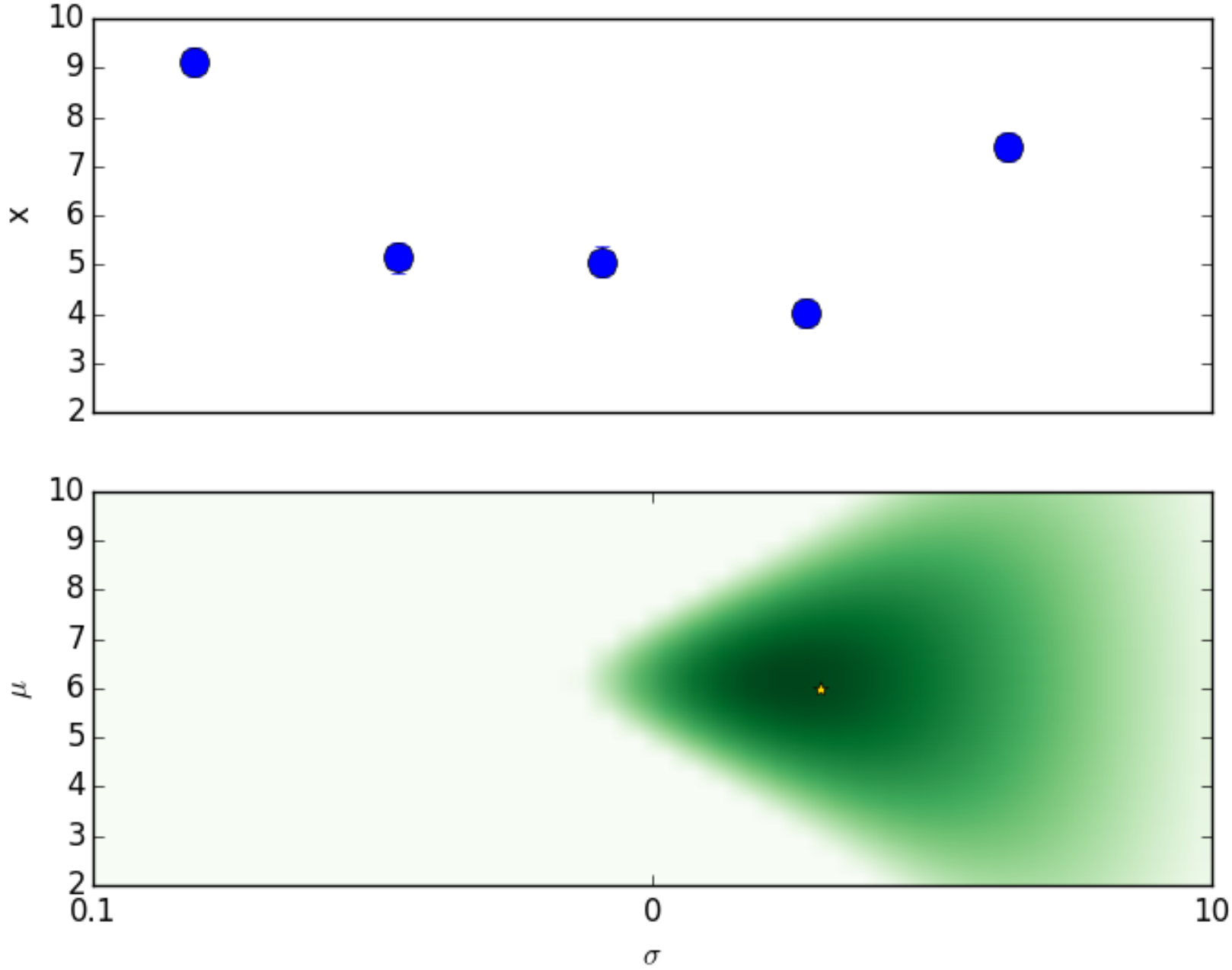

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

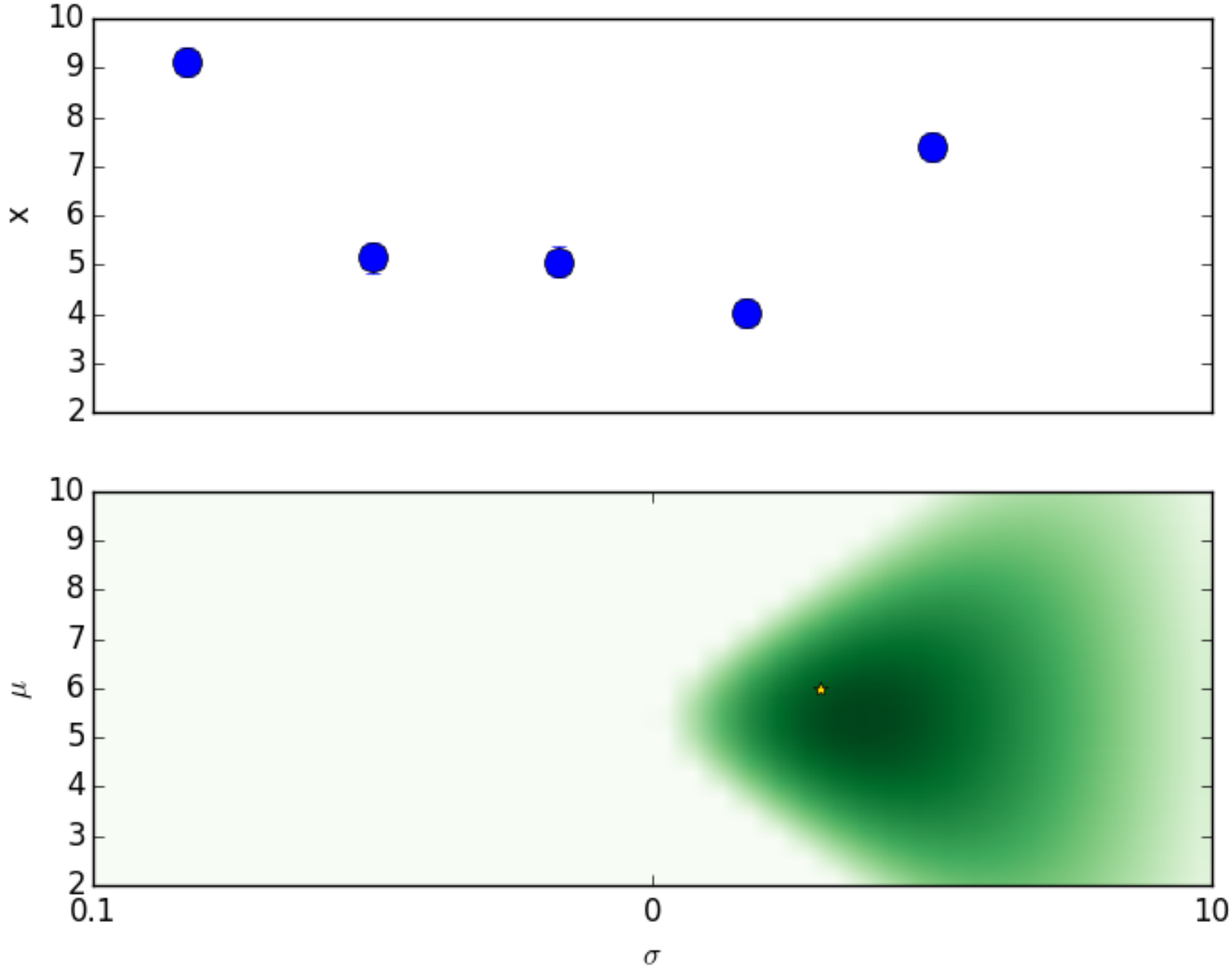

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

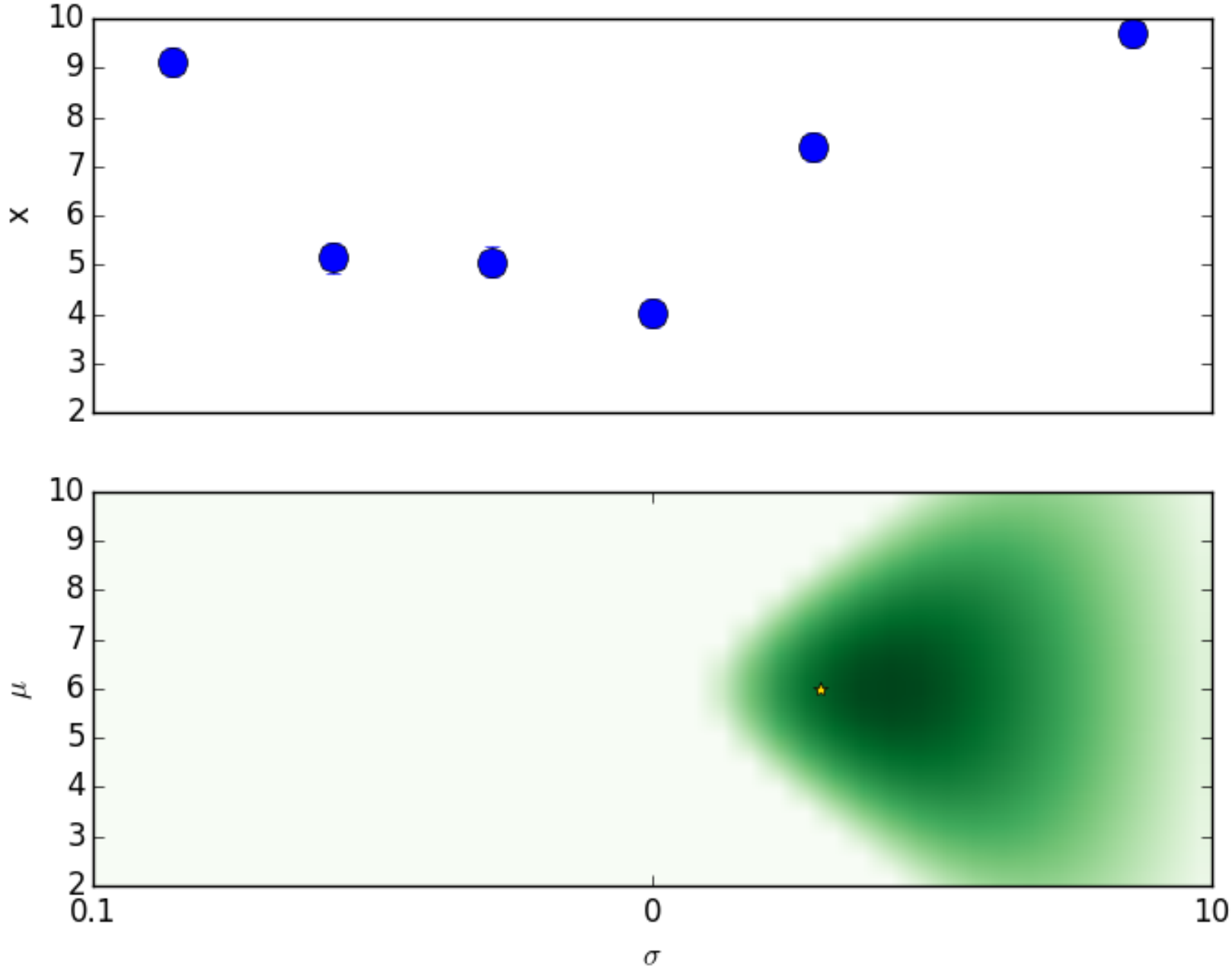

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

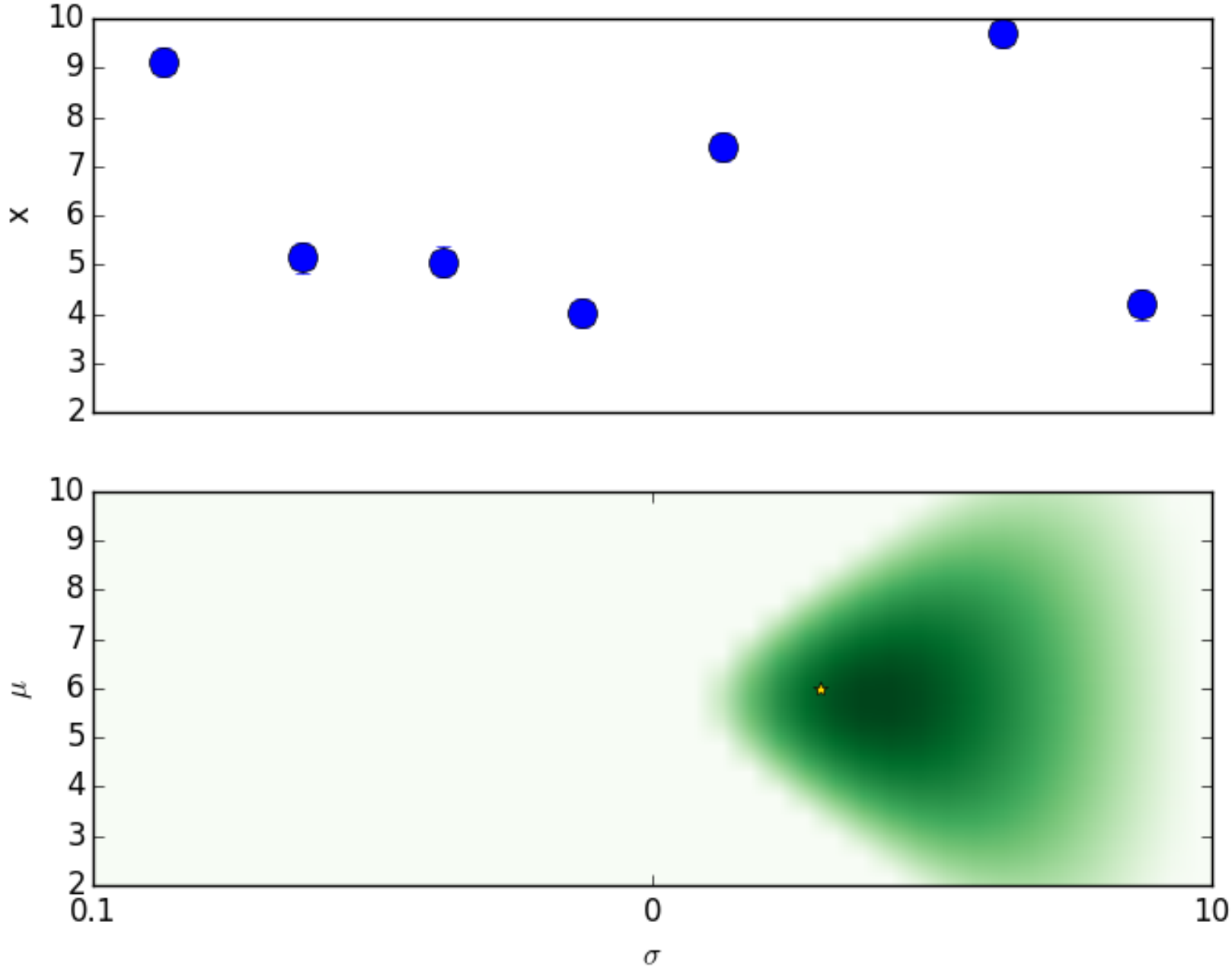

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

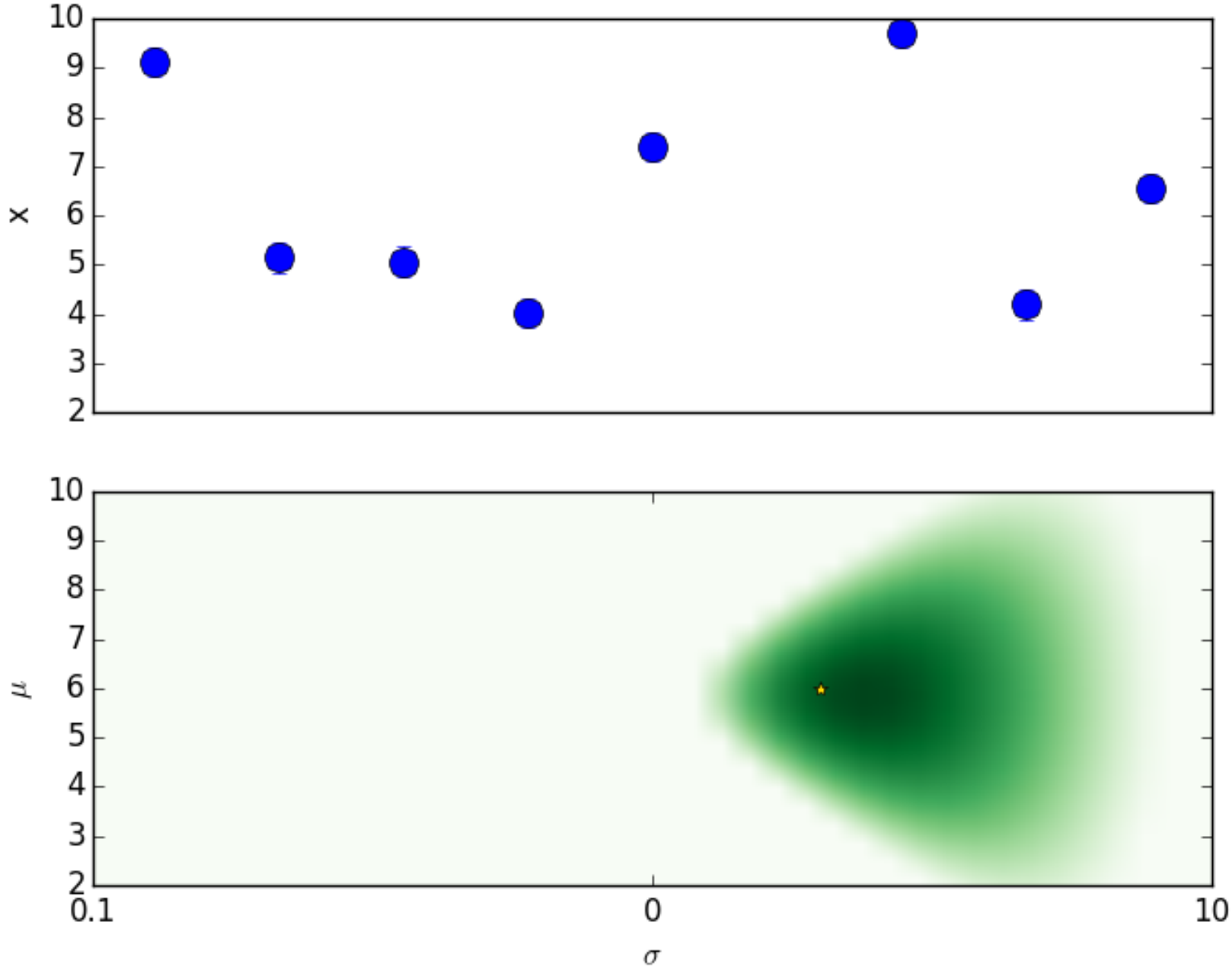

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

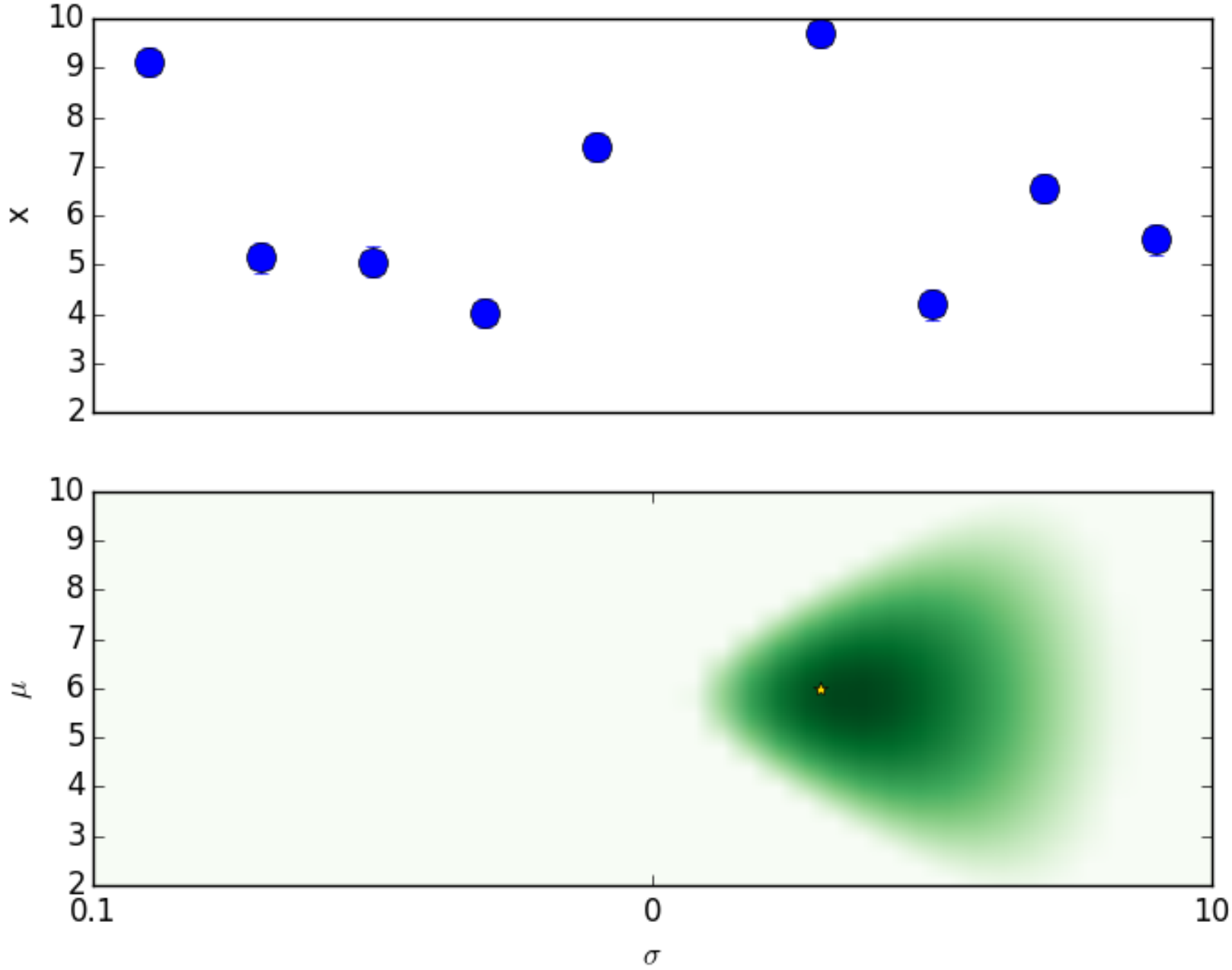

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

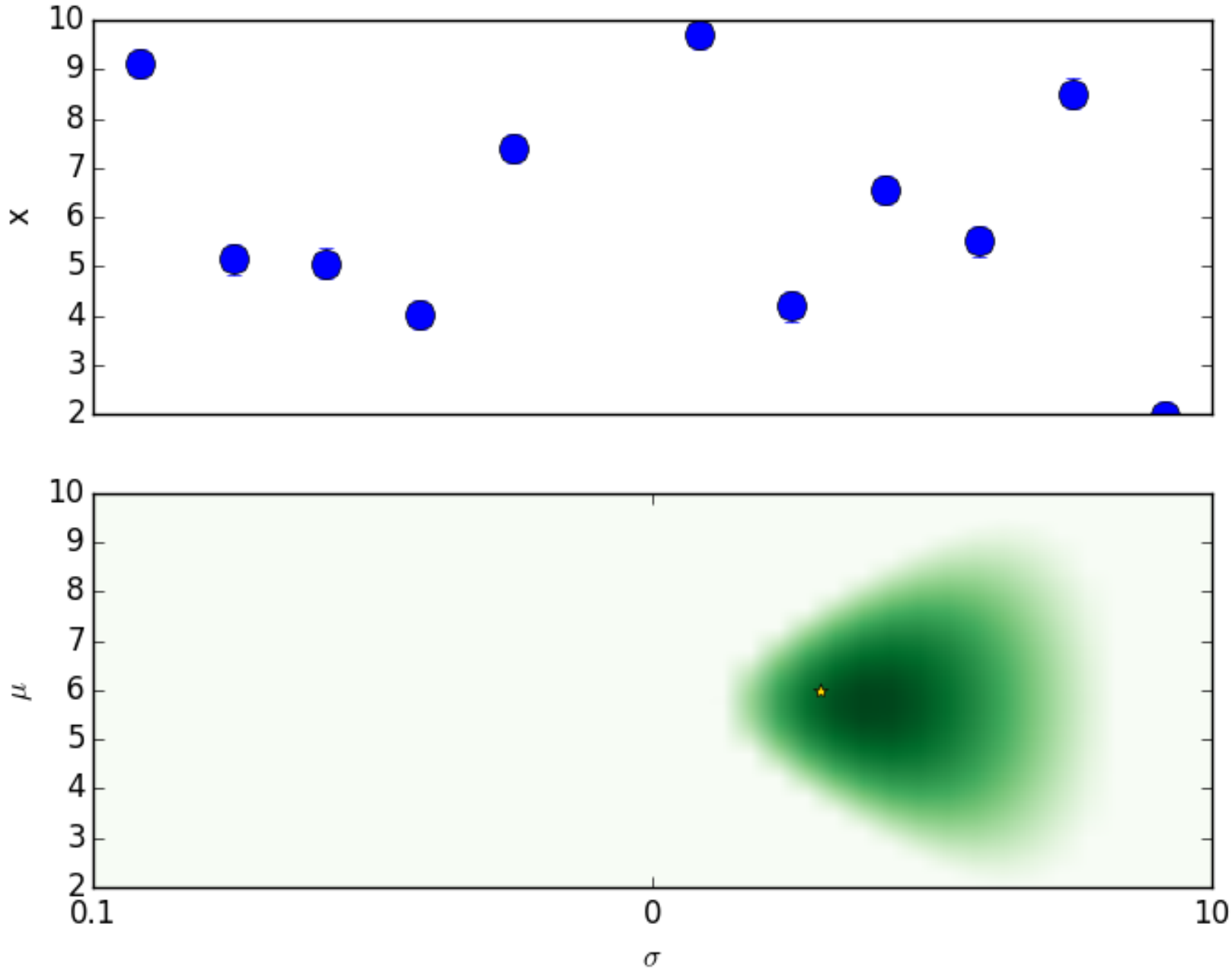

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

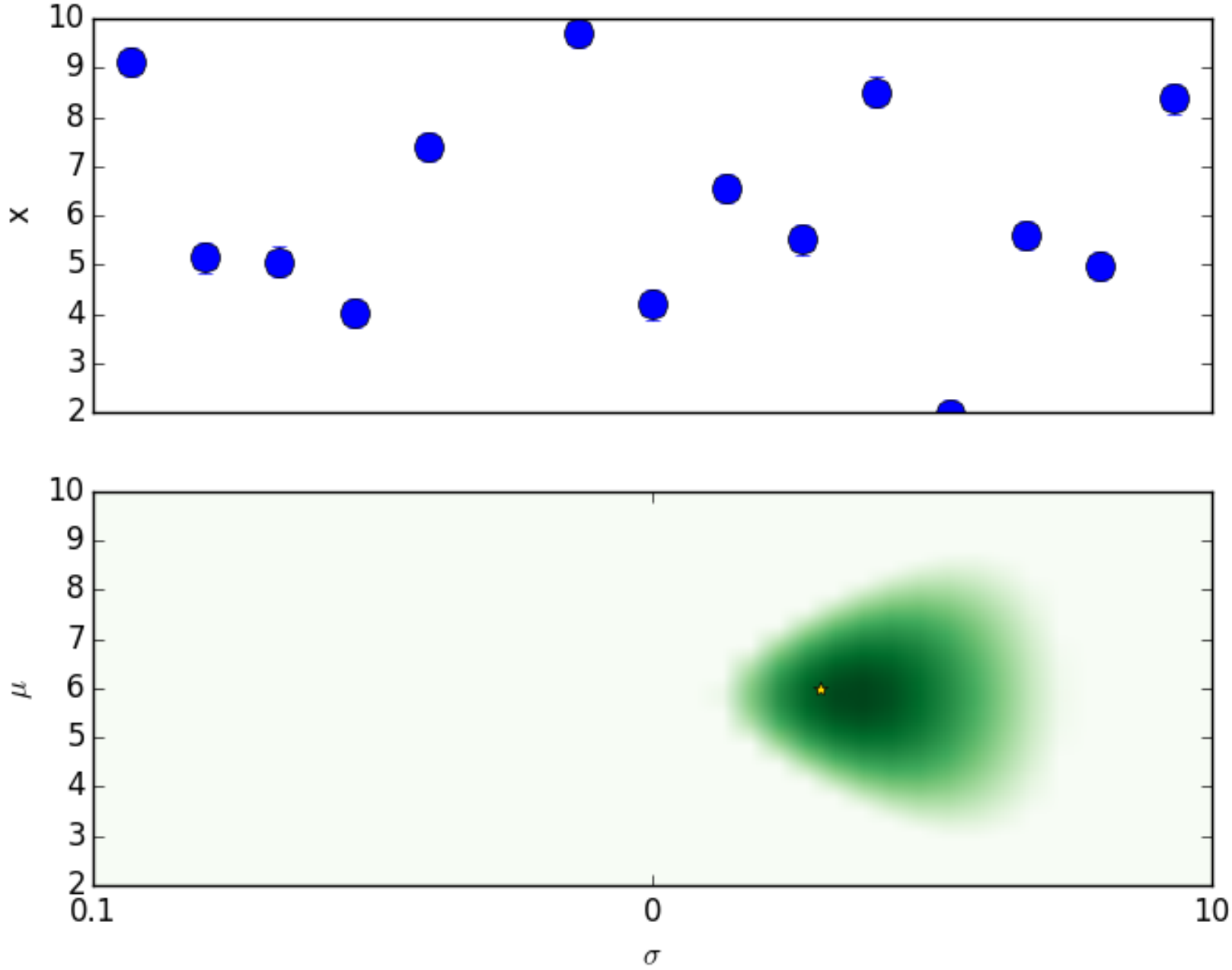

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

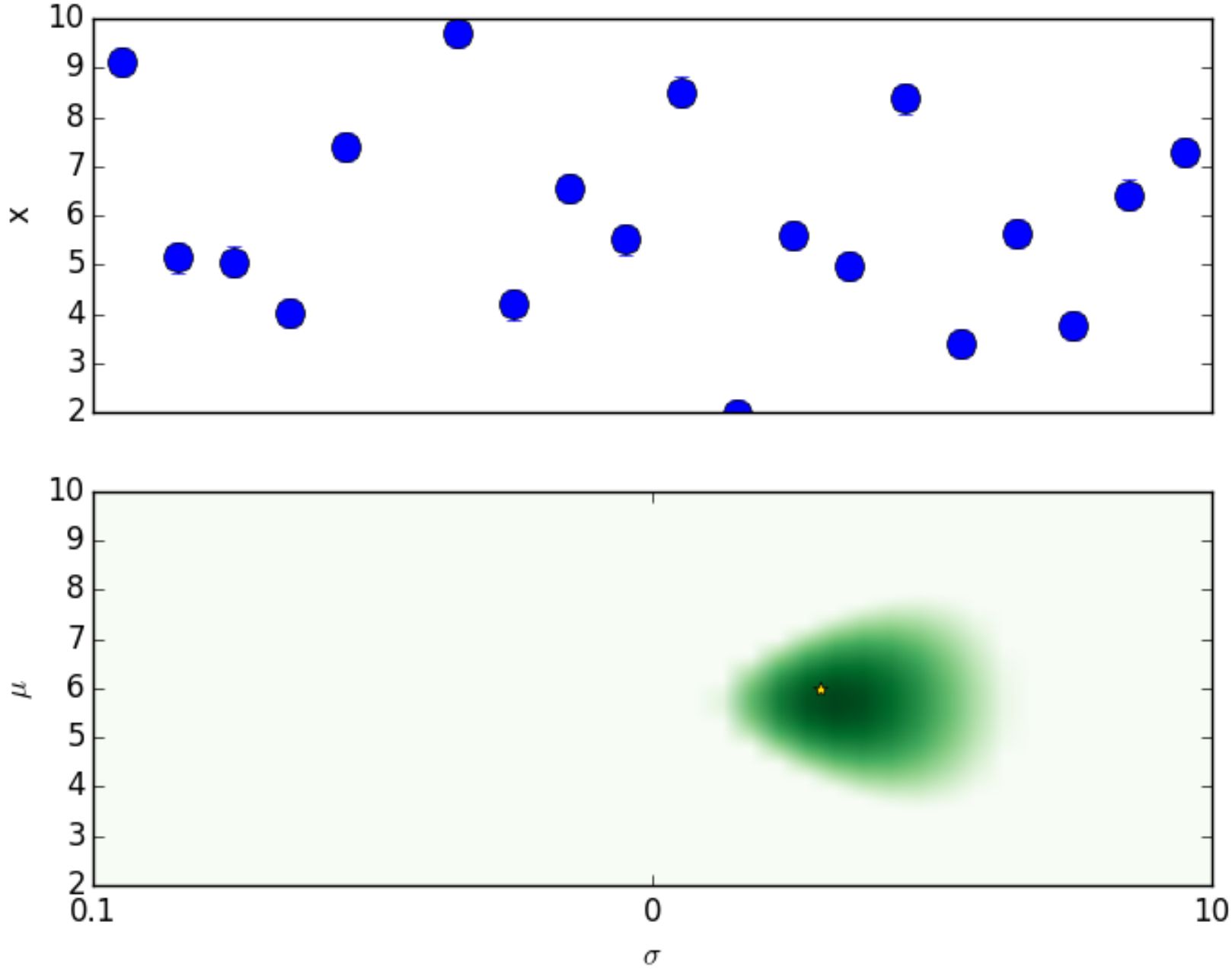

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

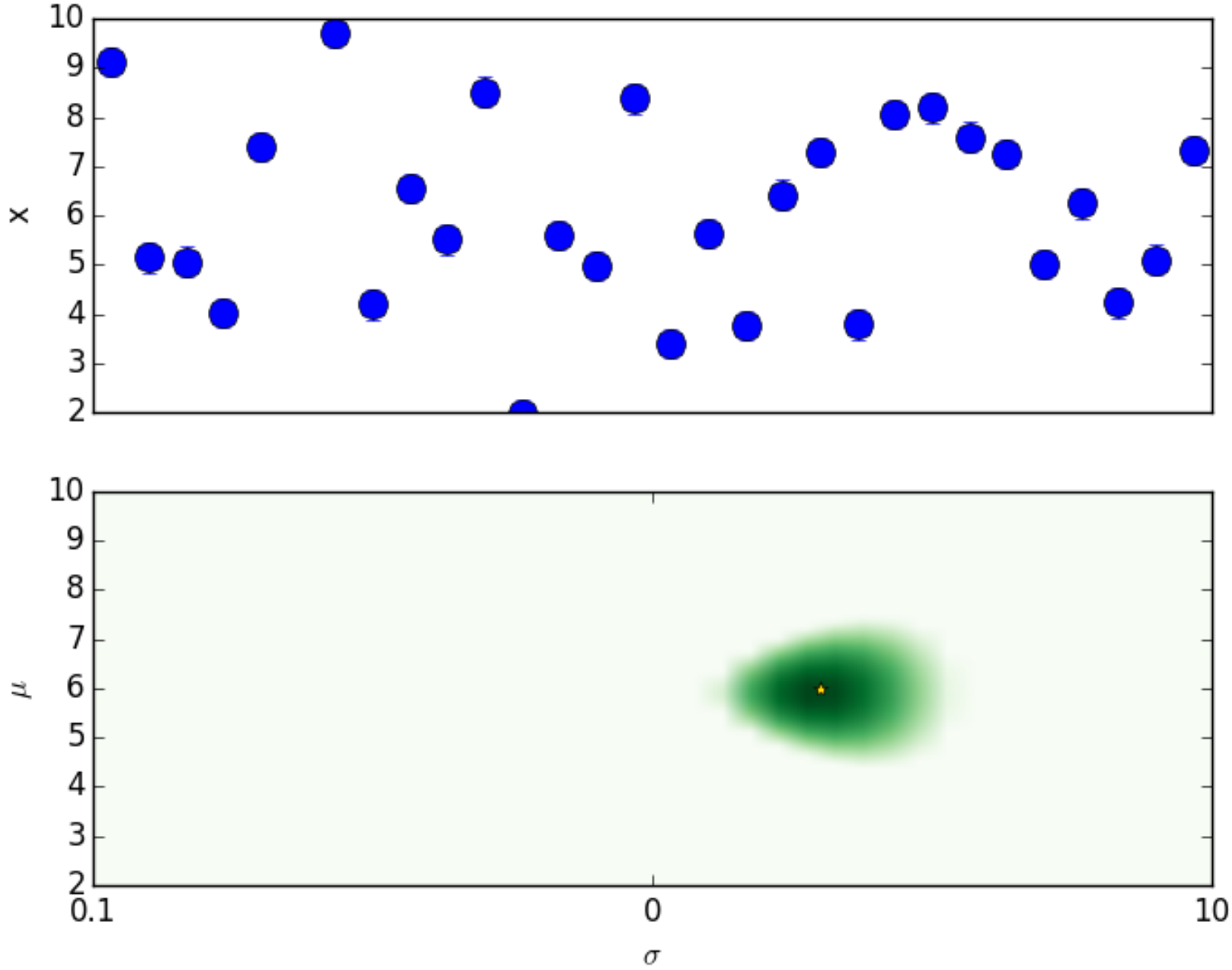

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

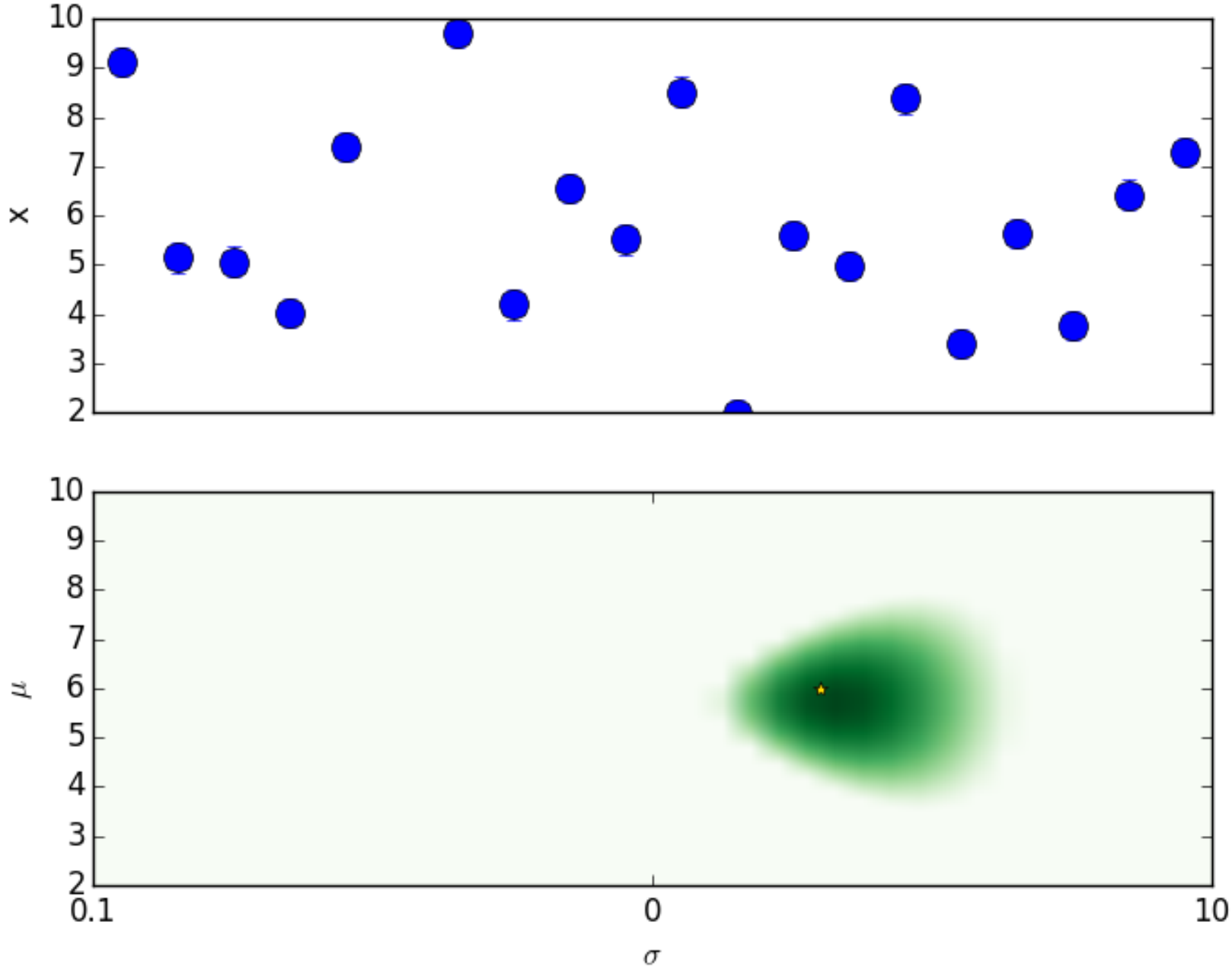

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

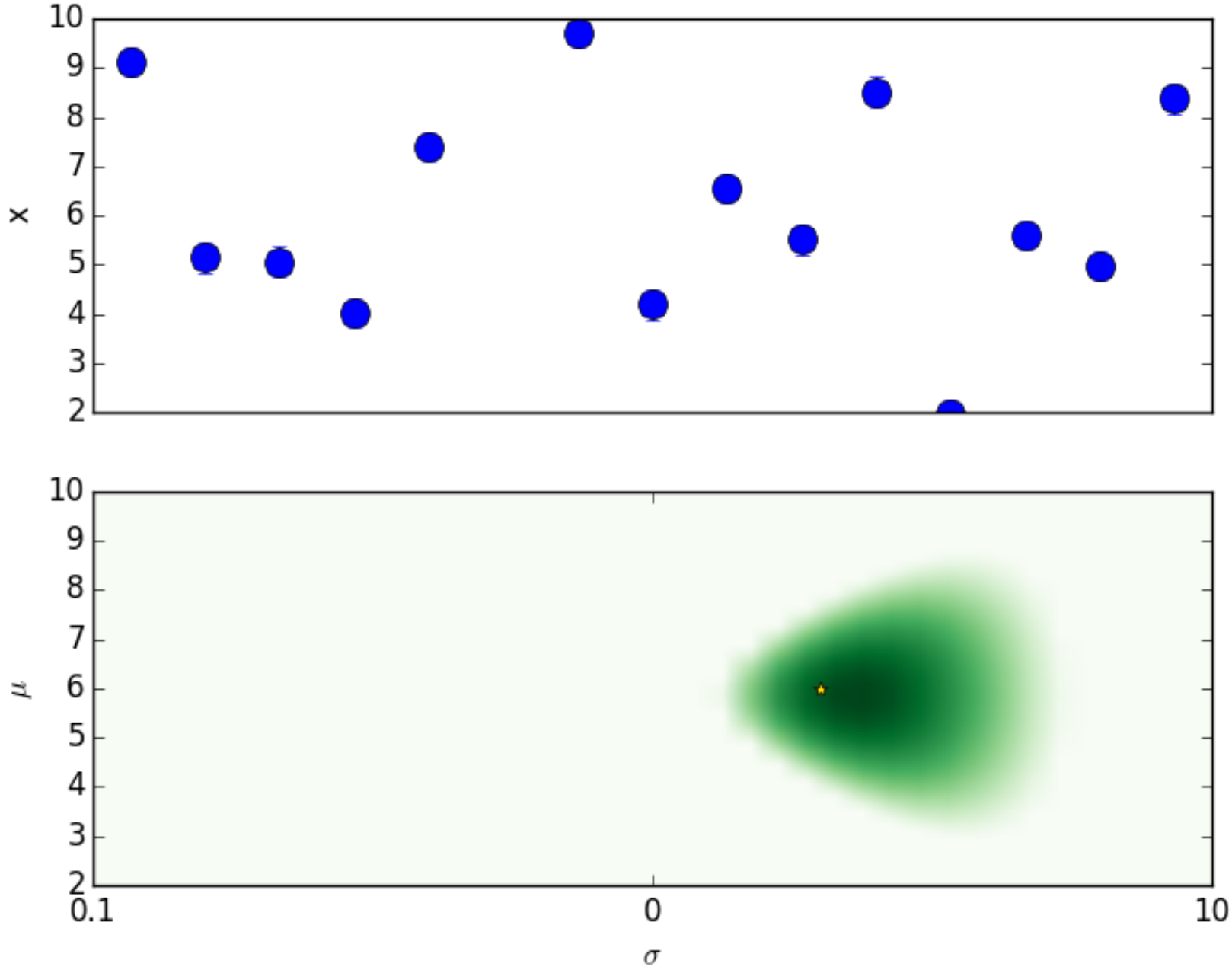

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

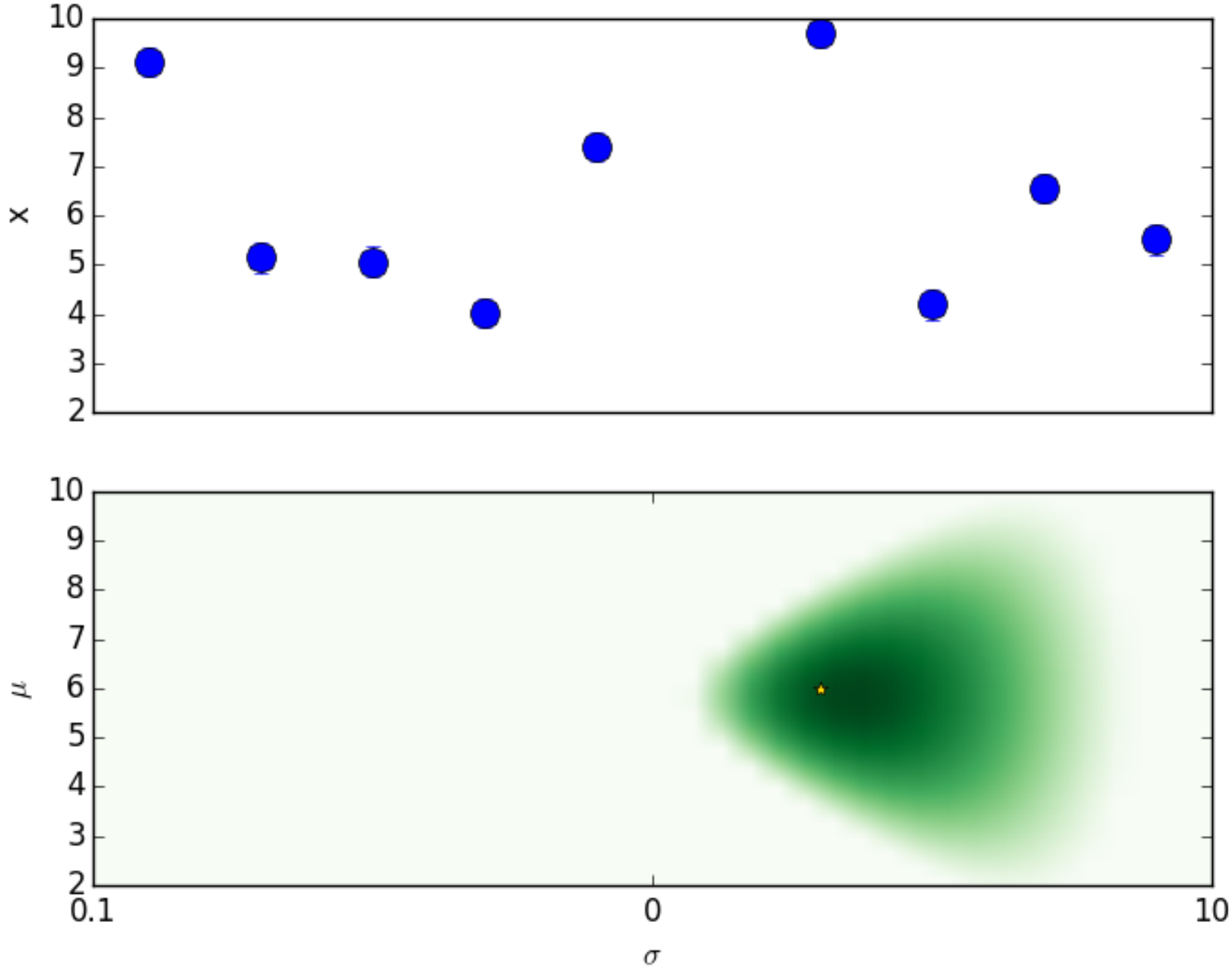

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

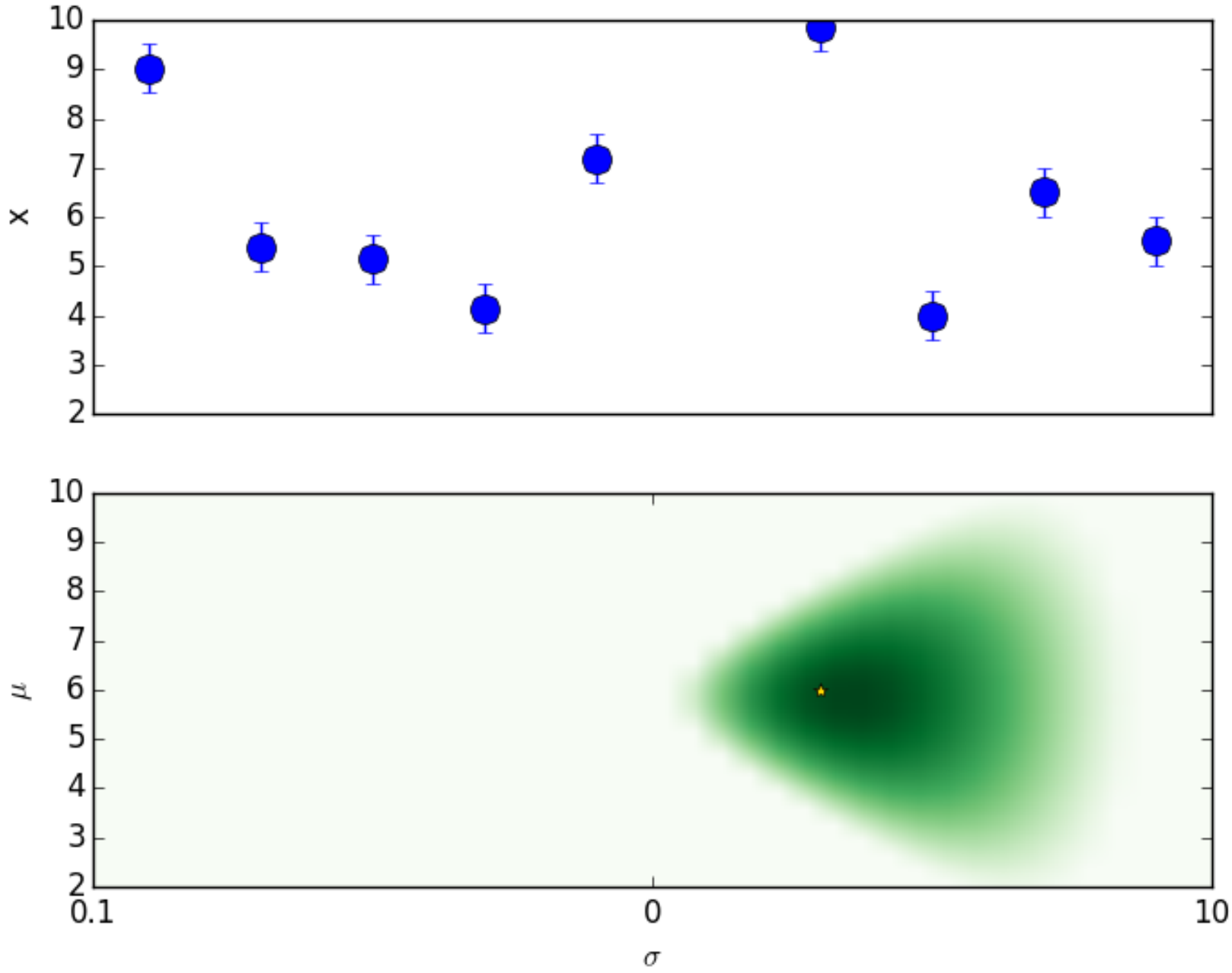

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

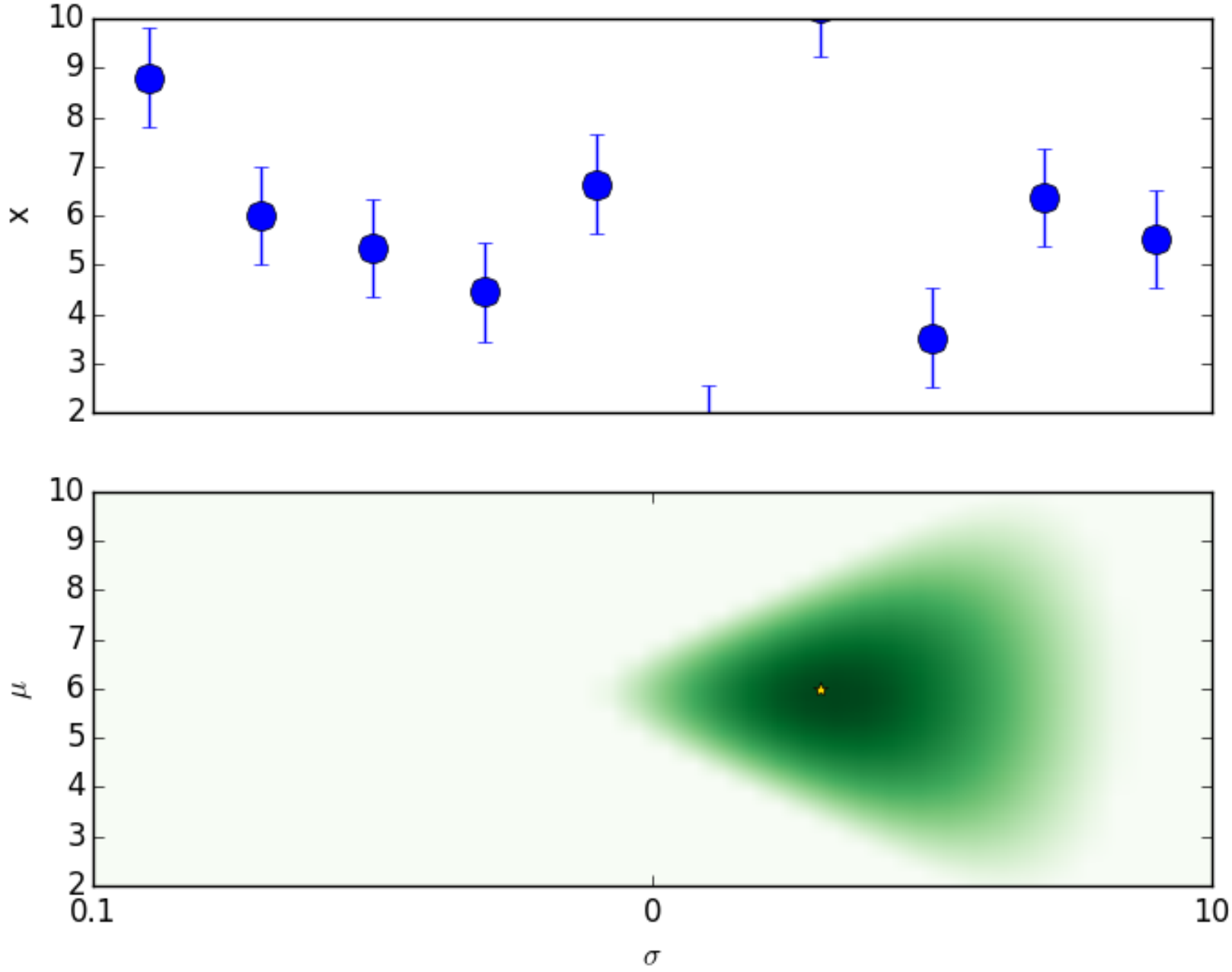

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

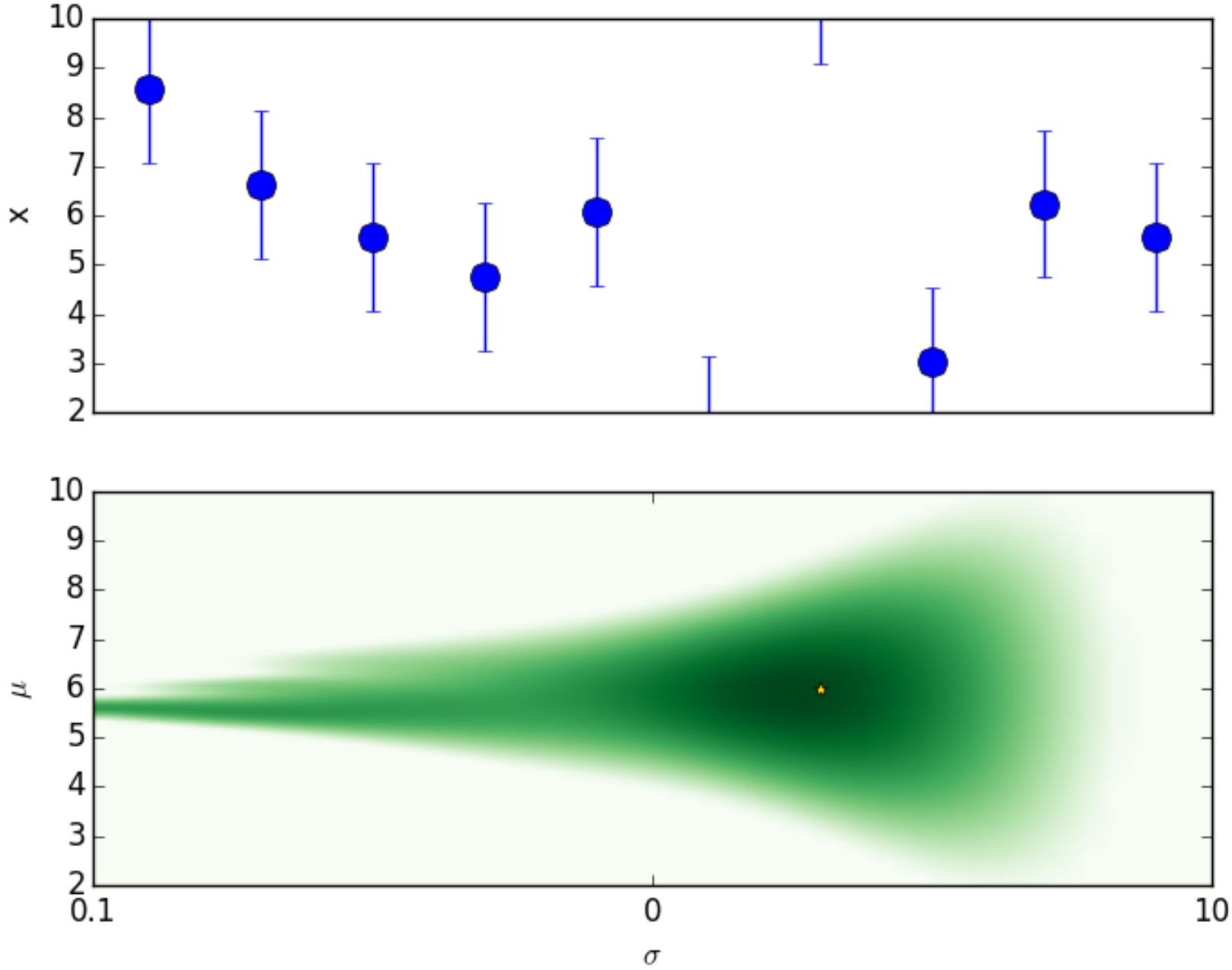

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

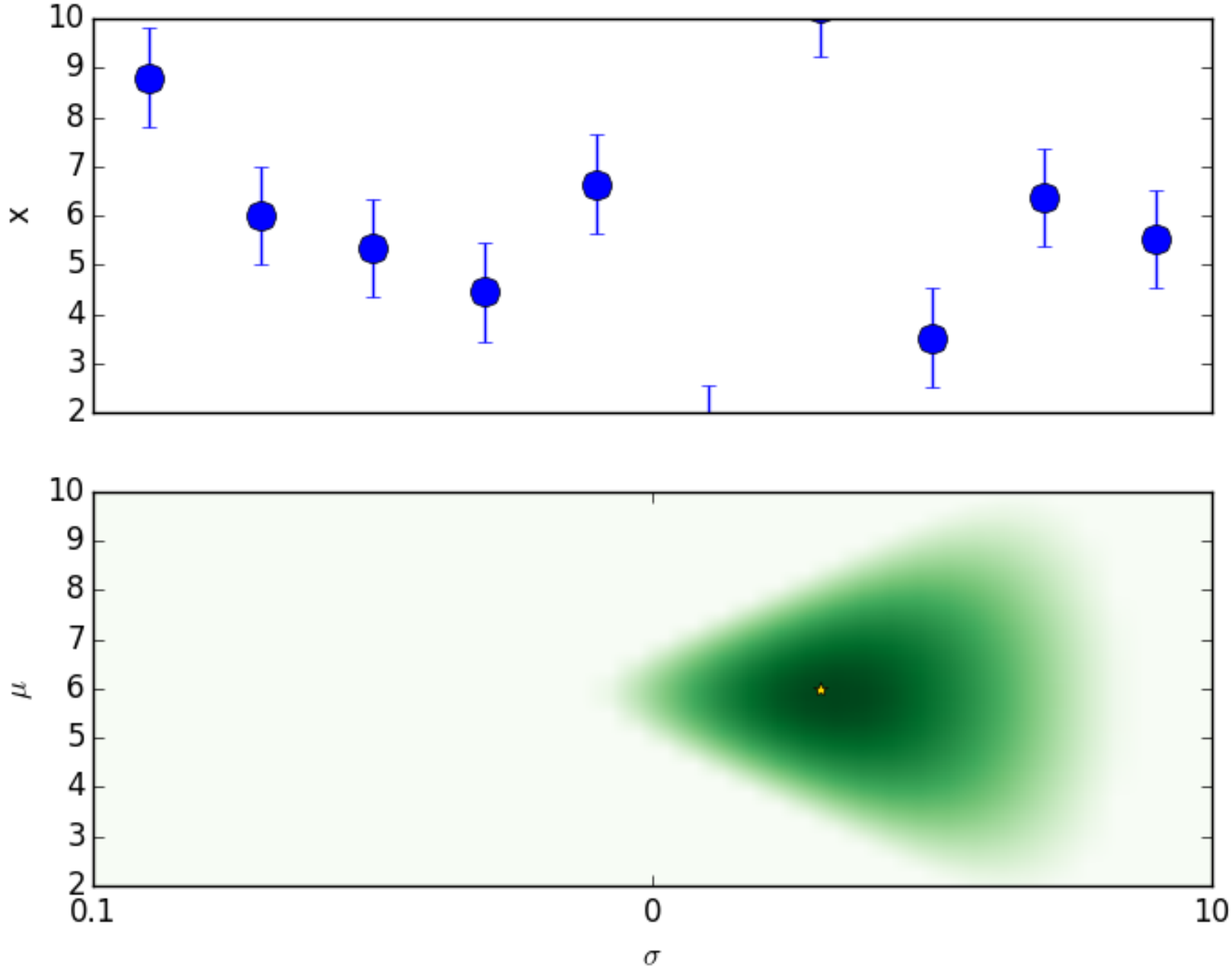

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

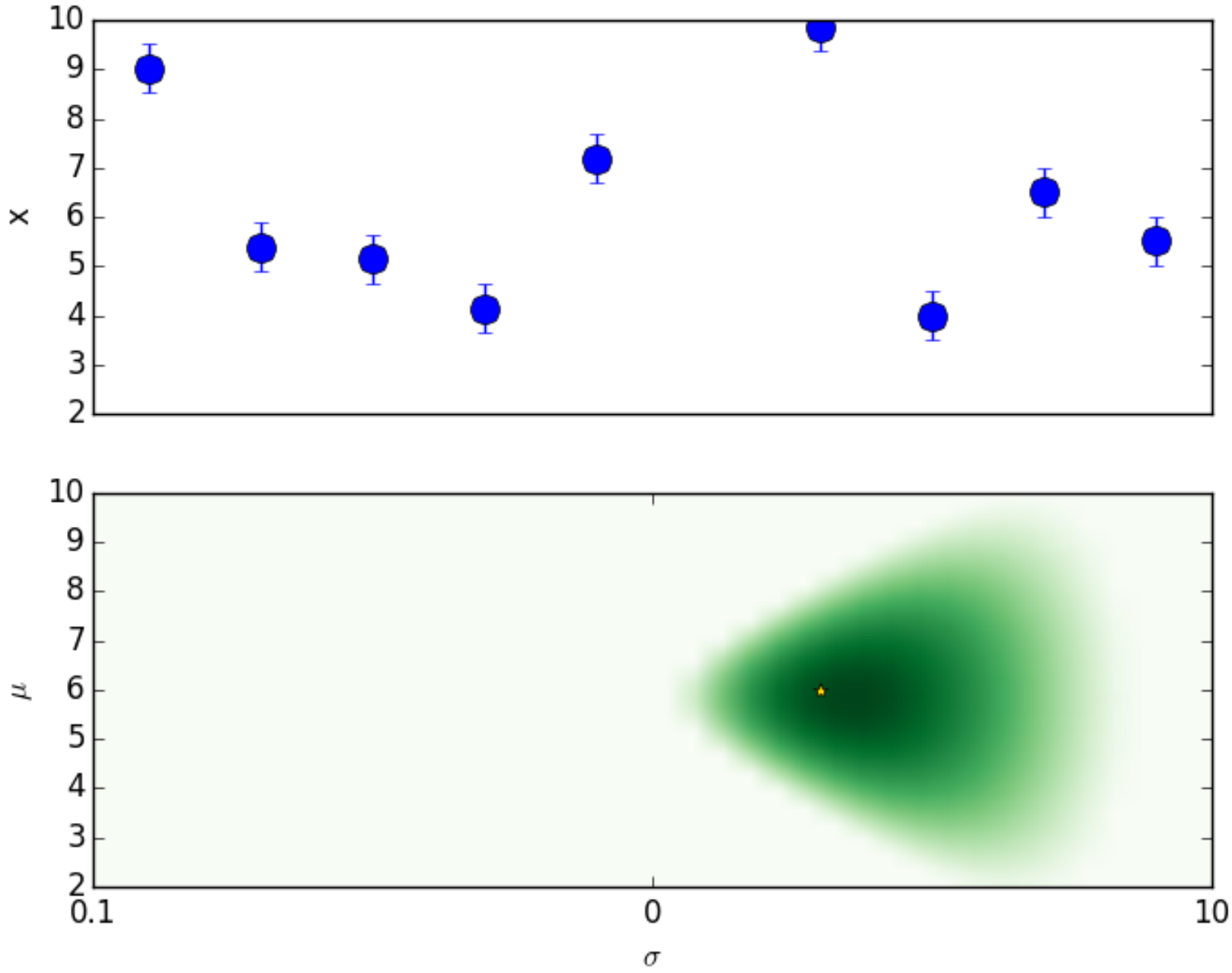

10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ


10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ


10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ


10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ


10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ


10
9
8
7
6
5
4
3
2
x
10
9
8
7
6
5
4
3
2
μ
0.1
0
10
σ

Practical pointers

Practical advice

- You can do this in any package!
- State what you are doing
- CStat (Poisson)
- Background with functions (check fit)
- Visualise, visualise, visualise
- Show posterior distributions & fits in data space
- Vary priors & assumptions
- Use nested sampling, MCMC with care
- Make simulations
- Ask for help

isis, sherpa, spex, xspec, 3ml, ...

Contact points for questions

- Ask a colleague
- Astrostatistics Facebook group
- XSPEC Facebook group
- @MPE
  - J Michael Burgess
  - Johannes Buchner