Strong Lens Modeling (II): Statistical Methods

Chuck Keeton

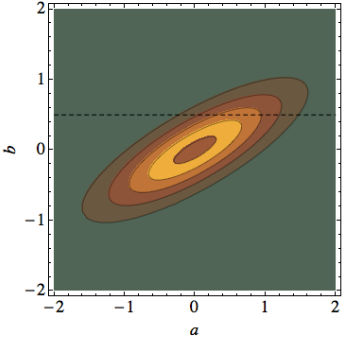
Rutgers, the State University of New Jersey

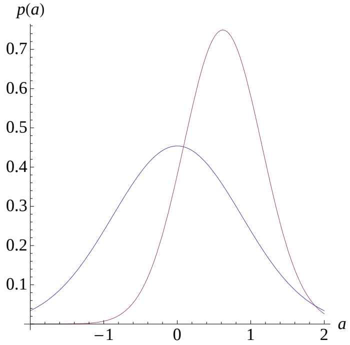
Probability theory

multiple random variables, a and b

joint distribution	$p(a, b)$
conditional distribution	$p(a\|b)$
marginal distribution	$p(a) = \int p(a, b)\ db$

note: $p(a, b) = p(a|b)\, p(b)$





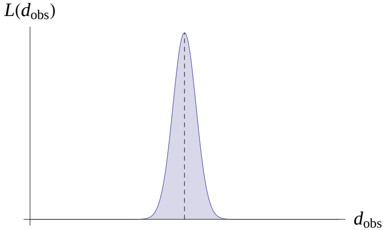
Likelihood

even if model is correct, measured data may not exactly match model predictions because of noise

1-d example: suppose model predicts d^{mod}, and measured value follows a Gaussian distribution with uncertainty σ

$$\mathcal{L}(d^{\mathrm{obs}}|d^{\mathrm{mod}}) \propto \exp\left[-\frac{(d^{\mathrm{obs}}-d^{\mathrm{mod}})^2}{2\sigma^2}\right] \propto e^{-\chi^2/2}$$

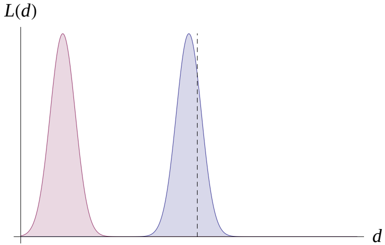
call this the "likelihood of the data given the model"



if model predictions depend on set of parameters $\mathbf{q}$, write this as

$$\mathcal{L}(d^{\rm obs}|\mathbf{q}) \propto \exp\left[-\frac{(d^{\rm obs} - d^{\rm mod}(\mathbf{q}))^2}{2\sigma^2}\right]$$

how to use? when model is wrong, $d^{\rm obs}$ is far from $d^{\rm mod}$ so χ^2 is high and $\mathcal{L}$ is low; adjust model to reduce χ^2 and increase $\mathcal{L}$ → **maximum likelihood method**



Bayesian inference

goal: see what we can infer about the parameters from the data; so shift from $p(d|q)$ to $p(q|d)$

note:

$$p(d,q) \;=\; p(d|q)\,p(q) \;=\; p(q|d)\,p(d)$$

Bayes's theorem:

$$p(q|d) = \frac{p(d|q)\,p(q)}{p(d)}$$

$p(d|q)$ = likelihood $\rightarrow \mathcal{L}(d|q)$

$p(q)$ = "prior" probability distribution for q

$p(d)$ = "evidence" *(more later)*

$p(q|d)$ = "posterior" probability distribution for q given d

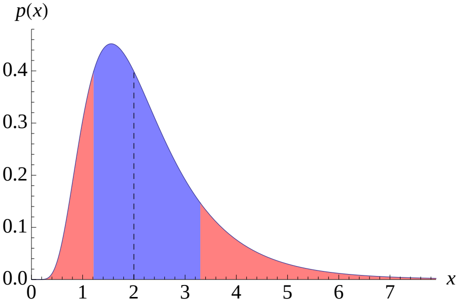
idea: use the "posterior" to quantify constraints on the parameters

Quantifying constraints

how do we use $p(q|d)$ to quantify parameter constraints?

could use μ and σ; but those have specific meaning only for Gaussian distributions

better to generalize: median and 68% confidence interval



Nuisance parameters

suppose we have some joint $p(a, b)$, but we are mainly interested in a

we say b is a “nuisance” parameter

probability theory lets us “integrate out” b to get **marginalized distribution** for a:

$$p(a) = \int p(a, b)\ db$$

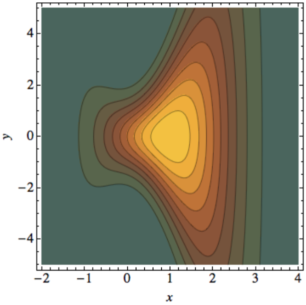
in general, this is not the same as *optimizing* the nuisance parameter example:

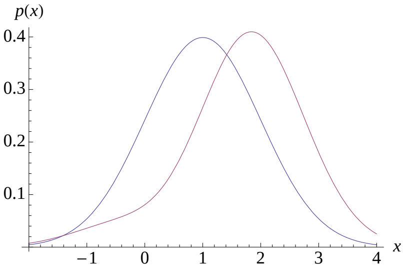
$$p(x,y) \propto \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \exp\left[-\frac{y^2}{2\sigma_y^2}\right]$$

with $\sigma_y = 1 + x^2$

$$\text{optimize:} \quad p(x) \quad \propto \quad \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right]$$

$$\text{marginalize:} \quad p(x) \quad \propto \quad (1+x^2)\exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right]$$



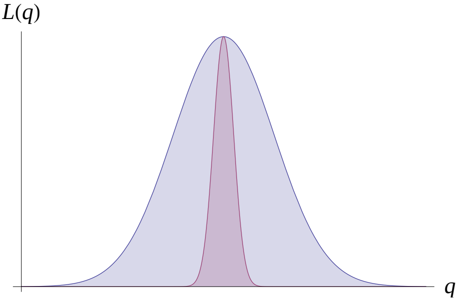


Model comparison

"evidence" quantifies overall probability of getting these data from this model:

$$Z \equiv p(d) = \int \mathcal{L}(d|q)\ p(q)\ dq$$

can be used to **compare different models** (even they have different numbers of parameters)



Monte Carlo Markov Chains

often it is inconvenient or even impossible to analyze full posterior

instead, turn to **statistical sampling**: set of points $\{\mathbf{q}_k\}$ drawn from the posterior

(for now: assume flat priors, so $p(\mathbf{q}) \propto \mathcal{L}(\mathbf{q})$)

MCMC method:

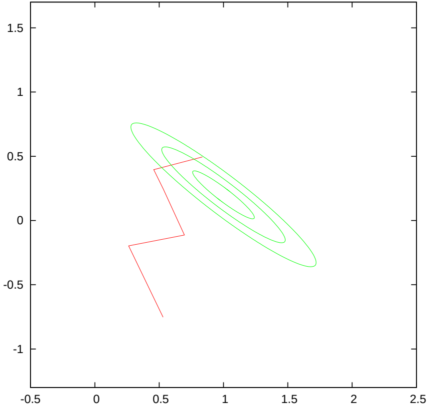
- pick some starting point $\mathbf{q}_1$
- postulate some *trial distribution*, $p_{\rm try}(\mathbf{q})$
- draw a *trial point*, $\mathbf{q}_{\rm try}$, from $p_{\rm try}$; probability to accept is

$$\min\left[\frac{\mathcal{L}(\mathbf{q}_{\rm try})}{\mathcal{L}(\mathbf{q}_1)}, 1\right]$$

- if accept trial point, put $\mathbf{q}_2 = \mathbf{q}_{\rm try}$; otherwise, put $\mathbf{q}_2 = \mathbf{q}_1$.
- iterate!

Simple steps

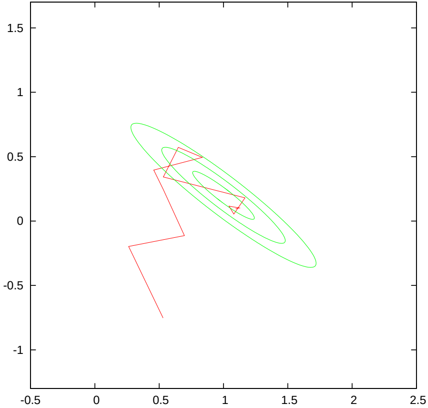
Let trial distribution be a simple Gaussian.



MCMC step 20.

Simple steps

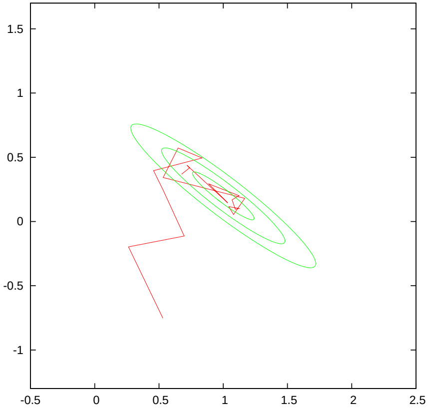
Let trial distribution be a simple Gaussian.



MCMC step 50.

Simple steps

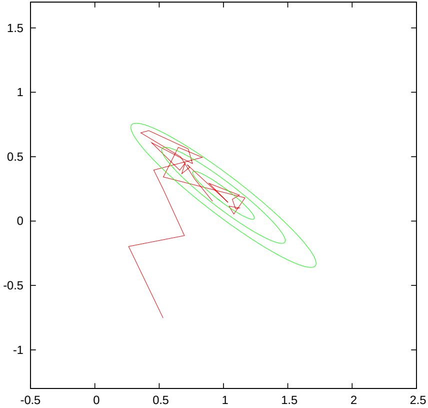
Let trial distribution be a simple Gaussian.



MCMC step 100.

Simple steps

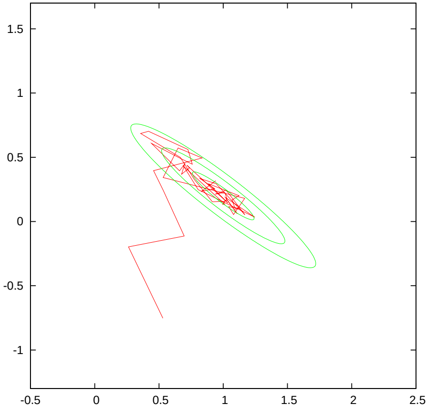
Let trial distribution be a simple Gaussian.



MCMC step 150.

Simple steps

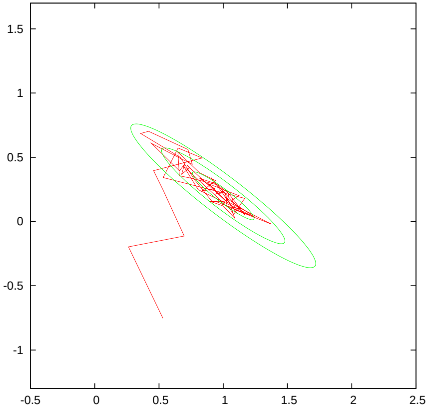
Let trial distribution be a simple Gaussian.



MCMC step 200.

Simple steps

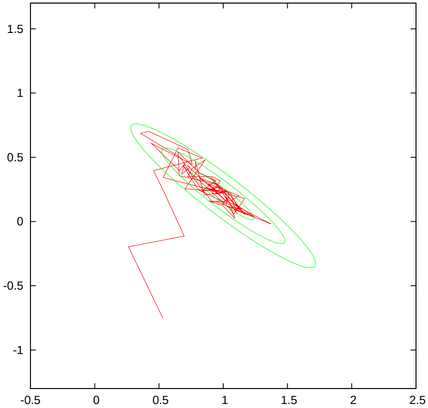
Let trial distribution be a simple Gaussian.



MCMC step 250.

Simple steps

Let trial distribution be a simple Gaussian.



MCMC step 300.

When to stop?

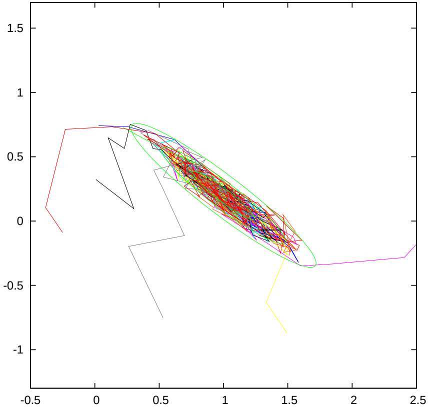
want:

- to sample $\mathcal{L}$ well
- to get results that are independent of starting point

solution:

- run multiple chains
- keep going until statistical properties of chains are equivalent
- throw away first half of each chain to eliminate “memory” of starting point

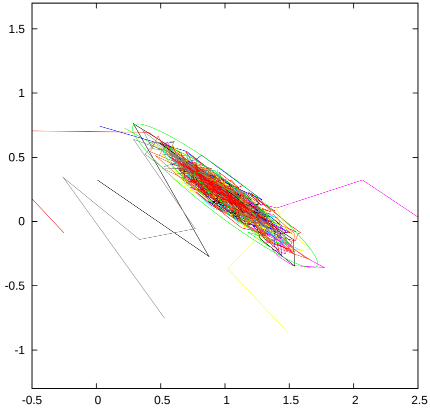
Multiple chains



MCMC, 10 chains, simple Gaussian steps.

Adaptive steps

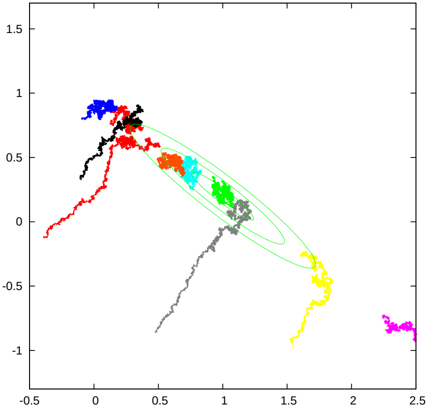
Q) Can we pick trial distribution to make MCMC more efficient?
A) Use covariance matrix of points so far.



MCMC with adaptive steps.

Step size

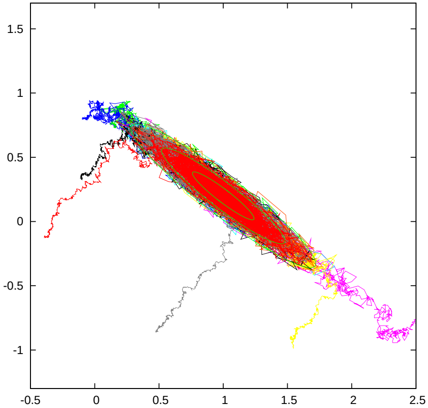
How big to make the steps?



MCMC with tiny steps.

Step size

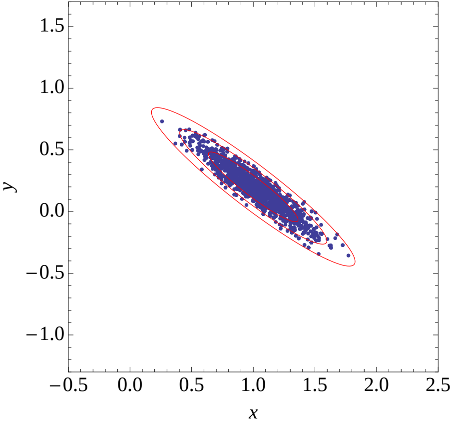
How big to make the steps?



MCMC with adjustable step size.

MCMC results

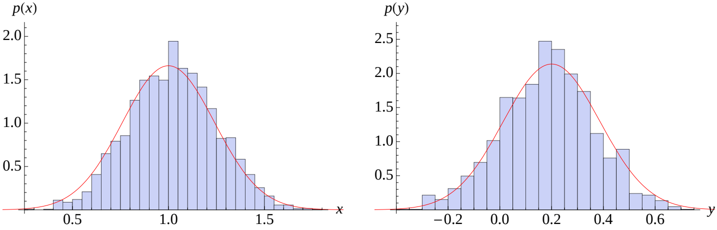
joint posterior, $p(x, y)$: just plot all the sampled points



MCMC results

marginalized posterior, $p(x)$: just plot a histogram of the x-values of all the sampled points

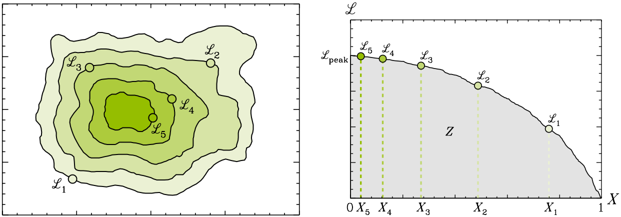
likewise for $p(y)$



Nested sampling

introduced by Skilling (2004, 2006)

- peel away layers of constant likelihood one by one
- estimate "volume" of each layer statistically
- combine $(\mathcal{L}_i, \mathcal{V}_i)$ values to estimate the Bayesian evidence
- get a sample of points as a by-product



(courtesy R. Fadely)

variants: Shaw et al. (2007), Feroz & Hobson (2008), Brewer et al. (2009), Betancourt (2010); statistical uncertainties: CRK (2011) given likelihood $\mathcal{L}(q)$ and prior $\pi(q)$, write evidence as

$$Z = \int \mathcal{L}(q)\ \pi(q)\ dq$$

define fractional volume with likelihood higher than L

$$X(L) = \int_{\mathcal{L}(q)>L} \pi(q)\ dq$$

in principle, can invert to find $L(X)$, then write

$$Z = \int_0^1 L(X)\ dX$$

discretize: if we can find a set of points (L_i, X_i) then we can write

$$Z = \sum_{i=1}^{N_{\text{nest}}} L_i (X_{i-1} - X_i)$$

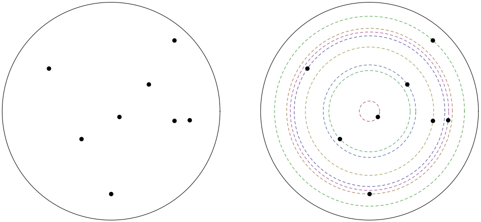
how to get the points?

L_i is "easy"

- draw uniformly (from prior) in region with $\mathcal{L} > L$

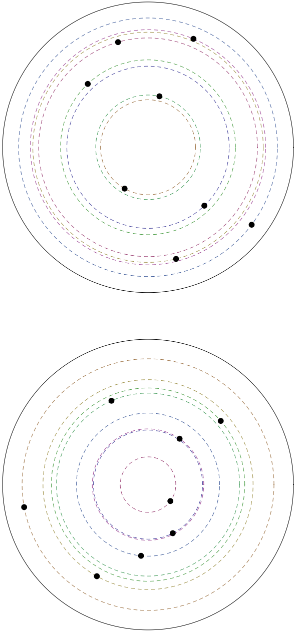
X_i is harder

- in principle, requires integration
- proceed statistically. . .



- consider M points drawn uniformly from region with $\mathcal{L} > L$
- draw likelihood contours through them, let enclosed volumes be $V_1 > V_2 > \ldots > V_M$ — these are *random variables*
- write $V_1 = V_0 t_1$ where $t_1 \in (0, 1)$
- then t_1 is the largest of M random variables drawn uniformly between 0 and 1 — characterized by probability distribution

$$p(t) = M t^{M-1} \quad \Rightarrow \quad \langle t_1 \rangle = \frac{M}{(M+1)}$$



Algorithm

begin with M “live” points drawn uniformly from full prior; let their likelihoods be $\mathcal{L}_\mu$ $(\mu = 1, \ldots, M)$

at step k:

- extract live point with lowest $\mathcal{L}$, call it k-th sampled point:

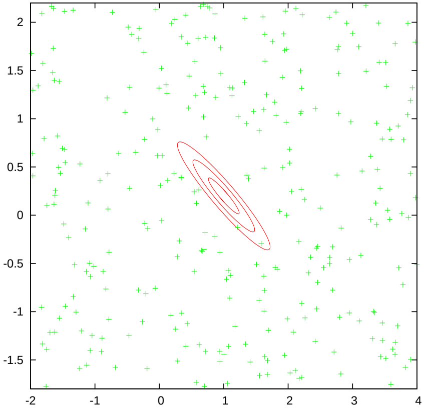
$$L_k = \min_\mu (\mathcal{L}_\mu)$$

- estimate the associated volume as

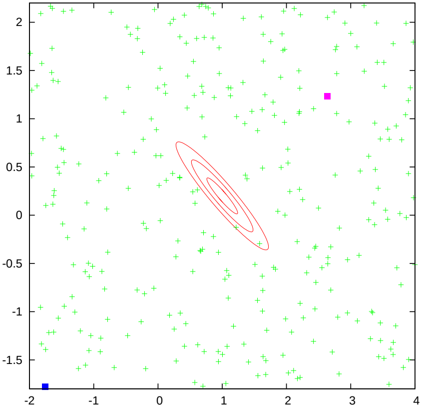
$$X_k = X_{k-1} t_k$$

  where t_k is a random number drawn from $p(t) = M t^{M-1}$
- replace extracted live point with a new point drawn from the priors but restricted to the region $\mathcal{L}(\mathbf{q}) \geq L_k$

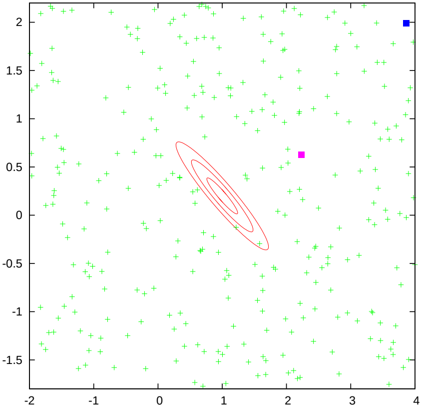
iterate for N_{nest} steps



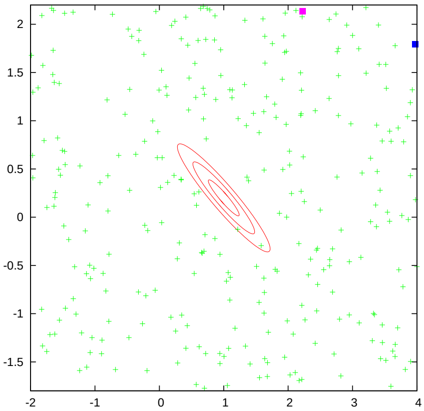
Nested sampling setup.



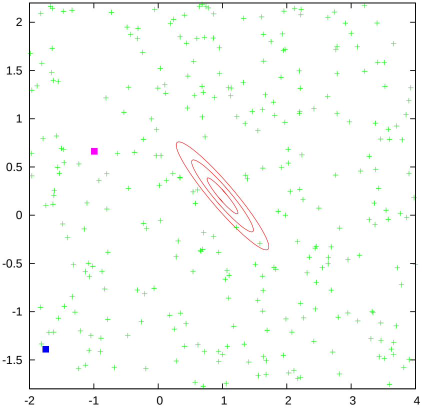
Nested sampling step 1.



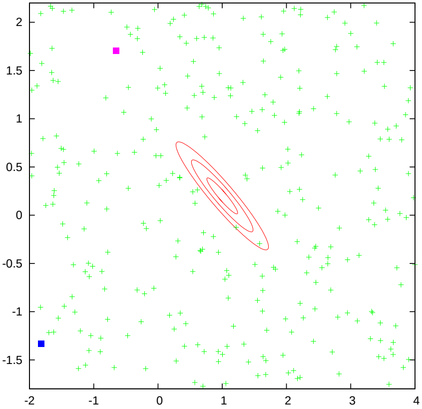
Nested sampling step 2.



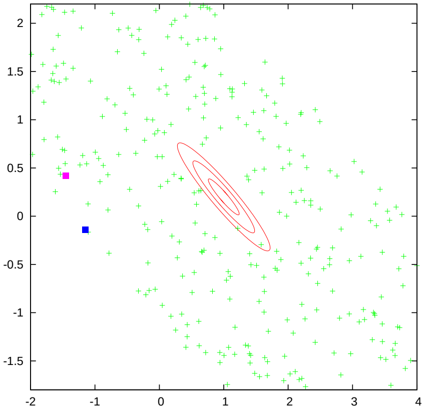
Nested sampling step 3.



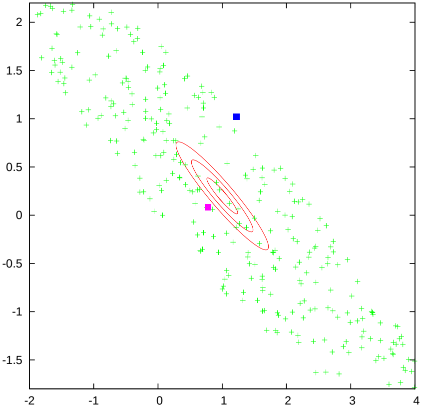
Nested sampling step 4.



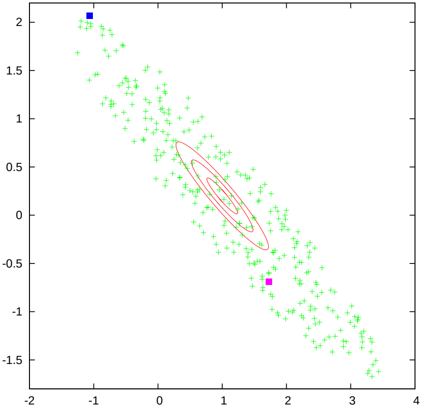
Nested sampling step 5.



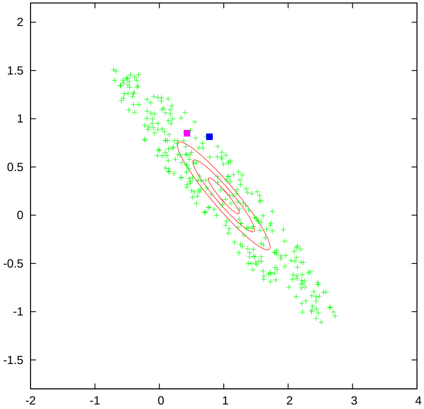
Nested sampling step 100.



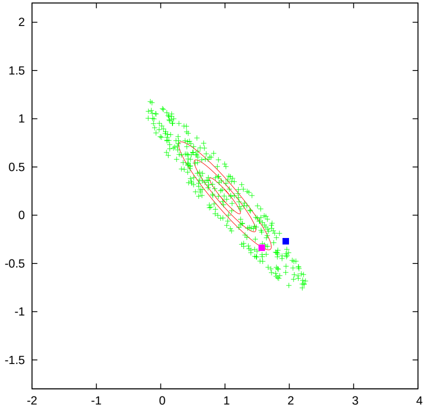
Nested sampling step 300.



Nested sampling step 500.

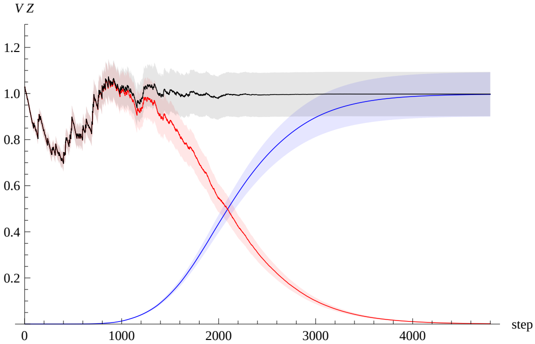


Nested sampling step 700.



Nested sampling step 900.

development of evidence: contributions from live points, sampled points, and **total**



(See CRK 2011 for statistical uncertainties.)