بسم الله الرحمن الرحیم

Data Modeling.

Main References:

① Selected Topic 10

② " " 11

③ " " 12_A

④ " " 13_A

⑤ 14031001

⑥ 14030216

⑦ IPM Workshop 2023

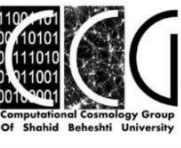
بسم الله الرحمن الرحیم


# Review on Data Analysis Methods & Cosmological Simulations

Seyed Mohammad Sadegh Movahed

Physics department, Shahid Beheshti University
Computational and Cosmological group (CCG-SBU)
www.smovahed.ir

Workshop on Computational Cosmology:
From Theory to Observation
1-2 August 2023







## Most relevant References: Books

1) Modern Cosmology, S. Donelson.
2) The Cosmic Microwave Background, R. Burrer
3) Physical foundation of Cosmology, V. Mukhanov
4) Cosmology, S. Weinberg
5) Neutrino Cosmology, J. Iesgourgues et.al
6) Galaxy Formation and Evolution, H. Mo, F.V. den Bosch and S. White
7) Statistics of Galaxy Distribution, V. Martinez, E, Saar
8) Cosmological Physics, J. A. Peacock
9) Topological Complexity of Smooth Random Functions, Adler, Robert, Taylor, Jonathan E., Springer, 2009.
10) Geometry, Topology and Physics, Mikio Nakahara, 1990.
11) Analysis and Data-Based Reconstruction of Complex Nonlinear Dynamical Systems, Using the Methods of Stochastic Processes, M. Reza Rahimi Tabar, Springer, 2019
12) Zomorodian, Afra. "Topological data analysis." Advances in applied and computational topology 70 (2012): 1-39.
13) Edelsbrunner, Herbert, and John Harer. Computational topology: an introduction. American Mathematical Soc., 2010.

## Most relevant References: Previous activities

Part A: Previous workshops
1) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/106-data-modeling-workshop
2) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/131-workshop-on-observational-data-analysis-in-cosmology-kurdistan-1398
3) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/144-school-and-workshop-on-statistical-analysis-of-cosmic-fields-1400
4) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/154-school-and-workshop-on-topological-based-data-analysis-1401

Part B: Some lectures:
1) http://facultymembers.sbu.ac.ir/movahed/index.php/courses/142-optimization-and-computational-approaches-fall-2021
2) http://facultymembers.sbu.ac.ir/movahed/index.php/courses/132-advanced-course-on-computational-physics
3) http://facultymembers.sbu.ac.ir/movahed/index.php/courses/139-stochasic-processes

Part C: Some of my talks:
1) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/102-my-talk-at-cosmology-meeting-ipm-96
2) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/148-my-talk-at-sharif-group-meeting-1401-2022
3) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/145-my-talk-at-conference-on-gravity-and-cosmology-1400
4) http://facultymembers.sbu.ac.ir/movahed/index.php/talks-a-presentations/130-data-science

## Most relevant References: Papers and Lectures

1) Matsubara, Takahiko. "Statistics of smoothed cosmic fields in perturbation theory. I. Formulation and useful formulae in second-order perturbation theory." The Astrophysical Journal 584.1 (2003).
2) Vafaei Sadr, A., and S. M. S. Movahed. "Clustering of local extrema in Planck CMB maps." Monthly Notices of the Royal Astronomical Society 503.1 (2021): 815-829.
3) Masoomy, H., et al. "Persistent homology of fractional Gaussian noise." Physical Review E 104.3 (2021): 034116.
4) Masoomy, H., S. Tajik, and S. M. S. Movahed. "Homology groups of embedded fractional Brownian motion." Physical Review E 106.6 (2022): 064115.
5) Lesgourges, J. "Cosmological perturbations." Searching for New Physics at Small and Large Scales: TASI 2012. 2013. 29-97.
6) Mostaghel, Behrang, Hossein Moshafi, and S. M. S. Movahed. "Non-minimal derivative coupling scalar field and bulk viscous dark energy." The European Physical Journal C 77 (2017): 1-22.
7) Pranav, Pratyush, et al. "Topology and geometry of Gaussian random fields I: on Betti numbers, Euler characteristic, and Minkowski functionals." Monthly Notices of the Royal Astronomical Society 485.3 (2019): 4167-4208.
8) Pranav, Pratyush. "Topology and geometry of Gaussian random fields II: on critical points, excursion sets, and persistent homology." arXiv preprint arXiv:2109.08721 (2021).
9) Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, Astro- phys. J., 304, 15
10) Bond, J. R., et al., The Astrophysical Journal 379 (1991): 440-460.

# Different parts of Data modeling

① Classification Data modeling
- Theory-Based approach ☆
- Data-Based approach

② Different types
- Frequentism approach
- Bayesian approach ☆

③ Mathematical Description and foundation ☆

④ Analytical and computation Algorithms for Data modeling
☆ MCMC & HMC, LHMC
- Canonic
- Microcanonic

⑤ Models Selection and Comparision ← Posterior ☆
Evidence

⑥ Goodness of fit & Confidence Interval on model's free ☆
Parameters.

⑦ Bayesian Model averaging BMA

⑧ Fisher Forecast or Fisher Information Matrix ← arXiv.0906.0943
Un-Biased Estimator
arXiv:2409.13583

⑨ Simulation Based Inference (SBI), Field-Level Inference.

Neural Likelihood Estimation, Emulators. (Data-Based Inference)

Likelihood free Inference

{D} observation or Experiments
Synthetic Data

## مراحل مدل کردن داده ها

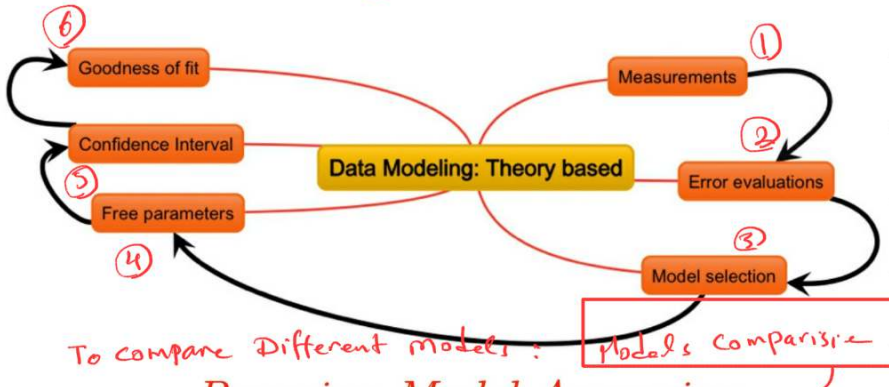
۱) اندازه گیری ( مستقیم یا شبیه سازی) ( Measurment )
۲) برآورد خطا و ارزیابی انتشار خطا بر روی کمیت های ثانویه (Error estimation and error propagation)
۳) تدوین مدل با توجه به تابع مناسب (model selection) (Merit function)
۴) انتخاب بینش تعیین مقادیر آزاد مدل (Bayesian or Frequentist)
۵) استفاده از روشهای تعیین مقادیر کمیت های آزاد مدل و البته حوزه اعتبار آنها
(Parameter estimation and confidence interval)
۶) تعیین خوبی مدل Goodness of fit

{θ}: پارامترهای آزاد مدل

① ☆ Theory_Based approach.

مدل خوش تعریف در نظر داریم.

## General view on Theory based
## a challenge: *a priori-that's it*




To compare Different models: Models comparisie {مقایسه مدل}

*Bayesian Model Averaging* (BMA)

Mostaghel, Behrang, Hossein Moshafi, and S. M. S. Movahed. "Non-minimal derivative coupling scalar field and bulk viscous dark energy." The European Physical Journal C 77 (2017): 1-22.

arXiv: 2403.02120
arXiv: 2310.06747
Lecture Note: 14030230

Ex:

$$U = \langle H \rangle = \frac{3K_BT}{2} + \frac{n}{2}\int d^3r\, u(r)\, g(r)$$

- $U$: Energy of system
- $u(r)$: Intractive potential.
- $g(r)$: Pair-Distribution

For Ideal Gas, we ignore this part

$$u(r) = 4\epsilon\left[\left(\frac{\alpha}{r}\right)^a - \left(\frac{\sigma}{r}\right)^b\right], \quad g(r) = e^{-\beta u(r)}$$

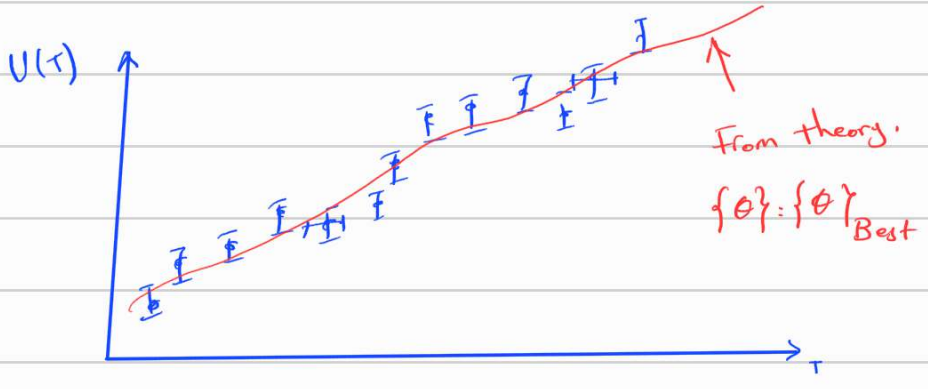
$$\{\theta\} = \{a, b, \epsilon, \alpha, \sigma, \beta\} \qquad \boxed{M = 6}$$

$\{\theta_i\}$, $i = 1 \cdots M$ → # of model's Free Parameter

$\{\theta\} = ?$

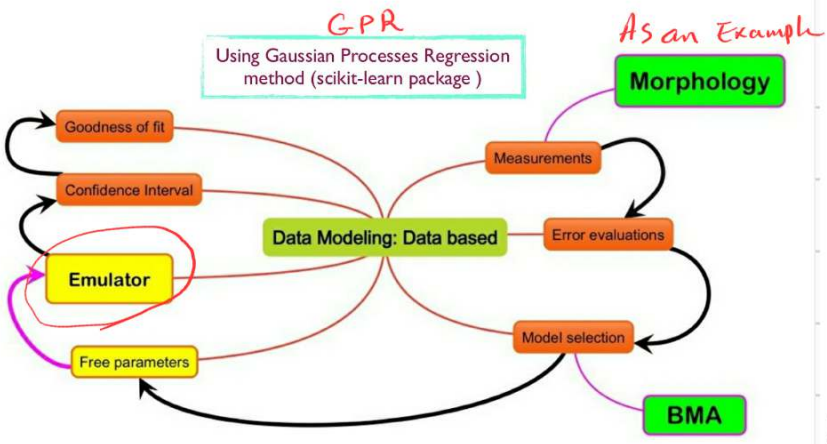
For experimental Setup we can Record

| T | U |
|---|---|
| $T_1$ | $U_1$ |
| $T_2$ | $U_2$ |
| $T_3$ | $U_3$ |
| ⋮ | ⋮ |



☆ Data-Based approach ⟶ Selected Topic II

## General view on Data based

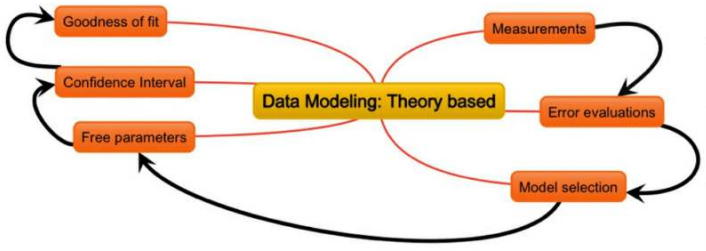




1) Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.
2) Heydenreich, Sven, Benjamin Brück, and Joachim Harnois-Déraps. "Persistent homology in cosmic shear: constraining parameters with topological data analysis." Astronomy & Astrophysics 648 (2021): A74.


## General view on Theory based

**a challenge:** *a priori-that's it*




*Bayesian Model Averaging*

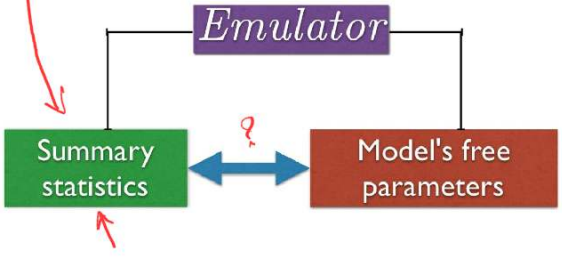
Mostaghel, Behrang, Hossein Moshafi, and S. M. S. Movahed. "Non-minimal derivative coupling scalar field and bulk viscous dark energy." The European Physical Journal C 77 (2017): 1-22.

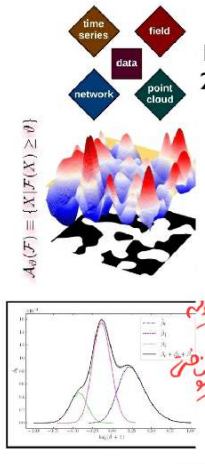
$U(T) = \ldots ?$

$\{\theta\} : \{M, S, \ldots\}$ : our interested parameters

## From Morphology to Cosmological Inferences (Part 1: Emulator)

1) Data preparation (Acquisition, reduction, generation)
2) Tracer selection (field, excursion sets, critical sets, morphological measure
3) Summary statistics



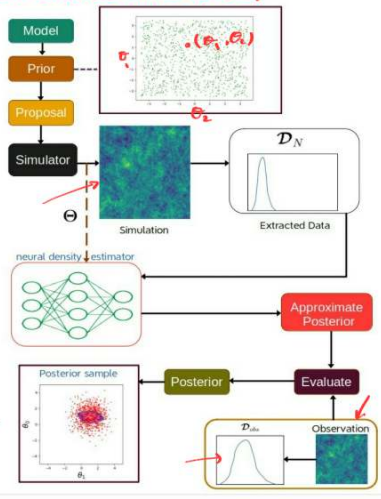

## From Morphology to Cosmological Inferences (Part 2: Simulation Based Inference)

1) Likelihood independent
2) No theoretical relation

Forward modeling

According to our First Principal and/or phenomenological theory Building

we are able to simulate our Cosmos




Credit: M.H. Jalali

Summary Statistics آمار و خلاصه

ابزاری هستند که با کمک آن ابزار دنیا را کمی می‌کنیم.

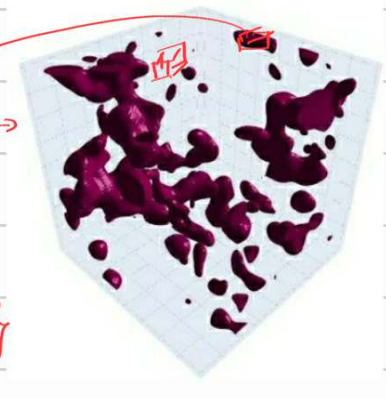
میدان جگالی ذرات اتم یا کهکشان

$\delta \equiv \frac{n(r) - \langle n \rangle}{\langle n \rangle}$

Density Contrast تباین چگالی

$\delta(\vec{r}) \xrightarrow{?} \{\text{size}, m, Q, T, P, \ldots\}$

$\{\theta\}$



(2) Bayesian approach with Mathematical Foundation

## A history about Bayesian strategy

- By Thomas Bayes (1702-1761)
- Pierre Simon Laplace (1812)
- Fisher, Neyman, Wald,...
- Gelfand and Smith (1990)
- Now

## تعیین مقدار برای کمیتهای آزاد

رهیافت Bayesian

۱) در این روش داده ها و نتایج بخشی از یک آنسامبل هستند

۲) کمیت های آزاد مدل مجهول هستند که ما تنها به صورت احتمالی می توانیم مقدار آنها را تعیین کنیم

۳) هر گونه اطلاعات اولیه در این رهیافت قابل استفاده هستند

۴) سازوکاری برای رهایی از کمیتهای اضافی (Nuisance) دارد

رهیافت Frequentist

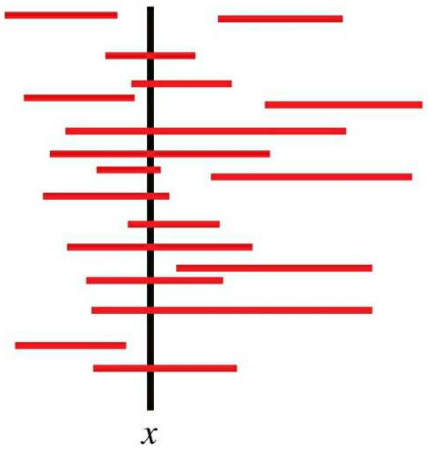
۱) در این روش داده ها و نتایج ریزحالت یک پیکربندی به حساب می آیند و داده ها تکرار پذیرند

۲) کمیت های مدل مجهول ولی ثابت هستند

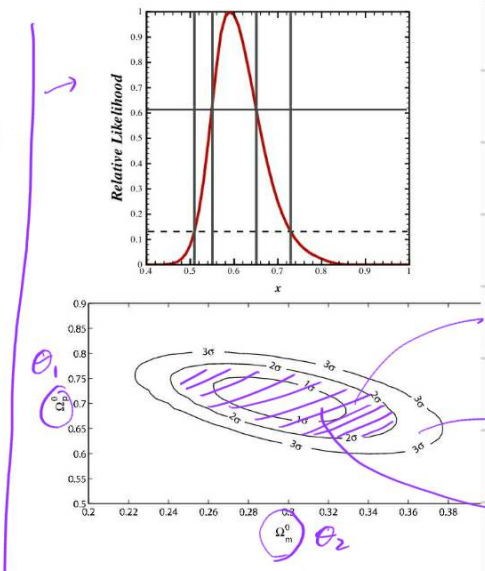
۳) هیچ گونه اطلاعی از مدل مورد استفاده نیست

۴) سازوکاری برای رهایی از کمیتهای اضافی (Nuisance) ندارد

5




بر اساس نگرش frequentist
برای مثال ۹۰ درصد
مشاهدات به مقدار ثابت x
منجر شده است




بر اساس نگرش Bayesian با
احتمال ۹۰ درصد مقدار دقیق
کمیت x دراین بازه قرار می گیرد

8

Mathematics:

Bayes theorem $P(x|y)P(y) = P(y|x)P(x)$

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

احتمال پیشین x بشرطی که y آمده باشد

احتمال پسین y بشرطی که x آمده باشد

Suppose that $\{D\} = \{D_1, D_2 - - \quad D_N\}$

$\{\theta\} = \{\theta_1, \theta_2 - \quad \theta_M\}$

نظریه ← مدل ← $Y(\{\theta\})$ Theo

$$P(\{\theta\}, M | \{D\}) = \frac{\mathcal{L}(\{D\}|\{\theta\}, M)\, P(\{\theta\}, M)}{\int d\{\theta\}\, \mathcal{L}(\{D\}|\{\theta\}, M)\, P(\{\theta\}, M)}$$

given a Model.

Posterior
احتمال پسین

☆ Probability of having $(\{\theta\}, M)$ given $\{D\}$

Likelihood Distribution توزیع درست نمایی

Probability of having $\{D\}$ given $(\{\theta\}, M)$

Prior Distribution توزیع پیشین

اطلاعاتی که در خصوص مدل و پارامترهای آن

$\{mass, Q\}$

If we have no Prior information Regarding $(\{\theta\})$,

Therefor we set $P(\{\theta\}, M) = cts$

$$P(\{\theta\}, M \mid \{D\}) \propto \mathcal{L}(\{D\} \mid \{\theta\}, M)$$

$$P(\{\theta\} = \{\theta\}_{Best}, M \mid \{D\}) \equiv \text{To be maximized}$$

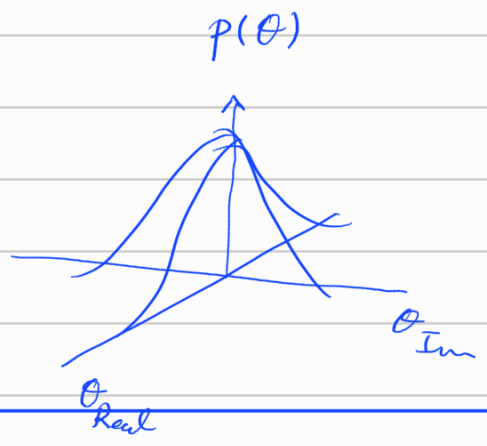
Prior = cts

$$\mathcal{L}(\{D\} \mid \{\theta\} = \{\theta\}_{Best}) \equiv \text{To be maximized}$$

$M_1$ , $\theta$ = imaginary variable

$P(\theta)$

$\theta = \theta_{Real} + i\theta_{I}$




Selected Topic 11

Mathematical Form of Posterior and Likelihood ?

$Y_{theory}(\{\theta\})$ ✓

$\mathcal{L}(\{D\}|\{\theta\}) = ?$

# بیشینه سازی تابع Posterior

$D:\{x_i,y_i\} \quad i=1,...,N$

$\Theta:\{\theta_j\} \qquad Y(x;\{\Theta\}) = \sum_{k=1}^{M}\theta_k f_k(x)$

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\int d\Theta\ p(D|\Theta)p(\Theta)}$$ ← Bayes theorem

در صورتی که تابع prior وجود نداشته باشد در نتیجه تحلیل posterior به تحلیل Likelihood تبدیل می شود

$if \quad p(\Theta) = cts$

$p(\Theta|D):\ p(D|\Theta)$

$$\mathcal{L}p(D|\Theta) \equiv \prod_{i=1}^{N}P_i = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(-\frac{\left[y_i - Y(x_i;\{\Theta\})\right]^2}{2\sigma_i^2}\right)$$

۱) با فرض اینکه اندازه گیری ها از یکدیگر مستقل باشند

۲) با توجه به قضیه حد مرکزی می توان در نظر گرفت که هر نقطه اندازه گیری حول مقدار بهینه اش به صورت گوسی توزیع شده است

$$\chi^2(\{\Theta\}) \equiv \sum_{i=1}^{N}\frac{\left[y_i - Y(x_i;\{\Theta\})\right]^2}{\sigma_i^2} = \sum_{i=1}^{N}\frac{\left[y_i - \sum_{k=1}^{M}\theta_k f_k(x_i)\right]^2}{\sigma_i^2}$$

$\mathcal{L}p(D|\Theta):\ e^{-\frac{\chi^2(\{\Theta\})}{2}}$

بیشینه شدن Likelihood معادل با کمینه شدن chi^2 است

12

$$\mathcal{L}(\{D\}|\{\theta\}_{1\to M}) = \frac{1}{\sqrt{(2\pi)^N \, Det\, Cov_D}}\ e^{-\frac{\Delta^T \cdot Cov_D^{-1}\cdot\Delta}{2}}$$

$$\chi^2 \equiv \Delta^T \cdot Cov_D^{-1}\cdot\Delta$$

Multi-Variate Gaus Func

For Diagonal Cov $\Rightarrow \sum_{i=1}^{N}\frac{\left[y_i - Y_{th}(\{\theta\},x_i)\right]^2}{\sigma_i^2}$

$\Delta^T = (y - Y_{th})^T_{1\times N}$

$\Delta = (y - Y_{th})_{N\times 1}$

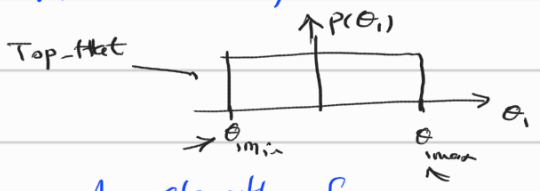
$Cov_D = \langle \Delta^T \Delta \rangle =$

بزرگترین خطای اندازه گیری

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & & \\ \vdots & & \ddots & & \\ \sigma_{N1} & & & & \sigma_N^2 \end{bmatrix}_{N\times N}$$

## ③ Computational Method to minimize $\chi^2$ or maximize Likelihood

Prior = cts




An algorithm for Deterministic scanning.

Ex 1: $\{\theta\} = \{\theta_1, \theta_2\}$ $M = 2$

Import $\{D\} = \{(x_i, y_i, \sigma_i)\}$ $i = 1, \ldots, N$

$\chi^2_{min} = 1e10$

loop $\theta_1 = \theta_{1min}, \theta_{1max}, \Delta\theta_1$

  loop $\theta_2 = \theta_{2min}, \theta_{2max}, \Delta\theta_2$

    $\chi^2 = 0$

    loop $i = 1, N$ (Computing $\chi^2$)

      $\chi^2 = \chi^2 + \dfrac{(y_i - Y(\theta_1, \theta_2, x_i))^2}{\sigma_i^2}$

    End loop

    If $\chi^2 < \chi^2_{min}$ then

      $\chi^2_{min} = \chi^2$

      $\theta_{1Best} = \theta_1$

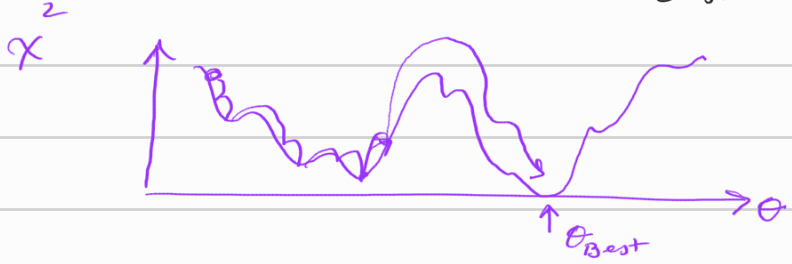
      $\theta_{2Best} = \theta_2$

    End if

  End loop

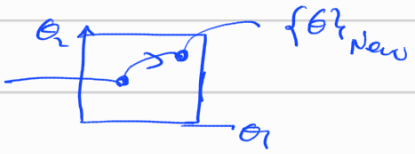
End loop

Write $\theta_{1Best}$, $\theta_{2Best}$, $\chi^2_{min}$



Ex 2

# MCMC Algorithm

$\{\theta_{old}\}$ 


Import Data $\{(x_i, y_i, \sigma_i)\}$ $i = 1, \dots N$

Select $\{\theta\}_{old}$ Randomly

$$\theta_1 = \theta_{1min} + R_1(\theta_{1max} - \theta_{1min})$$
$$\theta_2 = \theta_{2min} + R_2(\theta_{2max} - \theta_{2min})$$

loop $i = 1, N$

$$\chi^2_{old} = \chi^2_{old} + \frac{[y_i - Y_{th}(\{\theta\}_{old}, x_i)]^2}{\sigma_i^2}$$

End loop

$[M]$

loop on MCMC according to (Metropolis–Hastings Algorithm)

Select $\{\theta\}_{New}$ according to $\{\theta\}_{old}$ (Markov chain)

$$\{\theta\}_{New} = \{\theta\}_{old} + \{\Delta\theta\}$$

Random Number $\rightarrow R_1, R_2$ (Box–Muller algorithm)

$$\Delta\theta_1 = \epsilon_1 \sqrt{-2\ln R_1}\, \cos(2\pi R_2)$$

$$\Delta\theta_2 = \epsilon_2 \sqrt{-2\ln R_1}\, \sin(2\pi R_2)$$

$\chi^2$ for New location

$\chi^2 = 0$

loop $i = 1, N$

$$\chi^2 = \chi^2 + \frac{[y_i - Y_{the}(\{\theta\}_{New}, x_i)]^2}{\sigma_i^2}$$

End loop

$$\chi^2_{new} = \chi^2, \quad \Delta\chi^2 = \chi^2_{New} - \chi^2_{old}$$

$$AR \equiv \min\left\{1, \frac{\mathcal{L}_{New}}{\mathcal{L}_{old}}\right\} \quad e^{-\frac{(H_{New} - H_{old})}{K_B T}}$$

$$= \min\left\{1, \underbrace{e^{-\frac{\Delta\chi^2}{2}}}\right\}$$

Algorithm

$\xi \equiv$ call Random no.

Metropolis - Hastings method

If $\xi \leq e^{-\frac{\Delta\chi^2}{2}}$ then

$\chi^2_{old} = \chi^2_{New}$

$\{\theta\}_{old} = \{\theta\}_{New}$

End if

Write $\{\theta\}_{old}$, $e^{-\frac{(\chi^2 - \chi^2_{min})}{2}}$

Write $\{\theta\}_{old}$, $\chi^2_{old}$

$\{\chi^2_{min}, \{\theta\}_{Best}\}$

End loop MCMC