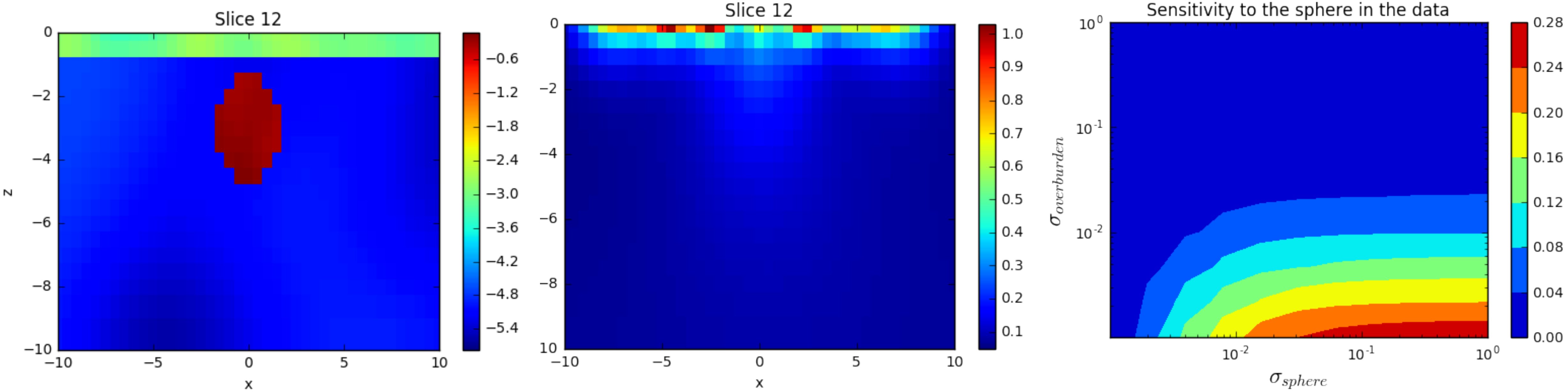
Designing objective functions: a probability approach

Thibaut Astic

SimPEG Meeting 28/02/2017

Motivation



Slice 12
z
x
−0.6
−1.2
−1.8
−2.4
−3.0
−3.6
−4.2
−4.8
−5.4
Slice 12
x
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Sensitivity to the sphere in the data
$\sigma_{overburden}$
σ_{sphere}
10^{0}
10^{-1}
10^{-2}
0.28
0.24
0.20
0.16
0.12
0.08
0.04
0.00

Basics

- Consider the data d as a realization of a multivariate probability distribution

$$d_{obs} \sim P(d|A, m, \Sigma)$$

- *m*: the model
- A: the physics operator
- Σ: The covariance matrix of your data

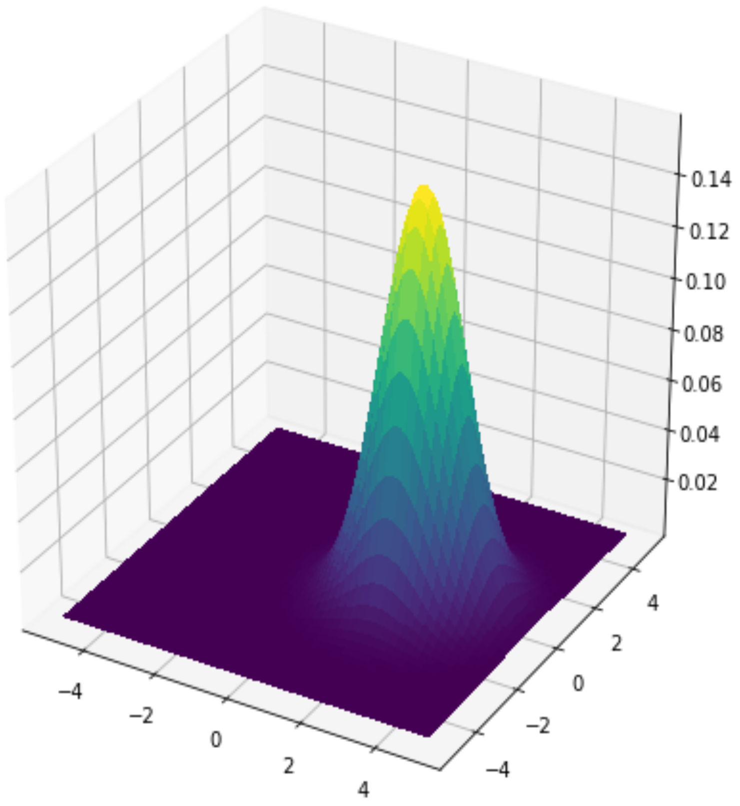
Maximum Likelihood Estimation

- Uncorrelated Gaussian Example

$$d_{obs} \sim P(d|A, m, \Sigma)$$

  - Assume:
    - P: multivariate Gaussian distribution
    - Mean is A**m*
    - Covariance: uncorrelated noise:

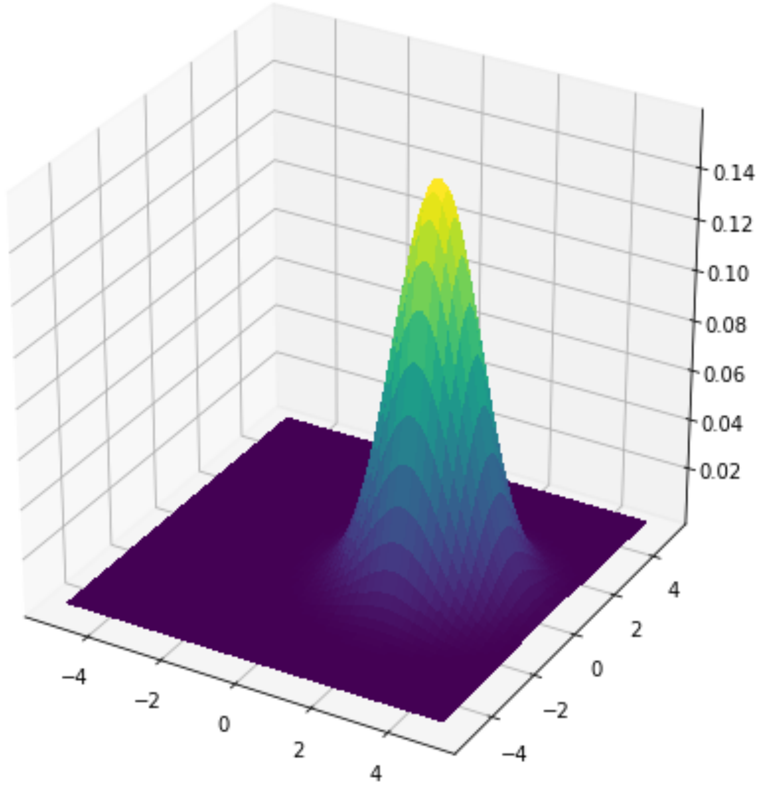
$$\Sigma = diag(\sigma_i^2)$$



$$P \sim \mathcal{N}$$

$$P(d|A, m, \Sigma) \propto \exp(-\frac{1}{2}(Am - d)^T \Sigma^{-1}(Am - d))$$

Maximum Likelihood Estimation



- Uncorrelated Gaussian Example

$$d_{obs} \sim P(d|A, m, \Sigma)$$

  - Assume:
    - P: multivariate Gaussian distribution
    - Mean is A**m*
    - Covariance: uncorrelated noise:

$$\Sigma = diag(\sigma_i^2)$$

$$P \sim N$$

Start to look familiar

$$P(d|A, m, \Sigma) \propto \exp(-\frac{1}{2}(Am - d)^T \Sigma^{-1}(Am - d))$$

Maximum Likelihood Estimation

- Uncorrelated Gaussian Example
  - Given *d*, we typically want to maximize the likelihood of *d* by playing on the model *m*. This is equivalent to minimize the negative log-likelihood,

$$\underset{m}{\operatorname{argmax}}\, P(d|A,m,\Sigma) = \underset{m}{\operatorname{argmin}} - \log(P(d|A,m,\Sigma))$$

  - which then give the familiar:

$$m_{opt} = \underset{m}{\operatorname{argmin}}\, \frac{1}{2}(Am-d)^T\Sigma^{-1}(Am-d)$$

$$= \underset{m}{\operatorname{argmin}}\, \frac{1}{2}||Am-d||^2_{\Sigma^{-1}}$$

Maximum Likelihood Estimation

- Uncorrelated Gaussian Example

$$m_{opt} = \underset{m}{\operatorname{argmin}} \frac{1}{2}(Am - d)^T \Sigma^{-1} (Am - d)$$

$$= \underset{m}{\operatorname{argmin}} \frac{1}{2} ||Am - d||^2_{\Sigma^{-1}}$$

  - But all models m are here equally good!
  - The real question is then not (only) to maximize the likelihood of *d*, but:
  - Given *d*, what is the most likely model? $P(m|A, d, \Sigma)$

Maximum A Posteriori Estimation

- How likely is my model?
  - Using Bayes rule:

$$P(m|A,d,\Sigma) = \frac{\overbrace{P(d|A,m,\Sigma)}^{\text{Data Fitting}}\overbrace{P(m)}^{\text{Prior/ Regularizer}}}{P(d)}$$

Normalization term

  - Need to choose a Prior:
    - For example: Multivariate Gaussian Prior

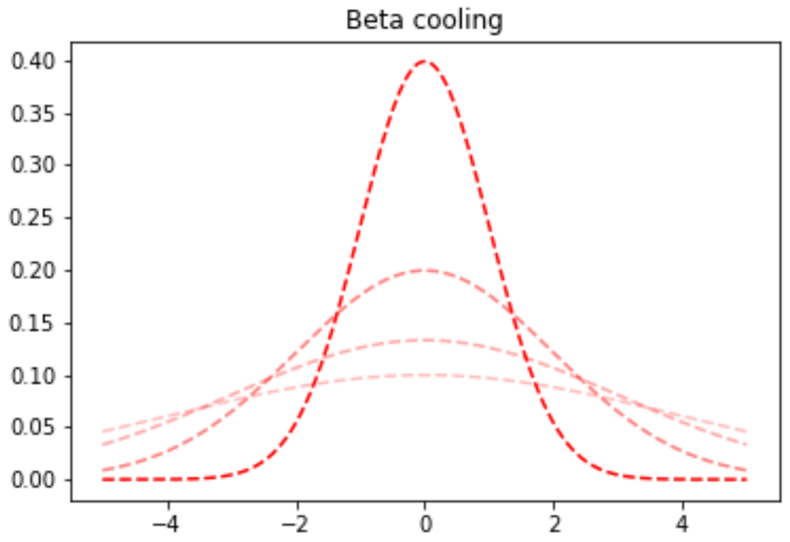
$$m \sim \mathcal{N}(m_{ref}, \frac{1}{\beta})$$

Maximum A Posteriori

- Uncorrelated Gaussian
  - Building on our previous example, we get the well-known Tikhonov regularization:

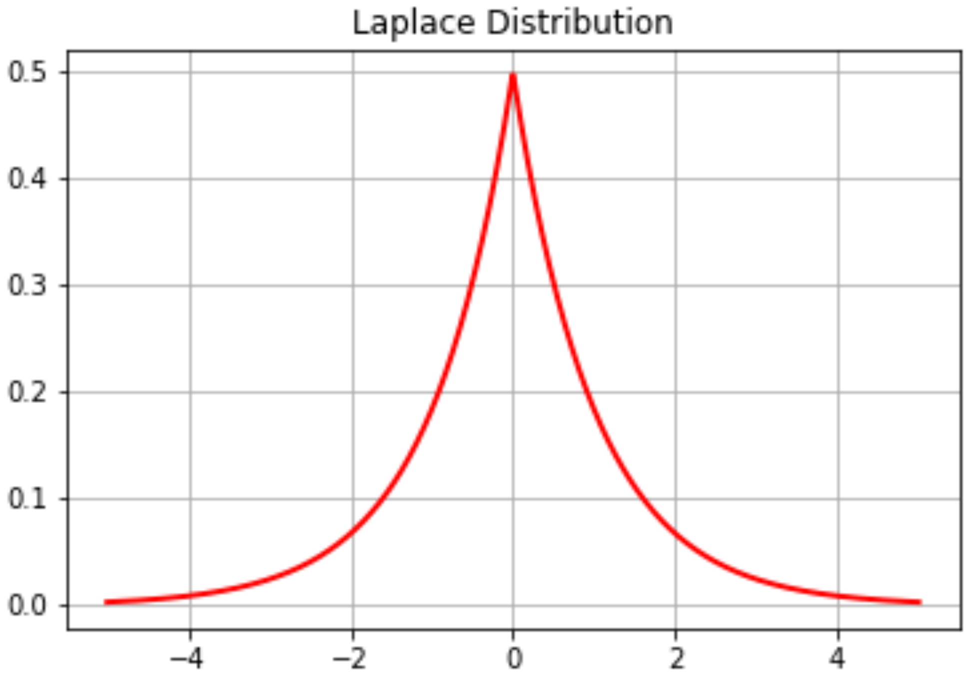
$$\underset{m}{\text{argmin}} - \log(P(m|A, d, \Sigma, m_{ref}, \beta)) = \frac{1}{2}||Am - d||^2_{\Sigma^{-1}} + \frac{\beta}{2}||m - m_{ref}||^2_2$$

- Take home:
  - Everything on the right side of 'P(m|…)' are our assumptions
  - L2-norm assumes a Gaussian Distribution of the parameters values
  - Beta-cooling: the smaller is Beta, the more likely become values far from the mean



What can we do with it?

- Use a different probability distribution



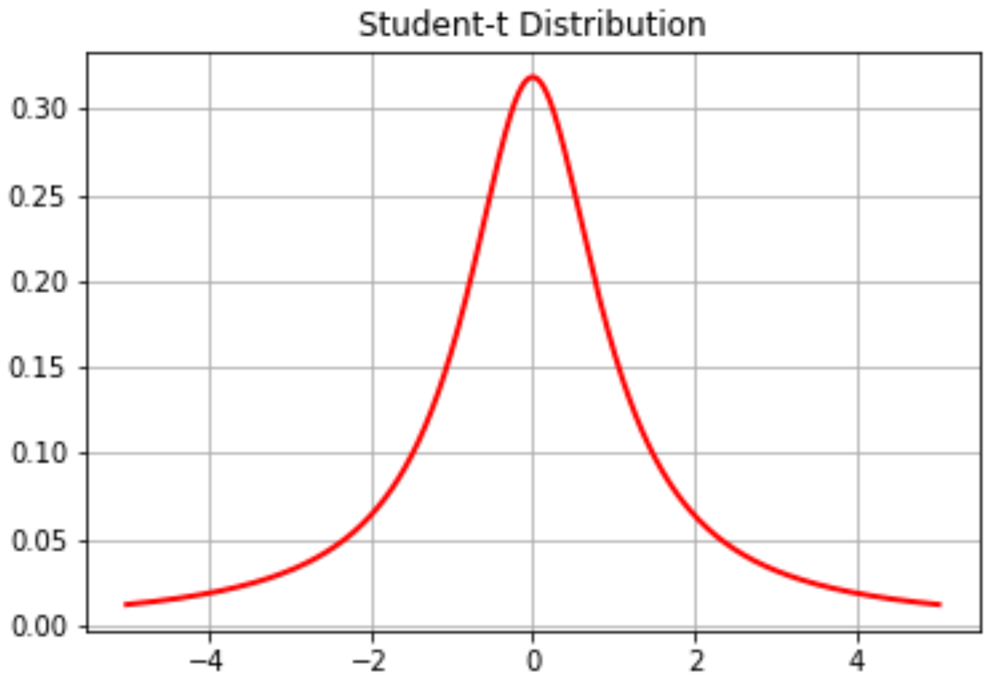
L1-Norm

Example: Regularizer: $||m||_1$

Very useful if your model is sparse in some basis (Space, Fourier…)

What can we do with it?

- Use a different probability distribution



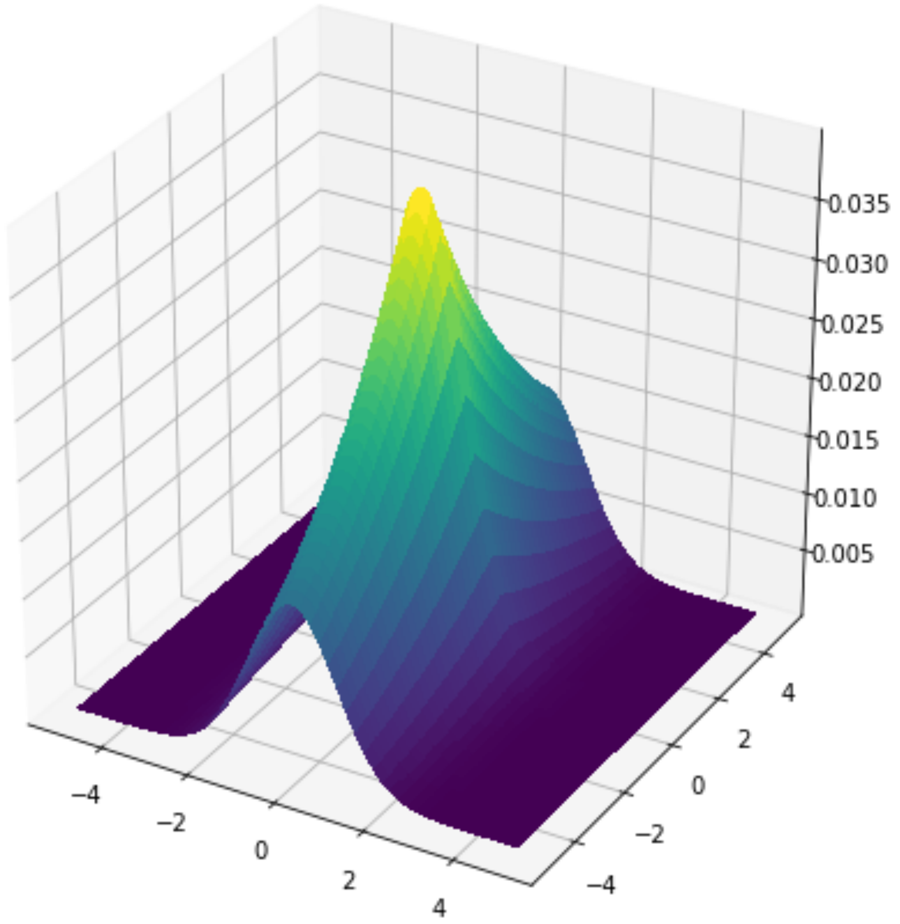
Very robust to outliers.

Example: Data fitting

$$\underset{m}{\operatorname{argmin}} \frac{v+1}{2} \sum_{i=1}^{n} \log\left(1 + \frac{(m^T a_i - d_i)^2}{v}\right)$$

What can we do with it?

- Compose different distributions to promote structures



- Some parts of the model will be sparse, others will be smooth
- Can also define dependencies between variables
  - Ex: If m1=0 then m2=0, else both are smoothed

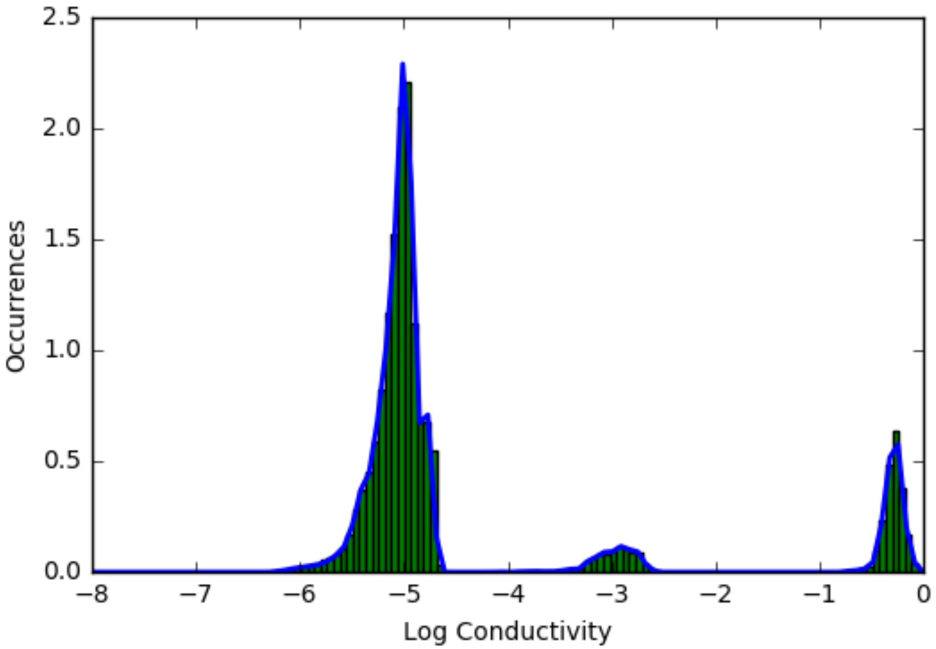
What can we do with it?

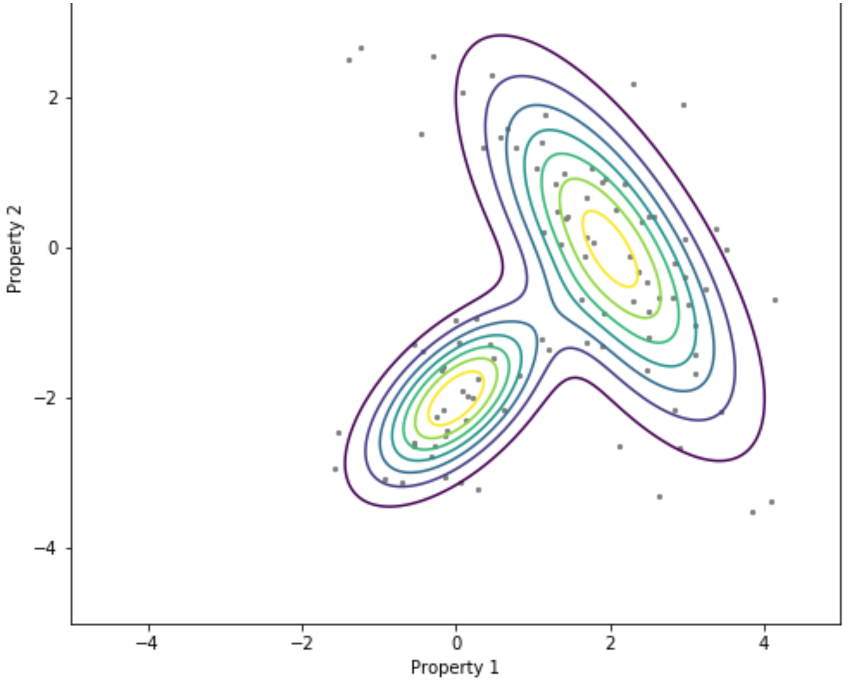
- My interest these days:

- Mixture of Gaussians prior

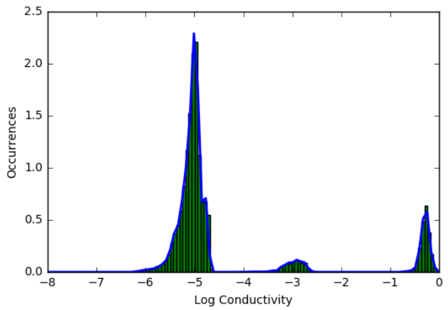
$$P(m) = \sum_{i=1}^{k} \pi_i p(m|z_i = c)$$

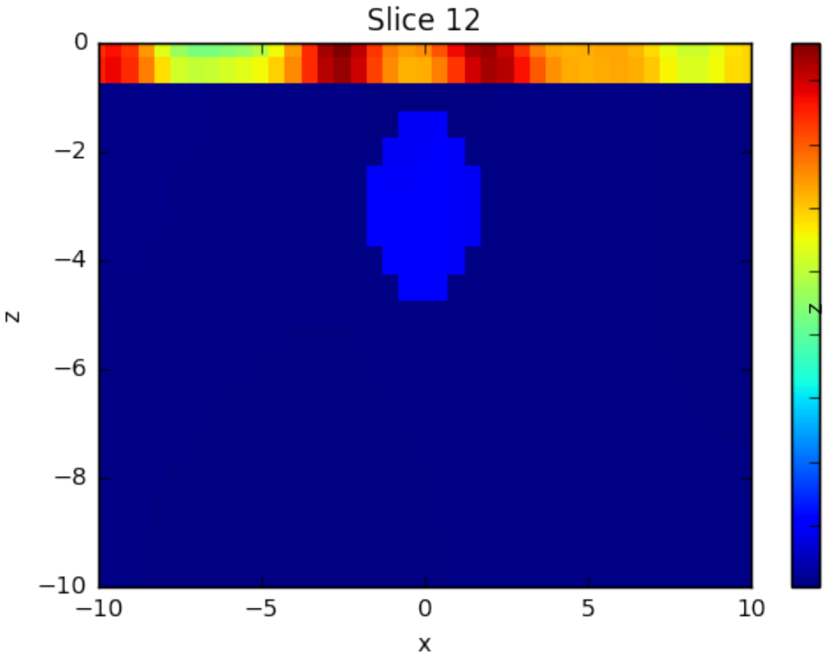
  - Can fit any parameter distributions
  - Can describe complex dependencies
  - Easily scalable (as long as data keep up with dimensionality)
  - Englobe k-means, Parzen Windows as special cases

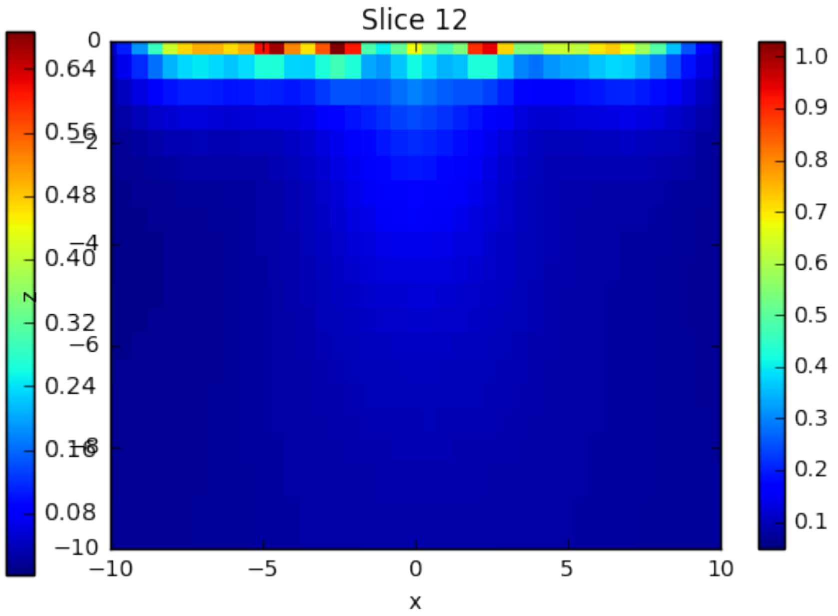


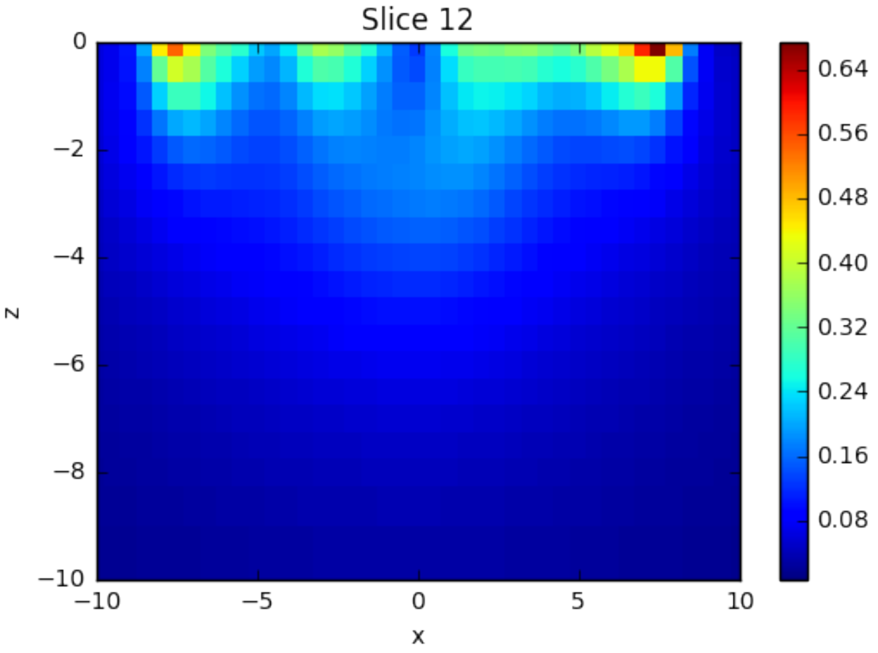


Motivation


Occurrences
Log Conductivity


Slice 12
z
x


Slice 12
z
x


Slice 12
z
x

Mixture of Gaussians as prior

- MAP estimator

$$\underset{m}{\operatorname{argmin}} \frac{1}{2}||Am - d||^2_{\Sigma^{-1}} - log(\sum_{i=1}^{k} \pi_i p(m|z_i = c))$$

- Where things get complicated numerically:
  - Log(sum(exp(…))) : well-known to underflow or overflow easily (there are some tricks fortunately)
  - Gradient and Hessian define everywhere (exp>0) but can get arbitrary large
  - No guarantee to have a positive negative log-likelihood

Mixture of Gaussians as prior