LA-UR-19-23704

Approved for public release; distribution is unlimited.

Title: An introductory note on Randomized Maximum Likelihood (RML) as it relates to quantitative inference from radiography

Author(s): Tobias, Benjamin John

Intended for: Report

Issued: 2019-04-24

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

An introductory note on
Randomized Maximum Likelihood (RML)
as it relates to
quantitative inference from radiography

B. Tobias, P-21

Purpose

- Identify the source citations for the RML method and clearly delineate the relevant analysis that is contained within
- Relate the RML paradigm to inference from radiography and thereby contextualize the source citations as *limited justification for an intuitive Monte Carlo technique*
- Identify a potential pitfall of applying RML as it is presented in the source citations and prescribe a deviation from techniques adopted by the AWE and presented by researchers there

2 original works, roughly consistent notation

D.S. Oliver, On conditional simulation to inaccurate data, *Mathematical Geology* **28**, pp. 811-817 (1996)

Describes the nature of the model and the data:

- Model parameters (quantities of interest, independent variables) constitute a field with a 'smooth' and a 'rough' part: inherent variability
- Defines 'conditional' and 'unconditional' simulation
- Data further includes measurement errors
- *Both sources of variability are covariant Gaussian distributions*

D.S. Oliver, H.E. Nanqun, and A.C. Reynolds, Conditioning permeability fields to pressure data, Proc. ECMOR V, pp. 259-269 (1996)

Describes the RML Markov Chain methodology:

- Adopts terminology of 'calibrated' simulation

Examines validity of accepting all transitions:

- AWE: 'Exact for linear, Gaussian problems' and 'Approximate for univariate nonlinear models'
 - *Not quite what the paper actually says*
- *Provides exact expression for Metropolis-Hastings acceptance criterion*
- *Provides 2 expressions for the cost function that are not necessarily equivalent*

‘Truth’ and estimates from data (definitions)

The ‘truth’ is a random field:

$$Z(x) = m(x) + \epsilon(x)$$

A ‘best estimate’ has a well-defined relationship to the smooth part of the truth and noisy data:

$$\hat{Z} = m + H^{-1}(d^{obs} - Fm)$$

Unconditional and conditional (or calibrated) fields and measurements are defined in relation to each other:

$$Z^{ucs} = m + C_E^{1/2} S$$

covariance of the field
a random number $\sim N(0,1)$

$$d^{ucs} = FZ^{ucs} + C_D^{1/2} S_D$$

covariance of the data
a random number $\sim N(0,1)$

$$\hat{Z}^{ucs} = m + H^{-1}(d^{ucs} - Fm)$$

$$Z^{cs} = \hat{Z} + [Z^{ucs} - \hat{Z}^{ucs}], \quad Z^{cs} = \hat{Z} + C_E^{1/2} S$$

‘Truth’ and estimates from data (properties)

Note changes in notation!

Unconditional fields and data:

- Have the same statistical properties as the random field and the data
- Are statistically independent samples of the noise and error
- Are distributed according to a Gaussian prior distribution

$$f(m_{us}, d_{us}) \propto \exp \left\{ -\frac{1}{2} (m_{us} - \mu)^T C_M^{-1} (m_{us} - \mu) - \frac{1}{2} (d_{us} - d_{obs})^T C_D^{-1} (d_{us} - d_{obs}) \right\}$$

Conditional fields:

- Also have the same statistical properties as the random field and the data
- Are also statistically independent samples of the noise and error
- Are distributed according to inference from the observed data

$$m_{us} = m + H^{-1} (d^{obs} - Fm) + C_E^{1/2} S$$

RML: the Oliver Markov Chain

- Sample (m_{us}, d_{us}) from the prior distribution
 - No memory of previous state!
- Obtain a ‘calibrated’ $(m, g(m))$ by fitting to (m_{us}, d_{us})

$$O(m) = (m - m_{us})^T C_M^{-1} (m - m_{us}) + [g(m) - d_{us}]^T C_D^{-1} [g(m) - d_{us}]$$

- Accept all transitions, efficiently generating a histogram $\pi(m_{cal})$

$$\frac{\pi_j q_i}{\pi_i q_j} = \int_{D_{us}} \exp\{-F(\epsilon, \eta)\} dd_{us} \quad F \text{ is at least linear in small terms } (\epsilon, \eta)$$

How is RML hard-wired in the BIE?

- Independent variables are not treated as a covariant random field:

- $Z(x) = m(x)$, $Z = Z^{ucs}$
- samples $d_{us} = g(\hat{Z}) + C_D^{1/2} S_D$
- $O(m) = \chi^2 / \sigma^2$

→ Algorithm looks exactly like random trials fitting a synthetic dataset!

- Validity of estimates from this method are dependent upon:
 - validity of assumption $\hat{Z} = Z(x)$ (PDF always biased by ML estimate)
 - validity of noise model, $C_D^{1/2} S_D$ (here Gaussian, but could be any prescribed model)
 - scale of systematics and noise, (ϵ, η) (fit and M.-H. acceptance criterion intertwined)

An error in implementation?

- In previous work, the cost function has been wired up as:

$$S(m) = \left[g(m) - (d_{obs} - C_D^{1/2} S_D) \right]^T C_D^{-1} \left[g(m) - (d_{obs} - C_D^{1/2} S_D) \right]$$

This comes directly from the 2nd reference (Eq. 3), but does not agree with expressions used in the fundamental analysis. Comparing to Eq. 14 of the same reference, it requires

$(d_{obs} - C_D^{1/2} S_D) = d_{us}$, which is not true. If the statistical properties of d_{us} are defined by d_{obs} , then $d_{obs} - C_D^{1/2} S_D$ has different variance than either d_{us} or d_{obs} .

Subtracting a random variable does not remove noise, it increases it

Summary

- When applied to problems of inference from radiography, RML amounts to Monte Carlo fitting to synthetic data
 - an intuitive solution to lacking lots of real data for ‘known objects’
 - subject to obvious limitations (e.g. validity of ML and noise model)
- Appropriate for parallel computing (as opposed to MCMC or HMC)
 - requires no memory of previous sample or state
- The source citations for RML provide rigorous justification for accepting all samples in the ‘Markov Chain’
 - M.-H. criterion approximately satisfied for many problems (*for proper cost function!*)
- Source citations DO NOT justify using a single real dataset in the Monte Carlo sampling process!
 - Real data has to be analyzed to generate ‘unconditional simulations,’ but *there is no analysis provided that justifies inserting the observed data into the cost function for calibration*