Glossary of important terms for GW1876

April 21, 2018

Notes on mathematics symbology

It would be really nice if a totally consistent set of symbols was used in the mathematics of parameter estimation and uncertainty analysis. But....no such luck! So many researchers in different fields over a long time have contributed to the work used in this class. As a result, this glossary is an attempt to highlight some general symbology and clarify some terms.

First of all, though, one thing that is consistent (mostly!) is the general linear algebra notation used throughout the class.

scalar values Lowercase, non-bold font indicates a scalar (single) value: x, y, z

vectors Lowercase, bold font indicates a vector of values: $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$

matrices Uppercase, bold font indicates a matrix of values: $\mathbf{X}$, $\mathbf{Q}$, $\mathbf{J}$ A matrix with $< \cdot >^T$ indicates a matrix transpose. A matrix with $< \cdot >^{-1}$ indicates a matrix inverse.

matrix multiplication Then, matrix multiplication (with either other matrices or vectors) is expressed simply by adjacent matrices: $\mathbf{Xy}$, $\mathbf{X}^T\mathbf{Q}^{-1}\mathbf{X}$

Glossary of terms and equations

Parameters Variable input values for models, typically representing system properties and forcings. Values to be estimated in the history matching process. Typically identified as k, p, or x, θ, ($\mathbf{k}$, $\mathbf{p}$, $\boldsymbol{\theta}$ or $\mathbf{x}$ for multiple parameters in a vector).

Observation Measured system state values. These values are used to compare with model outputs collocated in space and time. The term is often used to mean *both* field measurements and outputs from the model. When referring to a measured value, observations are typically identified by the variables y or o ($\mathbf{y}$ or $\mathbf{o}$ for multiple parameters in a vector)

Modeled Equivalent (aka Simulated Equivalent) A modeled value collocated in time and space with an observation. There are various ways to identify a single or multiple modeled equivalent values (and, to make things confusing, they are often *also* called "observations"!)

Single values

1. $f(x)$
2. $X(\beta)$
3. $M(p)$

Multiple values

1. $\mathbf{X}\beta$
2. **Mp**
3. **NOBS** Number of observations/simulated equivalents in the inverse model setup
4. **NPAR** Number of adjustable input parameters in the inverse model setup

Forecasts Model outputs for which field observations are not available. Typically these values are simulated under an uncertain future condition.

Phi Objective function, defined as the weighted sum of squares of residuals. Phi (aka Φ) is typically calculated as

$$\Phi = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i)}{w_i} \right)^2 \quad or \quad \Phi = (\mathbf{y} - \mathbf{Jx})^T \mathbf{Q}^{-1} (\mathbf{y} - \mathbf{Jx}) \tag{1}$$

Residuals The difference between observation values and modeled equivalents $r_i = y_i - f(x_i)$

Sensitivity The incremental change of an observation (modeled equivalent, actually) due to an incremental change in a parameter. Typically expressed as a finite-difference approximation of a partial derivative: $\frac{\partial y}{\partial x}$

Jacobian Matrix A matrix of the sensitivity of all observations in an inverse model to all parameters. This is often shown as a matrix by various names $\mathbf{X}$, $\mathbf{J}$, or $\mathbf{H}$. Each element of the matrix is a single sensitivity value $\frac{\partial y_i}{\partial x_j}$ for $i \in NOBS$, $j \in NPAR$

Regularization A preferred condition pertaining to parameters, the deviation from which, elicits a penalty added to the objective function. This serves as a balance between the level of fit or "measurement Phi" ($\mathbf{\Phi_M}$) and the coherence with soft knowledge/previous conditions/prior knowledge/regularization ($\mathbf{\Phi_R}$). These terms can also be interpreted as the likelihood function and prior distribution in Bayes' theorem (see below)

PHIMLIM A PEST input parameter the governs the strength with which regularization is applied to the objective function. A high value of PHIMLIM indicates a strong penalty for deviation from preferred parameter conditions while a low value of PHIMLIM indicates a weak penalty. The reason this "dial" is listed as a function of PHIM (e.g. $\mathbf{\Phi_M}$) is because it can then be interpreted as a limit on how well we want to fit the observation data.

FOSM *fill this in*

Gaussian (multivariate) The equation for Gaussian (Normal) distribution for a single variable (x) is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{2}$$

where μ is the mean and σ is the standard deviation The equation for a multivariate Gaussian for a vector of k variables ($\mathbf{x}$) is

$$f(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} e^{-\frac{1}{2}\left((\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mu)\right)} \tag{3}$$

where μ is a k-length vector of mean values, $\mathbf{\Sigma}$ is the covariance matrix, and $|\mathbf{\Sigma}|$ is the determinant of the covariance matrix

Weight or Epistemic Uncertainty A value by which a residual is divided by when constructing the sum of squared residuals. In principal, $w \approx \frac{1}{\sigma}$ where σ is an approximation of the expected error between model output and collocated observation values. While the symbol σ implies a standard deviation, it is important to note that measurement error only makes up a portion of this error. Other aspects such as structural error (e.g. inadequacy inherent in all models to perfectly simulate the natural world) also contribute to this expected level of error. The reciprocal of weights are also called Epistemic Uncertainty terms.

Weight Covariance matrix (correlation matrix) In practice, this is usually a $NOBS \times NOBS$ diagonal matrix with values of weights on the diagonal representing the inverse of the observation covariance. This implies a lack of correlation among the observations. A full covariance matrix would indicate correlation among the observations which, in reality, is present but, in practice, is rarely characterized. The weight matrix is often identified as $\mathbf{Q}^{-1}$ or ${\mathbf{\Sigma}_\epsilon}^{-1}$

Parameter Covariance matrix The uncertainty of parameters can be expressed as a matrix as well. This is formed also as a diagonal matrix from the bounds around parameter values (assuming that the range between the bounds indicates 4σ (e.g. 95% of a normal distribution). In `pyemu`, some functions accept a `sigma_range` argument which can override the 4σ assumption. In many cases of our applications, parameters are spatially distributed (e.g. hydraulic conductivity fields) so a covariance matrix with off-diagonal terms can be formed to characterize not only their variance but also their correlation/covariance. We often use geostatistical variograms to characterize the covariance of parameters. The parameter covariance matrix is often identified as $C(\mathbf{p})$, $\mathbf{\Sigma}_\theta$, or $\mathbf{R}$.

Measurement noise/error Measurement noise is a contribution to Epistemic Uncertainty. This is the expected error of repeated measurements due to things like instrument error and also can be compounded by error of surveying a datum, location of an observation on a map, and other factors.

Structural (model) error Epistemic uncertainty is actually dominated by structural error relative to measurement noise. The structural error is the expected misfit between measured and modeled values at observation locations due to model inadequacy (including everything from model simplification due to the necessity of discretizing the domain, processes that are missing from the model, etc.)

Monte Carlo Parameter Realization A set of parameter values, often but not required to be a multi-Gaussian distribution, sampled from the mean values of specified parameter values (either starting values or, in some cases, optimal values following parameter estimation) with covariance from a set of variance values, or a covariance matrix. Can be identified as θ

Monte Carlo Observation Realization A set of observation values, often but not required to be a multi-Gaussian distribution, sampled using the mean values of measured observations and variance from the observation weight covariance matrix. Can be identified as $\boldsymbol{d_{obs}}$

Monte Carlo Ensemble A group of realizations of parameters ($\mathbf{\Theta}$), observations ($\mathbf{D_{obs}}$) and the simulated equivalent values $\mathbf{D_{sim}}$. Note that these three matrices are made up of column vectors representing all of the $\boldsymbol{\theta}$, $\mathbf{d_{obs}}$, and $\mathbf{d_{sim}}$ vectors where $\mathbf{d_{sim}}$

Bayes' Theorem $P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\mathbf{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{D})}$ ... $\underbrace{P(\boldsymbol{\theta}|\mathbf{D})}_{\text{posterior pdf}} \propto \underbrace{P(\boldsymbol{D}|\boldsymbol{\theta})}_{\text{likelihood function}} \underbrace{P(\boldsymbol{\theta})}_{\text{prior pdf}}$

Posterior (multivariate distribution)

Schur Complement

Prior (multivariate distribution)

Likelihood (multivariate distribution)