Conditional and marginal probabilities in AEM inversions using multivariate Gaussian statistics

Aaron C Davis
CSIRO ESRE
26 Dick Perry Ave
Kensington WA 6151
Aaron.davis@csiro.au

Andrew King
CSIRO ESRE
26 Dick Perry Ave
Kensington WA 6151
andy.king@csiro.au

Niels Christensen
Arhus University
C.F. Moellers Allé 4
DK-8000 Aarhus C
nbc@geo.au.dk

Tim Munday
CSIRO ESRE
26 Dick Perry Ave
Kensington WA 6151
tim.munday@csiro.au

SUMMARY

We summarise and extend the concept of Gaussian, or normal, distributions into multivariate statistics over many dimensions. We demonstrate how multivariate statistics can be applied to probability distributions. Through assumptions in the linearisation of the inverse problem, we show that the best-fit inverse model parameters are normally distributed with mean values and associated variance and covariance values that obey Gaussian statistics. Variance and covariance values describe how the model parameters interact with each other. By changing one value in the model parameter vector, other parameters are changed through the covariance that links them. We apply Gaussian statistics over many dimensions to query our models for statistically meaningful questions that can only be answered by taking the integral of the multivariate distribution over the multidimensional space that contains the model parameter values. We illustrate this with an example of aquifer detection, using resistivity limits, for an electromagnetic transect adjacent to the Gascoyne River near Carnarvon, Western Australia.

Key words: Airborne electromagnetics, AEM, Gaussian, covariance, aquifer, probability.

INTRODUCTION

With emphasis being placed on uncertainty in groundwater modelling and prediction, coupled with questions concerning the value of geophysical methods in hydrogeology, it is important to ask meaningful questions of hydrogeophysical data and inversion results. For example, to characterise aquifers using electromagnetic (EM) data, we ask questions such as 'Given that the electrical conductivity of aquifer 'A' is less than x, where is that aquifer elsewhere in the survey area?' An answer to this question may be given by examining inversion models, selecting locations and layers that satisfy the condition 'conductivity <= x', and labelling them as aquifer 'A'. One difficulty with this approach is that the inversion model result is often considered to be the only model for the data. In reality it is just one image of the subsurface that, given the method and the regularisation imposed in the inversion, agrees with measured data within a given error bound. We have no idea whether the final model realised by the inversion satisfies the global minimum error, or whether it is simply in a local minimum. There is a distribution of inversion models that satisfy the error tolerance condition: the final model is not the only one, nor is it necessarily the correct one (Tarantola, 2005).

AEM inversions often involve linearised approximations to the calculation of the Jacobian and Hessian matrices in the forward solution. This results in a second order Taylor series expansion of the estimation of an error surface when calculating the misfit between forward model data and measured data. In the vicinity of a minimum in the error surface, the first-order terms drop out and only second-order terms are present. This guarantees that model parameters resulting from the inversion will be Gaussian distributed with mean model parameter values and model parameter variance terms. The end product is the output of the inversion that is most often used, and we produce conductivity-depth sections from the 'best-fit' model parameters. In this approach, however, we are neglecting the fact that those parameters contain variances and correlations, which are also of value. In reality, because of the way we have constructed the inversion scheme, the model covariance terms contain information about how each model parameter interacts with each other model parameter for a given inversion output result. The collection of variances is assembled in a posterior covariance matrix, where terms on the main diagonal are the autocorrelation values and terms off the diagonal are the cross-correlation or covariance terms.

We can examine the posterior covariance matrix terms, and exploit the well-known characteristics of Gaussian statistics to ask meaningful conditional and marginal probability questions from the inversion data. This is done so that we can ask questions such as: 'Given that the aquifer I am interested in has resistivity ranges between ρ_{low} and ρ_{high}, where, how deep and how thick is the aquifer?' The result is a probability map that shows the most likely location of the structure of interest given the conditional statements made when posing the question.

In this presentation, we use the posterior covariance matrices from maximum-gradient inversion schemes, such as Em1dinv (Auken et al., 2005), to answer quantitative statistically meaningful questions about aquifer location, depth and thickness in an AEM system selection process for groundwater exploration along the Gascoyne River in Carnarvon, Western Australia. We extend upon the results of Christensen and Reid (Christensen and Reid 2012) whereby, instead of using a stochastic approach to answer questions of the inversion results, we instead examine multivariate Gaussian integrals to evaluate our statistics. We show the most probable depth and thickness of a portion of one of the more productive aquifers located along the northern side of the extended borefield reticulation development at Carnarvon.

THE GAUSSIAN DISTRIBUTION

First introduced by DeMoivre, and then extensively improved by Gauss, the normal distribution is probably most recognisable in its standardised modern form by Fisher (Fisher, 1990). A variable x has a Gaussian, or normal, distribution when it can be expressed as $\mathcal{N}(\mu,\sigma^2)$, a distribution that has some mean value μ and a standard deviation of σ such that the functional differential form of the distribution is:

$$df = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

All forms of $\mathcal{N}(\mu,\sigma^2)$ can be re-expressed in the more convenient form of the standardised normal distribution of $\mathcal{N}(0,1)$, such that the new variable z is renormalised to having a mean value of 0 and a standard deviation of 1. This standard form leads us directly to the use of the z-statistic (whose integral over all values of z is normalised to 1), with the expression

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}z^2} dz,$$

and is shown graphically in Figure 1.

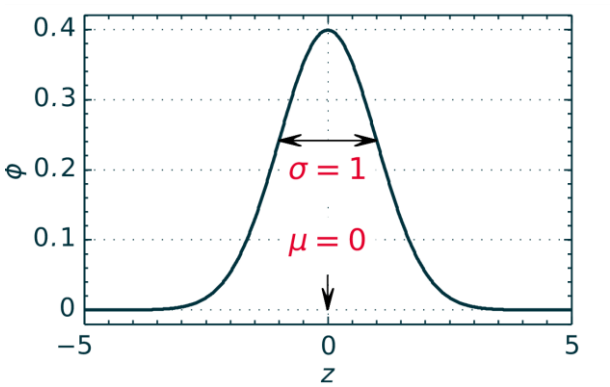

Figure 1. The z-distribution: a standardised normal distribution with mean 0 and standard deviation 1. All Gaussian distributions may be expressed like this.

The Gaussian distribution becomes important when we consider the function $\phi(z)$ to represent a probability distribution function of the variable z, and this is useful in considering random processes, curve fitting, error and, in this paper, inversion.

Extension to more than 1 dimension

We can easily extend the Gaussian distribution of variables to more than one dimension by considering the collection of variables $x_1, x_2, \ldots x_N$ as a vector of variables $\mathbf{x}$. Each variable still retains its own standard deviation σ, but these are now more conveniently expressed using the concept of variance (which is simply the square of the standard deviation). We therefore extend the notation of standard deviation into σ_{ij}, where it is implicit that this value represents the true variance of the variable x_i when $i = j$ (in σ_{ii}), and it represents the covariance of variable x_i and x_j when we examine σ_{ij}. We accumulate the variances and covariances into matrix $\mathbf{\Sigma}$, whereby

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & & \sigma_{21} \\ \vdots & & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{NN} \end{bmatrix},$$

and the mean value of each variable into vector $\boldsymbol{\mu}$. Each element μ_i of $\boldsymbol{\mu}$ is the mean value of x_i. The Gaussian probability distribution then takes the more general multivariate form of

$$\phi(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Sigma}|^{1/2}} \int e^{\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\right)} d\mathbf{x}.$$

AN EXAMPLE IN TWO DIMENSIONS

We consider an example of variable distributions in two dimensions, since it is easy to visualise. The variables x_1 and x_2 both have mean 0 and variance 1, but they are related to each other by the covariance $\sigma_{12} = 0.6$ such that

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \text{and } \mathbf{\Sigma} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}.$$

A realisation of their distribution is shown in Figure 2.

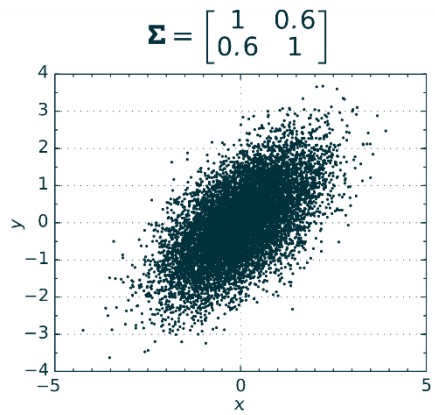

Figure 2. A realisation of the 2-D multivariate distribution of 2 variables, both of mean 0, variance 1 and covariance of 0.6.

From the probability distribution function, we can ask the question 'What is the probability of $x_1 \geq 1$?' (Answer: 0.157) or the question 'What is the probability of $x_2 \geq 0.5$?'(Answer: 0.31). However, by knowing the covariance between x_1 and x_2, we can also ask the question 'What is the probability of $x_2 \geq 0.5$, *given that we know that* $x_1 \geq 1$?' This is now a conditional probability, and the covariance between x_1 and x_2 indicates that their relationship is important. The answer, 0.109, is calculated by taking the integral of the probability distribution function over both dimensions 1 and 2.

THE NORMAL DISTRBUTION IN INVERSION

In electromagnetic geophysics, the relationship between survey data and a valid model of the conductivity structure of the earth, is problematic. There are a limited number of direct, closed-form expressions between conductivity and a transmitter-receiver array. Even a relatively simple model, such as a 1-dimensional layered earth with dipole transmitters and receivers (eg Wait, 1982) results in a non-linear relationship between data and model structure. In order to invert upon such functions, we often simplify the problem by

linearising the calculation of the derivatives of the sensitivity of the function to the model parameters. The resulting Taylor expansion, in the vicinity of a minimum point of the misfit calculation, ensures that the first-order terms are effectively zero, and that our estimation of the error surface is only second-order. The resulting error calculation ensures that each of our model parameters, that are sufficiently in the vicinity of a minimum, be distributed with mean values μ_i, variance σ_{ii}, and covariance with model parameter j as σ_{ij}. The inversion of the measured data, and the generation of the conductivity-depth model produces both the vector $\boldsymbol{\mu}$ (which is what we call the best-fitting model, and is the 'end-product' for most inversions); but it also produces the posterior covariance matrix $\mathbf{C}$ which is an expression of the variance and cross-correlation between model parameters in the inversion based on the assumptions that we have made regarding the calculation of the sensitivity of the forward function in the vicinity of a minimum in the error surface. We therefore can express the end result of an inversion as a collection of mean values normally distributed about the mean with multidimensional auto- and cross-correlations connecting the expression of the mean parameters. Changing one parameter necessarily changes another (eg Aster et al., 2005).

GASCOYNE RIVER, WESTERN AUSTRALIA

As part of its Water for Food Program, the Department of Agriculture and Food, Western Australia (DAFWA) has an interest in determining the full extent of the groundwater resources along the Gascoyne River near Carnarvon, Western Australia. To this end DAFWA is partnering with CSIRO to investigate the potential use of airborne electromagnetic (AEM) surveys for groundwater exploration and detection. As part of the AEM system selection process, ground TEM data were acquired along a 400 m north-south transect from borehole 1910 to the shoreline of the Gascoyne River (Figure 4). The target aquifer, which bore 1910 taps into, is located between 15 m to 30 m depth. It is expected to persist to the Gascoyne River where it is recharged during the wet season. Forward and inverse models for an AEM system are derived by using the ground TEM conductivity-depth profiles as 'truth' in the forward model, and inverting upon the 'true' forward data for the system being considered.

We characterise the target aquifer by defining the following parameters: depth to the top of the aquifer d_t; depth to the bottom of the aquifer d_b, lower limit of aquifer resistivity ρ_{Al} (10 Ωm); and upper limit of resistivity ρ_{Au} (∞). To determine detection of the aquifer, we ask the following question: 'For every layer i of the inverse model, and for every layer j of the same model (whereby j is equal to or greater than i up to layer N of my inverse model), what is the probability that layers i to j satisfy my aquifer conductivity range, given that layers j+1 to N do not satisfy this range?' This question, which relies heavily upon the auto- and cross-correlation values of the posterior covariance matrix that results from the inversion scheme used, is answered through integration of multivariate Gaussian distributions over N dimensions. Each station involves $N^2/2$ layer combinations of the inversion model whereby the probability distribution is queried (for example, layer 1 can have j = 1 to N, layer 2 can have j = 2 to N, etc). The probability value for each layer at the sounding location of 280 m is shown in Figure 3. By summing the probability value over each layer, we arrive at a cumulative probability distribution for an individual sounding. In the example shown in Figure 3, we see that the greatest probability of an aquifer satisfying our resistivity criteria occurs over layers 14 to 20. This cumulative probability distribution is then displayed in section for each sounding. The resulting probability transect is shown in Figure 5, whereby the most probable depths and thicknesses of the aquifers are shown in dark blue, whereas areas with lower probability are lighter in colour.

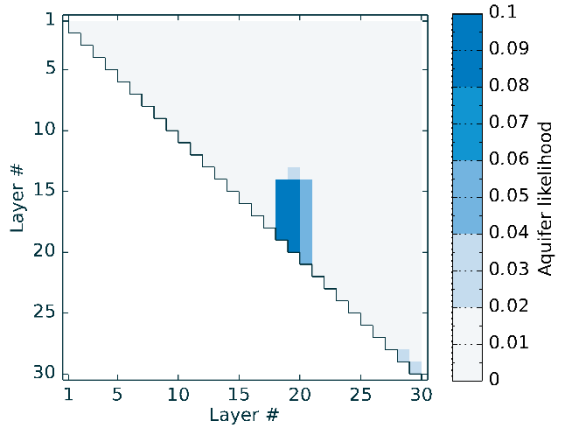

Figure 3. Probability value for each layer of the inverted model parameters of the 280 m AEM sounding based on resistivity limits of the aquifer characterisation question.

CONCLUSIONS

In this presentation, we have summarised the concept of Gaussian distributions, whereby a variable is statistically described as having a mean value and a variance, and extended to multivariate distributions. The mean value of the variables, whose distribution can be expressed as a vector of mean parameters is described with a variance (in the variable itself) and a series of covariances which describe how the variable interacts with other variables in the collection. We show how, through the assumptions in an inversion scheme, the concept can be adapted to the estimation of the model parameters. Every model parameter is then a Gaussian variable with mean value and a vector of variance and covariance. We exploit the multidimensional relationship of the model parameters to pose statistical questions of the inversion data, an example of which has been shown for aquifer detection in the Gascoyne River near Carnarvon, Western Australia. The resulting cumulative probability section affords a method of aquifer detection where the more likely areas of an aquifer, based on electromagnetic data, exist in the subsurface.

REFERENCES

Aster, R.C., Thurber, C.H. and Borchers, B., 2005, *Parameter estimation and inverse problems*: Elsevier Academic Press.

Auken, E. et al., 2005, Piecewise 1D laterally constrained inversion of resistivity data: *Geophysical Prospecting*, **53**(4), 497–506.

Christensen, N.B. and Reid, J., 2012, Assessing the presence of hard rock along a gas pipeline alignment with airborne EM: *ASEG Extended Abstracts*, **2012**(1), 1–4.

Fisher, R.A., 1990, *Statistical inference and analysis: Selected correspondence of R.A. Fisher*: Clarendon Press.

Tarantola, A., 2005, *Inverse problem theory and methods for model parameter estimation*: Society for Industrial and Applied Mathematics.

Wait, J.R., 1982, *Geo-electromagnetism*: Academic Press.

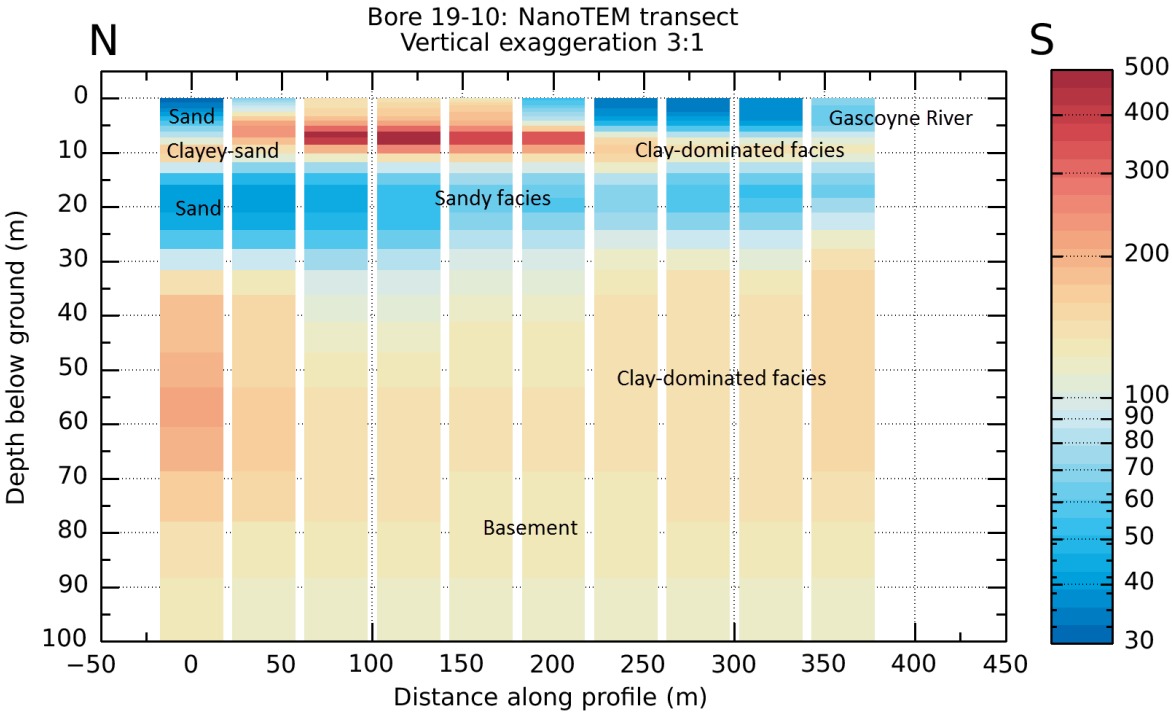

Figure 4. A 400 m NanoTEM transect used as a ground-model for AEM system selection. Interpretation of lithology and aquifer characteristics are overlain on the models.

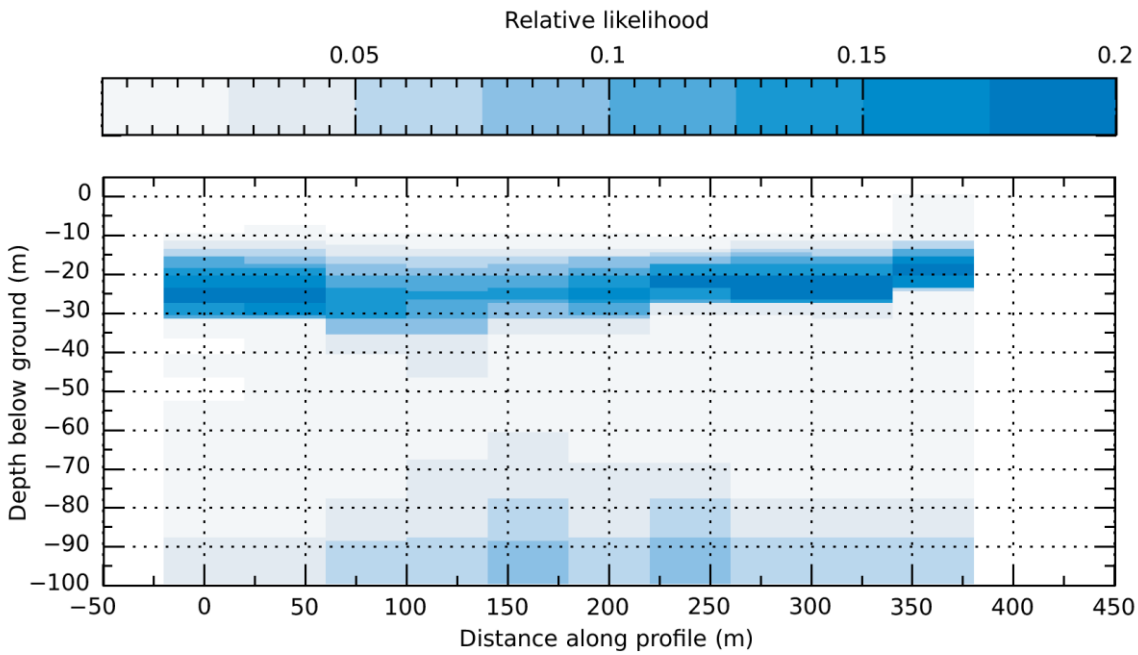

Figure 5. Aquifer likelihood based on consideration of the aquifer determination parameters and the cumulative conditional probability calculated from the inversion model mean resistivity vector and posterior correlation matrix. Multivariate Gaussian statistics are used to calculate the conditional probability for each layer, and then we accumulate layer probability to derive a relative likelihood of finding layers that satisfy the aquifer conditions.