NATIONAL TECHNICAL UNIVERSITY OF ATHENS

SCHOOL OF MINING AND METALLURGICAL ENGINEERING

DOCTORAL THESIS

THE C-ISR METHOD FOR FACIES INVERSION

Author: George J. Valakas

Advisor: Professor K. Modis

Athens 2018

Selection Board:

Apostolopoulos George: Associate Professor in National Technical University of Athens - School of Mining and Metallurgical Engineering.

E-mail: gapo@metal.ntua.gr

Boudouvis Andreas: Professor in National Technical University of Athens - School of Chemical Engineering.

E-mail: boudouvi@chemeng.ntua.gr

Hristopoulos Dionissios: Professor in Technical University of Crete - School of Mineral Resources Engineering.

E-mail: dionisi@mred.tuc.gr

Kyriakidis Phaedon: Professor in Cyprus University of Technology - Department of Civil Engineering and Geomatics.

E-mail: phaedon.kyriakidis@cut.ac.cy

Mantoglou Aristotelis: Professor in National Technical University of Athens - School of Rural and Infrastructure Development.

E-mail: mantog@central.ntua.gr

Modis Konstantinos: Professor in National Technical University of Athens - School of Mining and Metallurgical Engineering.

E-mail: kmodis@mail.ntua.gr

Vafidis Antonis: Professor in Technical University of Crete - School of Mineral Resources Engineering.

E-mail: vafidis@mred.tuc.gr

...to my parents, John and Anastasia,
for their love and support

Abstract

In the present Doctoral thesis, we develop and present the Cosimulated Iterative Spatial Resampling (C-ISR) method for stochastic solution of the inverse problem of hydrofacies characterization in a groundwater flow system and we establish the theoretical and mathematical background of the method. The development of the method stems from the need to characterize the geological formations before any other geostatistical estimation of hydrological parameters, due to the complexity of soil types as natural entities. The spatial distribution of hydrofacies could be estimated first by classical geostatistical simulations, such as Boolean methods, Sequential indicator simulation (SIS) and truncated Gaussian simulation (TGS) or Plurigaussian simulation (PGS), then they are populated with heterogeneous hydraulic and transport parameters.

However, the solution of the inverse problem is considered the most efficient practice to model the structure of hydrofacies while the physical law governing the groundwater flow system is taken into account. The parameters of the physical law, such as hydrofacies distribution, are defined by optimizing the response of the system while solving the physical law (forward problem), compared to the observations of a physical variable such as the hydraulic head. In most cases, the inverse problems are ill-posed, which means they have not a unique solution, the solutions do not depend continuously on data, or a solution does not exist. Moreover, the solutions of the problem may be affected significantly by small changes in the observations and cause large variations in the parameters estimation, making the system ill-conditioned. So, the limited number of observations and the nature of system may impose difficulties in solving an inverse problem.

Thus, the inverse problem is a classical optimization problem when the system of equations is linear, where the best-unbiased estimator can be found by the minimization of a least squares error criterion. In ill-conditioned or ill-posed problems, the objective is not only to find an unbiased estimator, but also one with a stable behavior. In weakly linear or non-linear systems, the objective function is linearized and iterative gradient methods are applied to find the estimator. However, the iterative process of optimization in those cases is often a demanding task with high computational cost, while gradient methods can get stuck at local

optima. Therefore, the use of stochastic methods allowing the sampling of model space, such as the Markov chain Monte Carlo (McMC) methods, is often preferable.

McMC methods are often used under the Bayesian perspective in solving the inverse problem. In Bayesian inference, the prior information is updated iteratively as new members are added to the chain. The members of the chain constitute the final a-posteriori distribution of the parameters. The McMC instead of Monte Carlo methods preferentially visit the model space where the posterior density is high, even if the dimension of the model space is large. The implementation of McMC requires the definition of a transition kernel from an accepted model to a new model, a criterion to accept a member in the chain and a criterion to interrupt the chain.

In this work, we adopt the iterative spatial resampling (ISR) technique as the transition kernel, either when the objective is to sample the posterior distribution of parameters or to reach an optimal solution. In sampling the posterior, the candidate models must be independent of the last accepted member of the chain, while the criterion of Metropolis-Hastings or Mosegaard and Tarantola (1995) could be used to accept a model to the chain. In the case of optimization, we adopt the implementation of independent Markov chains to reach an optimal solution. A candidate model is accepted in a chain when its likelihood is better than the last member of the chain, while the interruption of the chain is stochastic with the probability to interrupt becoming higher when the likelihood of a candidate model is high. This way, bias with the opposite direction is introduced to the samples of the posterior. Thus, an approach to posterior distribution is achieved. This mechanism reaches an optimal solution avoiding a large number of geostatistical simulations and forward problem solutions.

The novelty of the C-ISR algorithm is the iterative use of cosimulation between hydrofacies and the reference data as an auxiliary variable, in order to gradually improve the path to the optimal solution within a constantly improving Markov chain. Cosimulation for modeling a discrete variable such as the hydrofacies distribution has not been applied in inversion yet, due to the nonlinearity relation between the response variable and unknown parameters. In most cases, the response variable is used as indirect data to evaluate the prior models and drive the search path. The C-ISR method exploits the available information on the relationship between hydrofacies and the physical variable, to produce valid realizations by using cosimulation. More specifically, our method is based on the approach that,

instantaneously, the Normal scores transformed hydrological measurements can be correlated with the Gaussian variable of hydrofacies, through a linear coregionalization model, although the underground flow problem is not linear. The approach is used repeatedly within a Markov chain, while the members of the chain result from an iterative spatial resampling transition kernel. In this case, apart from their indirect use for inversion, groundwater pressure measurements are used directly, in order to evaluate the prior geological model of the subsurface. This, results in a narrower and more informed prior distribution, due to the support of the reference variable. The effectiveness of our method is demonstrated by an example application on a synthetic underdetermined inverse problem in aquifer characterization. The results show that the C-ISR method is faster and more accurate as compared to plain ISR.

Περίληψη

Στην παρούσα διδακτορική διατριβή αναπτύσσουμε και παρουσιάζουμε τη μέθοδο Cosimulated Iterative Spatial Resampling (C-ISR) για την στοχαστική επίλυση του αντίστροφου προβλήματος το οποίο σχετίζεται με τον χαρακτηρισμό των υδρογεωλογικών φάσεων του υπεδάφους και θέτουμε το θεωρητικό και μαθηματικό υπόβαθρο της μεθόδου.

Η ανάπτυξη της μεθόδου βασίστηκε στην ανάγκη του χαρακτηρισμού των υδρογεωλογικών φάσεων του υπεδάφους πριν από τον προσδιορισμό/εκτίμηση άλλων υδρογεωλογικών παραμέτρων, όπως της υδραυλικής αγωγιμότητας, του πορώδους και της ζώνης κορεσμού. Η πολυπλοκότητα της γεωλογικής διαδικασίας, οι φυσικές και χημικές αντιδράσεις, όπως επίσης η δυσκολία αναπαράστασης της υδραυλικής αγωγιμότητας ως συνεχούς μεταβλητής (Matheron, 1967; Emsellem and De Marsily, 1971), καθιστούν απαραίτητη τη μοντελοποίηση των υδρογεωλογικών φάσεων εκ των προτέρων. Επιπλέον, ο χαρακτηρισμός των πετρωμάτων είναι πάγια πρακτική στις γεωλογικές μελέτες σε σχέση με άλλες παραμέτρους όπως η αγωγιμότητα και αυτό οδηγεί σε πιο ακριβή προσδιορισμό της χωρικής κατανομής των φάσεων, αφού υπάρχει μεγαλύτερη διαθεσιμότητα δεδομένων.

Η χωρική κατανομή των υδρογεωλογικών φάσεων μπορεί να εκτιμηθεί με κλασικές μεθόδους γεωστατιστικής προσομοίωσης όπως οι Boolean μέθοδοι, η Sequential indicator simulation (SIS) και η truncated Gaussian simulation (TGS) ή Plurigaussian simulation (PGS). Στη συνέχεια άλλες υδρογεωλογικές παράμετροι μπορούν να προσδιοριστούν. Η προσομοίωση της χωρικής κατανομής των υδρογεωλογικών φάσεων μπορεί να γίνει πιο ακριβής, λαμβάνοντας υπόψη βοηθητικές μεταβλητές μέσω από κοινού προσομοίωσης, όπως για παράδειγμα την υδραυλική πίεση, καθώς συχνά υπάρχουν διαθέσιμα δεδομένα της πίεσης σε υδρογεωλογικές λεκάνες ενδιαφέροντος. Ωστόσο, η από κοινού προσομοίωση απαιτεί τη γνώση της συσχέτισης μεταξύ των φάσεων και της βοηθητικής μεταβλητής, ενώ δεν αξιολογείται η πιθανοφάνεια των παραμέτρων, καταλήγοντας έτσι σε μη αποδεκτές λύσεις. Κατά κανόνα, παρά το πλήθος των εκάστοτε διαθέσιμων δειγμάτων από γεωτρήσεις, ο αριθμός αυτός δεν είναι ποτέ αρκετός για την επίτευξη της επιθυμητής ακρίβειας στην απεικόνιση του υπεδάφους.

Η πιο σύγχρονη πρακτική αντιμετώπιση του προβλήματος αυτού, είναι η επίλυση του φυσικού νόμου που διέπει το φαινόμενο μεταφοράς ροής (ευθύ πρόβλημα) υποθέτοντας ότι

οι παράμετροι του συστήματος είναι γνωστές, ενώ στη συνέχεια γίνεται η ρύθμισή τους. Η πρακτική αυτή είναι γνωστή ως επίλυση του Αντίστροφου Προβλήματος, καθώς οι υδρογεωλογικές φάσεις αντιμετωπίζονται πλέον ως παράμετροι ενός συστήματος και η φυσική μεταβλητή (π.χ. υδραυλική πίεση) ως τυχαία μεταβλητή. Στο αντίστροφο πρόβλημα, οι μετρήσεις της φυσικής μεταβλητής και η απόκριση του φυσικού νόμου χρησιμοποιούνται για την κατανόηση της συμπεριφοράς των παραμέτρων και του προσδιορισμού τους.

Αν και η πρακτική του προσδιορισμού των φάσεων είναι πιο αποτελεσματική επιλύοντας το αντίστροφο πρόβλημα καθώς χρησιμοποιείται μία επιπλέον πληροφορία, δηλαδή ο φυσικός νόμος, μία σειρά από άλλα προβλήματα δημιουργούνται στα οποία οι ερευνητές καλούνται να δώσουν απαντήσεις. Πιο συγκεκριμένα, τα αντίστροφα προβλήματα είναι συνήθως ασθενώς/κακώς τεθειμένα (ill-posed), δηλαδή επιδέχονται από καμία λύση έως άπειρες λύσεις. Επίσης, τα αντίστροφα προβλήματα μπορεί να είναι μη επιλύσιμα εξαιτίας της υπολογιστικής ακρίβειας και τότε χαρακτηρίζονται ως ασταθή συστήματα (ill-conditioned systems). Τα προβλήματα αυτά δημιουργούνται κυρίως λόγω του περιορισμένου αριθμού δειγμάτων της φυσικής μεταβλητής αλλά και της φύσης του συστήματος.

Στην περίπτωση που το σύστημα εξισώσεων είναι γραμμικό, το αντίστροφο πρόβλημα είναι ένα κλασσικό πρόβλημα βελτιστοποίησης, στο οποίο ο βέλτιστος αμερόληπτος εκτιμητής των παραμέτρων μπορεί να δοθεί από το κριτήριο των ελαχίστων τετραγώνων. Στην περίπτωση γραμμικών ασθενών ή ασταθών συστημάτων, πέραν από την αναζήτηση ενός αμερόληπτου εκτιμητή, αντικείμενο είναι η εύρεση ενός εκτιμητή ο οποίος μπορεί να είναι μεροληπτικός αλλά να δίνει πιο δυνατές ή σταθερές λύσεις στο σύστημα των εξισώσεων. Στην περίπτωση (ασθενώς) μη γραμμικών συστημάτων, η αντικειμενική συνάρτηση γραμμικοποιείται και η διαδικασία εύρεσης του βέλτιστου εκτιμητή γίνεται επαναληπτικά με τους αλγόριθμους κλίσης.

Η επαναληπτική διαδικασία εύρεσης τους βέλτιστου εκτιμητή στις περιπτώσεις ασθενώς μη γραμμικών συστημάτων ή τελείως μη γραμμικών συστημάτων συχνά είναι μία επίπονη διαδικασία με μεγάλο υπολογιστικό κόστος (μεγάλος αριθμός παραμέτρων), ενώ είναι πολύ πιθανό το αποτέλεσμα του βέλτιστου εκτιμητή να αναφέρεται σε ένα τοπικό βέλτιστο της αντικειμενικής συνάρτησης. Για τον λόγο αυτό η χρήση στοχαστικών μεθόδων που επιτρέπουν το συστηματικό δειγματισμό του πεδίου τιμών των παραμέτρων, όπως οι

μέθοδοι Markov chain Monte Carlo, είναι προτιμότερη, ξεπερνώντας τα προβλήματα των αιτιοκρατικών ή άλλων στοχαστικών μεθόδων.

Στην αντιστροφή του προβλήματος οι μέθοδοι McMC χρησιμοποιούνται συνήθως σε συνδυασμό με τη στατιστική κατά Bayes, δηλαδή η εκ προοιμίου (a-priori) κατανομή των παραμέτρων αναθεωρείται επαναληπτικά όσο νέα μέλη προστίθενται στην αλυσίδα. Τα μέλη της αλυσίδας αποτελούν δείγματα της μεταγενέστερης (a-posteriori) κατανομής των παραμέτρων. Οι μέθοδοι McMC έχουν την ικανότητα να «επισκέπτονται» το χώρο των παραμέτρων όπου η πυκνότητα της μεταγενέστερης κατανομής είναι μεγάλη ακόμα και όταν ο χώρος των παραμέτρων είναι πολλών διαστάσεων.

Η εφαρμογή των McMC απαιτεί τον ορισμό ενός πυρήνα μετάβασης από το ένα μέλος της αλυσίδας σε ένα υποψήφιο μέλος, ενός κριτηρίου επιλογής των μελών της αλυσίδας και ενός κριτηρίου για να διακοπεί η αλυσίδα. Ο καθορισμός των παραπάνω κριτηρίων πρέπει να είναι τέτοιος ώστε τα μέλη της αλυσίδας να είναι ανεξάρτητα μεταξύ τους και να αποτελούν δείγματα της a-posteriori κατανομής. Η διακοπή της αλυσίδας συνήθως εξαρτάται από τον επιθυμητό αριθμό των δειγμάτων και τον διαθέσιμο χρόνο για την υλοποίηση της αλυσίδας.

Στη παρούσα διατριβή, τα υποψήφια μέλη της αλυσίδας δημιουργούνται με γεωστατιστική προσομοίωση και με τυχαίο χωρικό δειγματισμό του αμέσως προηγούμενου μέλους της αλυσίδας και των a-priori δεδομένων, δηλαδή ένα προσαρμοσμένο αλγόριθμο Gibbs Sampling. Όταν ο στόχος της έρευνας είναι ο ορισμός της πρότερης κατανομής, τότε ένα υποψήφιο μέλος της αλυσίδας κρίνεται βάσει του κριτηρίου που χρησιμοποιείται στον αλγόριθμο Metropolis-Hastings ή του κριτηρίου που προτείνεται από τους Mosegaard and Tarantola (1995). Τα υποψήφια μέλη της αλυσίδας πρέπει να κρίνονται αφού έχει επιτευχθεί η ανεξαρτησία τους από το προηγούμενο μέλος της αλυσίδας.

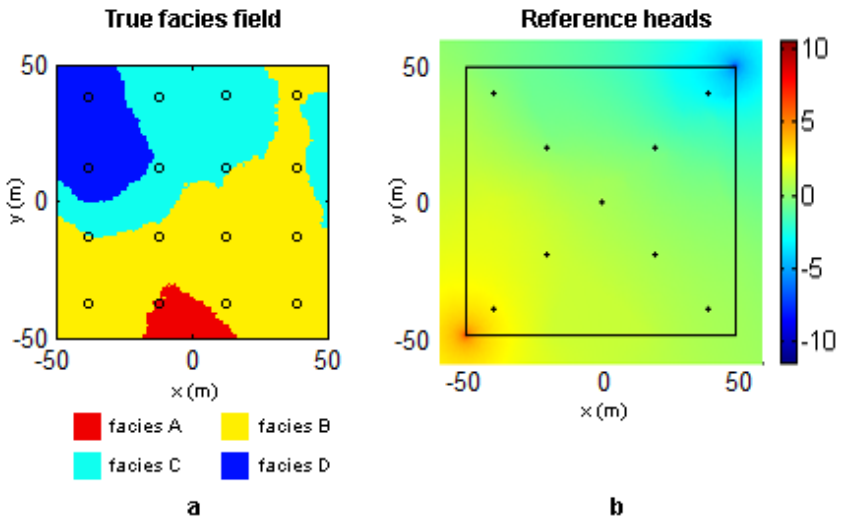
Στην περίπτωση που ο στόχος είναι η εύρεση ενός βέλτιστου εκτιμητή, υιοθετούμε την υλοποίηση ανεξάρτητων Μαρκοβιανών αλυσίδων, ενώ κάθε μέλος της αλυσίδας γίνεται αποδεκτό εφόσον η πιθανοφάνειά του είναι μεγαλύτερη από το προηγούμενο μέλος της αλυσίδας. Αυτό οδηγεί στην επιλογή μεροληπτικού δείγματος της ύστερης κατανομής, αφού τα τελευταία μέλη κάθε αλυσίδας που αποτελούν την κατανομή έχουν προκύψει με υψηλή πιθανοφάνεια. Η διακοπή κάθε αλυσίδας γίνεται με στοχαστικό κριτήριο, στο οποίο η πιθανότητα διακοπής της αλυσίδας αυξάνεται όσο η πιθανοφάνεια των μελών μίας αλυσίδας αυξάνεται. Με αυτόν τον τρόπο προκύπτει ένα μεροληπτικό δείγμα της a-posteriori κατανομής

αλλά με αντίθετη κατεύθυνση της μεροληψίας που εισάγεται με την επιλογή των μελών κάθε αλυσίδας. Έτσι, επιτυγχάνεται μία προσέγγιση της ύστερης κατανομής. Αυτός ο μηχανισμός οδηγεί στην εύρεση του βέλτιστου εκτιμητή αποφεύγοντας έναν μεγάλο αριθμό γεωστατιστικών υλοποιήσεων και της επίλυσης του ευθέως προβλήματος.

Σε προηγούμενες μελέτες οι οποίες χρησιμοποιούν τις μεθόδους McMC, δεν λαμβάνεται υπόψη η συσχέτιση της φυσικής μεταβλητής και των παραμέτρων για την παραγωγή των γεωστατιστικών υλοποιήσεων. Σε αυτή την περίπτωση, οι μετρήσεις της πίεσης του υπόγειου νερού χρησιμοποιούνται μόνον εμμέσως, για την αξιολόγηση των πρότερων γεωλογικών μοντέλων του υπεδάφους. Ο περιορισμός αυτός οφείλεται στην μη γραμμική συσχέτιση μεταξύ υδραυλικών μετρήσεων και παραμέτρων του υπεδάφους. Ως εκ τούτου, οι ευρέως γνωστές μέθοδοι των από κοινού kriging και από κοινού προσομοίωσης που βασίζονται στην γραμμική εκτίμηση, δεν μπορούν να χρησιμοποιηθούν απευθείας, γιατί καταλήγουν σε μη αποδεκτές λύσεις. Αντίθετα, η μέθοδος C-ISR εκμεταλλεύεται την πληροφορία των μετρήσεων της φυσικής μεταβλητής για την παραγωγή των γεωστατιστικών υλοποιήσεων.

Πιο συγκεκριμένα, η μέθοδος στηρίζεται στην προσέγγιση ότι, στιγμιαία, οι μετασχηματισμένες σε Normal Scores υδρογεωλογικές μετρήσεις, μπορούν να συσχετιστούν με την Γκαουσιανή μεταβλητή των υδρογεωλογικών φάσεων, μέσω γραμμικού μοντέλου συμμεταβλητότητας, παρότι το πρόβλημα της υπόγειας ροής δεν είναι γραμμικό. Η προσέγγιση αυτή γίνεται επαναληπτικά εντός μιας Μαρκοβιανής αλυσίδας, της οποίας τα μέλη προκύπτουν χρησιμοποιώντας ως πυρήνα μετάβασης έναν επαναληπτικό χωρικό δειγματισμό του προηγούμενου μέλους. Στην περίπτωση αυτή, οι μετρήσεις της πίεσης του υπόγειου νερού, εκτός από την έμμεση χρήση τους για την επίλυση του αντίστροφου προβλήματος, χρησιμοποιούνται και άμεσα, για την υλοποίηση των πρότερων γεωλογικών μοντέλων του υπεδάφους. Με τον τρόπο αυτό προκύπτει μια στενότερη και πιο ενημερωμένη πρότερη κατανομή, λόγω της συνδρομής της μεταβλητή αναφοράς.

Η αποτελεσματικότητα της μεθόδου αποδεικνύεται από την εφαρμογή της σε ένα συνθετικό παράδειγμα υπο-ορισμένου αντίστροφου προβλήματος για τον χαρακτηρισμό ενός υδροφόρου ορίζοντα (Εικόνα 0.1).



Εικόνα 0.1: Η πραγματική κατανομή των φάσεων στον υδροφόρο ορίζοντα (αριστερά). Οι τιμές αναφοράς και οι μετρήσεις της πίεσης (δεξιά).

Η μέθοδος C-ISR συγκρίνεται με την ISR μέθοδο χωρίς τη χρήση της από κοινού κατανομής των υδρογεωλογικών φάσεων και της πίεσης, ενώ για τις υλοποιήσεις χρησιμοποιείται η Plurigaussian προσομοίωση. Από τα αποτελέσματα προκύπτει ότι η C-ISR είναι πιο γρήγορη μέθοδος καθώς από τις 250 ανεξάρτητες αλυσίδες Markov που δημιουργήθηκαν, το ευθύ πρόβλημα επιλύθηκε 28914 φορές έναντι 48390 στην περίπτωση της ISR (Εικόνα 0.2).

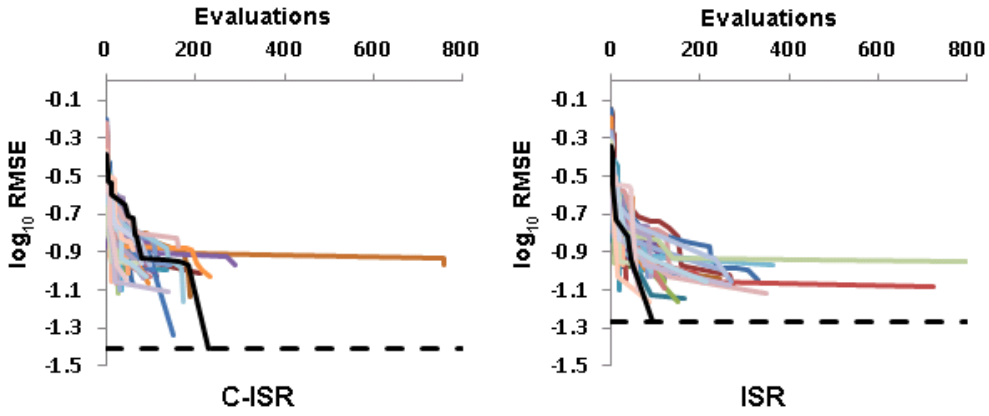


Figure 0.2: Η σύγκλιση 50 τυχαίων ανεξάρτητων αλυσίδων των μεθόδων C-ISR και ISR.

Επιπλέον, η C-ISR κατέληξε σε πιο ακριβείς προβλέψεις έχοντας μέση τετραγωνική ρίζα σφάλματος (RMSE) 0,0383 έναντι 0,0537 για την ISR (Εικόνα 0.3).

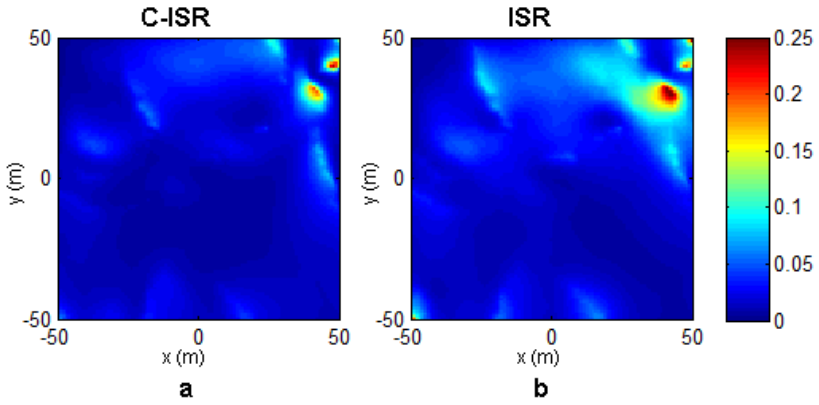


Figure 0.3: Ο χάρτης του RMSE από 250 ανεξάρτητες αλυσίδες της C-ISR (a) και ISR (b).

Επίσης, η βέλτιστη λύση της C-ISR έφτασε σε 85,69% ποσοστό ομοιότητας των υδρογεωλογικών φάσεων με την πραγματικότητα έναντι του 81,76% ομοιότητας της ISR (Εικόνα 0.4). Τα αντίστοιχα μέσα ποσοστά ομοιότητας από τις 250 αλυσίδες είναι 80,82% έναντι 71,04%.

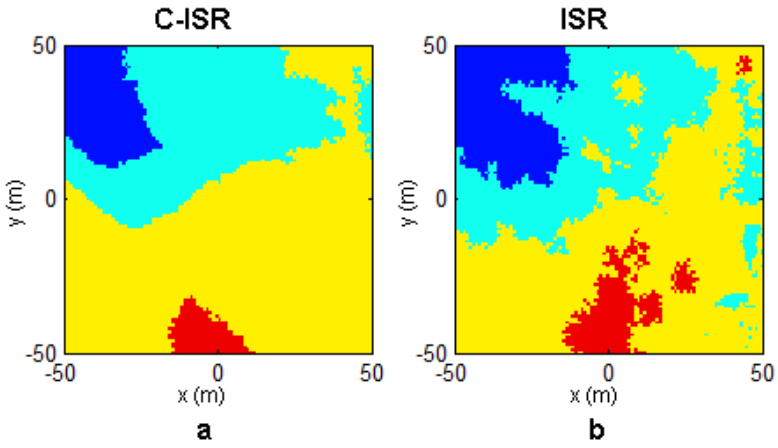


Figure 0.4: Ο βέλτιστος εκτιμητής των φάσεων της C-ISR (a) και ISR (b) μετά από 250 ανεξάρτητες αλυσίδες.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Professor K. Modis for the encouragement, the continuous support, the advices and the time he dedicated to this research. I would like to note that the research would be impossible without his experience, immense knowledge, patience, motivation and insist.

Besides my advisor, I would like to thank Associate Professor G. Apostolopoulos, for his insightful comments.

In addition, a thank to Professor A. Boudouvis, who introduced me to the finite element method for solving partial differential equations in his class of School of Chemical Engineering at NTUA.

Finally, I am grateful to the anonymous reviewers of Advances in Water Resources journal who made critical remarks and detailed comments in the submitted research and gave me and my co-author K. Modis the opportunity to substantially improve it.

Contents

List of Figures

1 Introduction

1.1 Statement of the Problem

Numerical modeling of groundwater flow and mass transport are important tools for predicting the behavior of a hydrogeological system. Nevertheless, in order to produce reliable hydrologic predictions, parameter values that determine the response of the model must be appropriately chosen for a specific aquifer. Direct measurements of hydrologic parameters, however, are scarce and fraught with uncertainty. The scarcity of direct measurements of the hydrologic parameters has been long ago recognized as a major impediment to the use of groundwater models and to their full utilization (Frind and Pinder, 1973).

To address this problem of parameter uncertainty, hydrologic models can be used in applications “opposite” or “inverse” to their original use, i.e., parameter values are treated as system unknowns and are determined by extracting information from observations of system-response variables (Kitanidis and Vomvoris, 1983). The procedure is conveniently called inverse problem-solving. However, the subsurface reservoir is normally very heterogeneous due to complex geologic processes and physical and chemical reactions, which makes model parameter identification a demanding task. On the other hand, a crucial issue is that the problem of specifying conductivity in every block from sparse head observations is underdetermined, i.e., there are many solutions that are consistent with the data. The ambiguity is largely due to the scarcity of the data but is also inherent in the mathematics of typical inverse problems: a small range of values in the observed head is consistent with a larger range of conductivity values (Kitanidis, 2007). This characteristic is known as ill-posedness and results in non-uniqueness in the solution of the inverse problem (see also section 3.23).

Numerous deterministic and stochastic methods have been proposed to overcome the ill-posedness of the inverse problem and specify the hydraulic conductivity in an aquifer field. Over the recent years, researchers have developed novel approaches to solve the problem by inverting directly for the spatial distribution of hydrofacies, instead of conductivity. Admittedly, specifying the hydrofacies distribution is a more efficient and direct way to predict the behavior

of a hydrogeological system (see literature review). But, there are problems with this approach, since facies is a categorical variable and is a challenge to incorporate it in the geostatistical formalism. And, apart from all complications, the “curse of dimensionality” adds in computer time and complexity in general.

Taking all the above in account, the present work lies in the McMC category of methods and attempts to solve the facies inversion problem by selectively sampling a chain of successive models driven by cosimulation with the response variable.

1.2 Objectives

The main goal of this research is to develop a method for direct inversion of the spatial distribution of hydrofacies, using geostatistical information such their covariance, the proportions and their contacts and to deal with the theoretical and practical difficulties arising in this approach in the non-linear case. For this scope, we use the Plurigaussian simulation, one of the recently developed facies modeling methods in geostatistics.

Our original idea is to exploit the information resulting from the response of the hydrological model not only as likelihood function, as the most researches in Bayesian perspective propose, but also as a linear approximation of the association between the state variables. This association can be expressed by their experimental cross-variogram and it is embedded in the prior ensemble by cosimulation with the reference variables.

An objective of this study is to build the theoretical background behind the proposed method and explain the effectiveness of cosimulation, which is the basic characteristic of our method, by comparing it to a similar procedure without cosimulation.

1.3 Innovation

Cosimulation for modeling a discrete variable such as the facies distribution has not been applied in inversion yet, due to the nonlinearity relation between the response variable and unknown parameters. In most cases, the response variable is used as indirect data to evaluate the prior models and drive the search path.

The cosimulation of facies and the hydrological response variable using iterative spatial resampling as transition kernel is used in hydrology for the first time. Our method creates an important sampling effect that steers the process to selected areas of the prior facies distribution. The results demonstrate the effectiveness of our method producing a better approximation of the facies spatial variability and, at the same time, less computation cost than other methods.

The straightforwardness of implementation, the simplicity of tuning the initial parameter and the requirement of less amount of information (hard data), in combination with the accuracy, the computational cost in large scale-problems and the realistic facies reproduction that honors production data, make C-ISR perform better than other methods.

1.4 Structure

This work contains six chapters in total. Chapter 1 introduces the reader to the general inverse problem in groundwater modeling and analyzes the objectives and the innovation of this research as previously stated. In the next section, the fundamental topics of inverse problem are presented followed by an extended literature review, while the developments of previous works that guided our research are finally presented. Chapter 2 explains the most common methods in geostatistical simulation for discrete variables and presents the theoretical basis of Plurigaussian simulation in detail. Moreover, the basic outlines of multivariate analysis in geostatistics, such as cokriging, the linear coregionalization model, and Cosimulation are given. Chapter 3 focuses on the general discrete inverse problem using the least square criterion to solve underdetermined and overdetermined optimization problems, while the regularized estimator in ill-conditioned systems is explained**.** Furthermore, the theory of the transition from the deterministic to the stochastic solution of the inverse problem and the solution of the non-linear problem with deterministic iterative algorithms that use the least square criterion as the objective function is presented. In Chapter 4, the Bayesian formulation of linear and non-linear inverse problems is established and the Monte Carlo and McMC methods to sample the posterior distribution are presented. The C-ISR method (Valakas and Modis 2015; Valakas and Modis 2016), which is the main achievement of this work, is presented in Chapter 5, explaining the outline, the kernel transition, the role of cosimulation and the detailed algorithm of the method. The effectiveness of the proposed approach is demonstrated by an example

application on a synthetic underdetermined inverse problem in aquifer characterization. Finally, the conclusions and contribution of the study are summarized in Chapter6. The last chapter details the progress achieved towards the objectives and suggests plans for future research.

1.5 Literature Review

1.5.1 Inverse Problem Approaches in Groundwater Modeling

In recent decades, numerous methodologies have been proposed to solve the inverse problem in groundwater modeling, with the general aim to estimate the spatial distribution of hydraulic conductivity in an aquifer field. A review of the most recently proposed methods can be found in Carrera et al. (2005). Because of the inherent difficulties associated with the estimation of spatial functions from limited and imperfect data, the crux of the problem is how to parameterize the distributed parameter system (Emsellem and De Marsily, 1971; Neuman, 1973; Gavalas et al., 1976), that is, how to transform the question into a well-posed estimation problem. In general, the principal conceptual differences among available methods lie in the parameterization (Kitanidis, 1995).

Certain formalisms (e.g., Tikhonov and Arsenin, 1977; Guidici et al., 1995; Tikhonov et al., 1997) focus on finding a single solution to the inverse problem by minimizing the error between model results and measurements. This is done by imposing restrictions, or by making assumptions about available information, such as that head is measured without error at every node of the model. Such methods have appeal in practice only when sufficient information is truly available (Kitanidis, 2007).

Stochastic methods, on the other hand, use statistical conditioning in which the covariance between parameters and system-response variables is utilized to condition the parameter values, using measurements information. By representing the spatial parameter function with a random field, the idea is that although the exact value is not known, one should be able to identify an interval that contains the true value with a high degree of assurance. In general, stochastic methods do not adopt the classic statistical but rather the Bayesian view, where the probabilities represent a state of knowledge or available information (Christakos, 1990). The idea is that the unknown function, such as the conductivity over a region, is

modeled as a random function (**RF**) mainly because there is insufficient information to model it as deterministic, rather than because repeated measurements indicate a statistical regularity.

Many stochastic parameter estimation methods employ geostatistical models to define the spatial distribution (structure) of conductivity (e.g., Kitanidis and Vomvoris, 1983; Hantush and Marino, 1997; Chen and Zhang, 2006; Hendricks Franssen and Kinzelbach, 2008), under the assumption that aquifer conductivity in regional systems can generally be described using such models (Kitanidis and Vomvoris, 1983; Hoeksema and Kitanidis, 1985; Carrera et al., 2005). There are difficulties, anyhow, in the representation of conductivity as a continuous variable. Matheron (1967) had already pointed out that conductivity is not a point but a set function, while Emsellem and De Marsily (1971) concluded that conductivity is a parameter with no punctual value but with an average in a region of a given size. In this line, the usual continuous and multi-Gaussian methods have shown that they do not allow modeling a sufficiently wide range of connectivity patterns for the high (or low) permeable structures (Journel and Alabert 1990; Zinn and Harvey 2003; Renard et al., 2005; Kerrou et al., 2008; Renard 2007). To overcome this problem, Fienen et al. (2008) propose an interactive zonation method, where candidate zones are implied from the data and are evaluated using cross-validation and expert knowledge.

An alternative is to use a two-step approach in which, first, the geological facies are modeled, and second, they are populated with heterogeneous hydraulic and transport parameters. This approach is flexible and allows modeling structures at different scales (Mariethoz et al., 2009). Therefore, before the geostatistical estimation of a soil property such as conductivity, the knowledge of geological formation should take into account, due to the complexity of soil types as natural entities (Ibáñez and Saldaña, 2008; Modis and Sideri, 2013).

In general, deterministic models for structure identification, such as graphical methods (Doveton, 1986), neural networks methods (Rogers et al., 1992) and fuzzy neural networks methods (Chang et al., 1997), have limitations and may introduce larger bias and uncertainty than an inappropriate choice of facies hydraulic parameters (Ye et al., 2004; Lu and Robinson, 2006). On the other hand, stochastic models have been found to be effective in overcoming the above problems. Following this approach, Winter and Tartakovsky (2000 and 2002) provide a general framework for modeling flow and transport in high heterogeneous porous media consisting of multiple materials, by quantifying uncertainty in both spatial arrangement of

geological facies and hydraulic properties within each facies. Guadagnini et al. (2004) utilize stratigraphic and sedimentological data to reconstruct the spatial extent of the aquitard zone applying indicator geostatistical techniques, in order to solve the equations for ensemble moments of hydraulic head. Mariethoz et al. (2009) assess contaminant migration, applying the truncated plurigaussian method that allows integrating a geological conceptual model (using a lithotype rule) within the framework of a mathematically consistent stochastic model.

Instead of inverting for conductivity, a more efficient approach is to solve the inverse problem directly on the hydrofacies distribution. This tactic allows for simultaneous estimation of the optimum aquifer structure from geological or geophysical data and hydrologic response measurements in a single step. In this line, Chen and Rubin (2003) propose a Bayesian model for lithofacies estimation, assimilating geophysical data through a likelihood function. Liu (2005) generates geological facies maps using Plurigaussian simulation in order to solve the problem of automatic history matching of facies and explore the gradient method and the ensemble Kalman filter (**EnKF**) method, after the parameterization of the geostatistical model. Harp et al. (2008) follow a genetic meta-algorithm to select between equiprobable structure realizations according to a fitness criterion, while Mariethoz et al. (2010a) employ an iterative spatial resampling procedure to steer the search for solutions among simulated images that preserve the same spatial structure. Furthermore, Cardiff and Kitanidis (2009) developed a flexible, yet sensitive to the initial guess framework for defining zone boundaries using the level set method.

1.5.2 Bayesian Perspective and McMC inversion

In a Bayesian perspective, the terminology is different. Ill-posedness is not defined anymore and the question can be phrased as whether the likelihood of the data is sufficient or not to constrain the posterior distribution to be different from the prior. Thus, the solution to a Bayesian inverse problem is an a posteriori probability density. Using sampling methods, the goal is to generate a sample from the posterior, such that statistics consistent with the posterior distribution can be inferred.

Yet, in addition to the classical optimization methods, a possible approach for the solution to inverse problems is the use of Markov chain Monte Carlo (**McMC**) techniques (Zhou et al., 2014). This is an alternative way to achieve the same results without resorting to a

deterministic optimization setup, but rather sampling a multivariate probability distribution that converges to the posterior. McMC methods generate model realizations that match the state observations while reproducing some prior statistics and obey Bayes' rule. It is noted that these requirements are only partly fulfilled by most gradient-based optimization techniques (Gomez- Hernandez et al., 1997; Mariethoz et al., 2010a). Thus, McMC techniques, which are mainly prior driven approaches, do not produce models with zero prior probability.

Howbeit, the use of informative priors in a Bayesian context has been a crucial issue of disagreement (Jaynes, 1985; Scales and Sneider, 1997; Mosegaard, 2011). Some authors suggest that the use of prior information may bias the solution of an inverse problem in an unwanted way. On the other hand, non-informative priors may lead to unsolvable problems or to solutions with high uncertainty (Hansen et al., 2012). Different a priori priors, that is, probability distributions that are somehow justified by the nature of the uncertainty of a situation, are many times found. But this subject is a matter of philosophical controversy, with the Bayesians divided into two schools: "objective Bayesians," who believe that such priors are justified in many useful cases and "subjective Bayesians" who believe that in practice the priors usually represent subjective judgments that cannot be justified (Williamson 2010). Objective prior means a state of knowledge coming from a logical inference and the strongest arguments for objective Bayesianism were given by Jaynes (2003) based on the principle of maximum entropy, which says that a belief function should be a probability function, from all those that are calibrated to evidence.

2 Geostatistical Simulation of Categorical Variables

2.1 Introduction

The solution of a physical system (e.g. groundwater flow transport) requires detailed knowledge of the physical parameter spatial distribution (e.g. hydraulic conductivity). The spatial distribution of the parameter may be estimated by interpolation methods, such as kriging. The stochastic solution of a physical system requires more than one realization of the spatial parameter (random field) to incorporate its structural uncertainty, which is reflected in its response. Such realizations of the parameter set can be generated by a geostatistical simulation. The parameters may concern either continuous variables or categorical variables, for example, hydraulic conductivity and rock type respectively. General, the continuous variables have an infinite number of values, while categorical variables contain a finite number of categories or distinct groups.

Various approaches of geostatistical simulation have been developed to generate these realizations. Approaches to simulate continuous variables include the turning bands method (Journel, 1974), spectral methods (Borgrnan et al., 1984; Gutjahr, 1989), lower upper (**LU**) or Cholesky decomposition (Alabert, 1987; Davis, 1987), sequential Gaussian simulation (Journel, 1989), and fractal approaches (Hewett and Behrens, 1990). For categorical variables, the approaches include Boolean methods (Serra, 1982; Jeulin, 1987; Chautru, 1989), Sequential indicator simulation (Journel and Alabert, 1989) and truncated Gaussian simulation or Plurigaussian simulation (Matheron et al. 1987; Galli et al. 1994; Amstrong, 2011), while Genetic algorithms (Whitley, 1994) and simulated annealing (Kirkpatrick et al., 1983; Deutsch and Journel, 1992) are used for both continuous and categorical variables.

The implementation of a geostatistical simulation requires specification of a multivariate probability model for the spatial process. Then, the algorithm produces equiprobable (equally likely to be drawn) realizations of a random field in order to capture the attributes of the phenomenon (Deutsch and Journel, 1997; Chiles and Delfiner, 1999; Modis and Sideri, 2013). However, the probability density values of the realizations are different in general.

In the next sections, we present the sequential simulation process and we focus on Sequential indicator simulation (**SIS**), and Plurigaussian simulation (**PGS**), which are widely used to model discrete or categorical variables.

2.2 Sequential Conditional Simulation

The aim of sequential conditional simulation is to produce realizations that honor the sample data as well as the spatial covariance of the attributes being simulated. Thus, in order to reproduce a realization, the simulated values are conditioned to the sample data along with the previous simulated values. The sequential process requires the definition of a random path by which all query points are visited sequentially. Thus, the realizations are different due to the path randomness and the conditioned simulated values.

Let $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_\nu$ be the random variables of interest at the locations $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\nu$, where *v* is the query points, then the *v*-dimensional probability density function (**pdf**) of a RF $f_\nu(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_\nu)$ can be decomposed as the product of one dimensional marginal and a series of one dimensional conditional densities:

$$f_\nu(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_\nu) = f_1(\mathbf{z}_1) f_{1|1}(\mathbf{z}_2|\mathbf{z}_1) f_{1|2}(\mathbf{z}_3|\mathbf{z}_1, \mathbf{z}_2) \cdots f_{1|\nu-1}(\mathbf{z}_\nu|\mathbf{z}_1, \dots, \mathbf{z}_{\nu-1}) \quad (2.1)$$

After generating the first simulated value $\mathbf{z}_1$ the next simulated value $\mathbf{z}_2$ will be generated by the conditional distribution $f_{1|1}(\mathbf{z}_2|\mathbf{z}_1)$. The process continues until one generates a value from $f_{1|\nu-1}(\mathbf{z}_\nu|\mathbf{z}_1, \dots, \mathbf{z}_{\nu-1})$. At the end of the process, the simulated values in the realization are a sample of $f_\nu(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_\nu)$. The above mathematical formulation is referred to as sequential simulation process while in sequential conditional simulation, the sample data are also taken into account by conditional distributions.

2.3 Sequential Indicator Simulation

The concept of SIS is based on indicator kriging, an algorithm which is widely used to transform a continuous random field $\mathbf{z}(\mathbf{x})$ to a binominal random function with values 0 and 1 in order to produce probability and risk maps (Journel, 1983). The coding of a datum into either 1 or 0 depends upon its relationship to a cut-off value, $\mathbf{z}_\mathrm{k}$. For a given value $\mathbf{z}(\mathbf{x})$

$$\mathbf{I}(\mathbf{x}, \mathbf{z}_k) = \begin{cases} 1 & \text{if } \mathbf{z}(\mathbf{x}) > \mathbf{z}_k \\ 0 & \text{if } \mathbf{z}(\mathbf{x}) < \mathbf{z}_k \end{cases} \quad (2.2)$$

where $\mathbf{x}$ are the spatial coordinates of the random variable. After this non-linear transformation, an indicator variogram can be modeled. The multivariate indicator kriging involves calculating and modeling multiple indicator variograms at a set of k thresholds covering the whole range of $\mathbf{z}(\mathbf{x})$. Multivariate indicator kriging is used in the same manner for a categorical variable such as geological facies, in order to produce estimation maps of this variable (e.g. each facies is transformed into 0 or 1 at every sample point, then the indicator variogram is calculated and modeled for each facies). After the modeling of variograms and cross-variograms, indicator kriging proceeds in the same manner as ordinary kriging: the results of coordinate queries (unsampled points) are obtained by kriging the indicator variables. The values are generally between 0 and 1 representing the probability of occurrence of a specific category at the particular coordinates. By the definition of probability, the acquired values must lie in the interval [0,1]. This relation may not be satisfied because the kriging estimate is a non-convex linear combination of conditioned data. Goovaerts (1997, pp. 399-400) lists the ways to avoid the violation and possible corrections to be applied to overcome this problem.

Implementation of the previous process under the (multivariate) indicator kriging modeling is referred to as SIS. The steps of SIS algorithm can be described as follows:

1. Chose a random location and perform indicator kriging for each category: Taking into account the conditioned data (sample data for the first query point), apply indicator kriging to estimate k probabilities of occurrence.
2. Correct these probabilities to sum to 100%.
3. For each category draw a simulated value from the Gaussian distribution with mean and variance the respective values from indicator kriging.
4. Define a cumulative distribution of categories using any ordering of the k categories. Then, k intervals are drawn and the probability of category occurrence lies in these intervals. For example, $[0, \mathrm{P}_1], (\mathrm{P}_1, \mathrm{P}_1 + \mathrm{P}_2], \dots, (\mathrm{P}_1 + \mathrm{P}_2 + \cdots + \mathrm{P}_k, 1]$, where the index of probability identifies the category.
5. Draw a random value P_u from the uniform distribution $\boldsymbol{U}(0,1)$. The interval in which P_u falls determines the simulated category.

6. Add the simulated value to the conditioning data and repeat the step one until all query points are simulated. When all query points are simulated, a full realization is reproduced. Multiple realizations are reproduced by repeating the above steps and defining different random paths.

2.4 Truncated Gaussian and Plurigaussian Simulation

2.4.1 The idea of Plurigaussian Simulation

Over the recent years, TGS is increasingly used in geological modeling where categorical variables such as rock facies occur. A sequential ordering of facies allows the researchers to define a simple lithotype rule. For example, the sand is followed by shaly sandstone then shale. Plurigaussian is simply an extension of TGS and allows for more complicated types of contacts between facies in order to produce a much wider range of geological patterns. TGS or PGS is an alternative method to SIS, in order to overcome the lack of geological realism and incorporate geological knowledge on the simulated field.

The idea behind the TGS is to represent the facies using a secondary continuous Gaussian RF at every point of the research area and then convert it to facies using the lithotype rule. The proportions of facies are used as prior information to define the thresholds of that Gaussian distribution. The thresholds can be calculated by the space facies proportions occupy in the distribution. In the case of PGS, more than one independent Gaussian distributions are defined either to represent more complicated contact relationships between facies or to identify an anisotropic behavior of facies contacts in space.

2.4.2 Theoretical formulation

Consider a standard Gaussian RF $\mathbf{Z}(\mathbf{x})$ where $\boldsymbol{x} \in \boldsymbol{R^3}$, with variogram $\gamma(\mathbf{h})$. Let $(D_1, \dots, D_k)$ be a partition of $\boldsymbol{R}$ into k disjoint subdomains. A categorical random field with k categories (facies) is obtained by putting

$$\forall\, \mathbf{x} \in \mathbf{R^3}, \mathbf{I}(\mathbf{x}) = i, \text{if and only if } \mathbf{Z}(\mathbf{x}) \in D_i \tag{2.3}$$

while the indicator random field for each facies F_i is defined as:

$$\forall\, \mathbf{x} \in \mathbf{R^3}, \mathbf{I}_{F_i}(\mathbf{x}) = \begin{cases} 1 \text{ if } \mathbf{I}(\mathbf{x}) = i \\ 0 \text{ otherwise} \end{cases} \tag{2.4}$$

The choice of the partition $(D_1, \dots, D_k)$ has implications on the spatial relationships between the facies, as it defines the permissible contacts between pairs of facies and also allows reproducing the chronological ordering of the facies (Armstrong et al., 2011). Practically, in order to transform between the Gaussian RF and the facies indicators, a proper set of $k-1$ truncation thresholds t_i has to be defined so as:

$$\mathbf{I}_{F_i}(\mathbf{x}) = 1 \Leftrightarrow t_{i-1} \leq \mathbf{Z}(\mathbf{x}) \leq t_i \tag{2.5}$$

The thresholds are in increasing order $t_1 \leq t_2 \leq \cdots \leq t_{i-1} \leq t_i \leq t_{i+1} \leq \cdots \leq t_{k-1}$.

Approximating the proportion of a particular facies F_i at point **x** by the probability of having this facies F_i at that point **x**:

$$P_{F_i}(\mathbf{x}) = \mathrm{P}(\text{facies at point } \mathbf{x} = F_i) = \mathbf{E}\{\mathbf{I}_{F_i}(\mathbf{x})\} \tag{2.6}$$

We assume that the facies at location **x**, $F(\mathbf{x})$ can be described as a function of the Gaussian RF by the following relation:

$$F(\mathbf{x}) = \sum_{i=1}^{k} c_i \mathbf{I}_{F_i} isTrue(t_{i-1} \leq \mathbf{Z}(\mathbf{x}) \leq t_i) \tag{2.7}$$

where k is the number of distinct facies, c_i is the value of the parameter in facies i and *IsTrue*() is a function which returns 1 if its argument is true and 0 otherwise. More Gaussian variables could also be added according to facies spatial associations (Armstrong et al., 2011).

Due to equation 2.4, we have:

$$P(\mathbf{x}) = \mathbf{G}(t_{\mathrm{i}}) - \mathbf{G}(t_{\mathrm{i}-1}) \tag{2.8}$$

where $\mathbf{G}(t)$ is the cumulative distribution function (**cdf**) for the standard normal distribution $\boldsymbol{N}(0,1)$. As the proportions of each facies are known experimentally, one has just to invert this relationship to deduce the thresholds:

$$t_i = \mathbf{G}^{-1}\left(P_{F_1}(\mathbf{x}) + P_{F_2}(\mathbf{x}) + \cdots + P_{F_k}(\mathbf{x})\right) \tag{2.9}$$

2.4.3 The Plurigaussian algorithm

First, one has to infer the variogram of **Z** by using the indicator variograms and the proportions of each facies, plus a truncation rule. The proportions of each facies are calculated based on the observations. The lithotype rule is usually based on available data. In practice, information concerning this rule might also be obtained by in situ inspection, or it might be in the form of empirical knowledge in the wider area. The model variogram of **Z**, which represents the facies distribution, can be determined by a trial and error procedure, using for example the program VMODEL (Emery and Silva, 2009). The criterion of this procedure is the optimal fit of the simulated facies variograms to the observed ones. After the definition of model variogram of **Z**, the Gibbs sampler (see section 4.2.2.2) can be used to generate gaussian values at sample points that have the right covariance and belong to the right intervals. Then, any conditional simulation can be used to generate a realization conditional to gaussian values at sample points. At the end of the process, the Gaussian values of the realization are converted to facies.

2.5 Multivariable Simulation

2.5.1 Introduction

Most natural phenomena involve multiple RFs at spatial locations with possible dependencies between them. Considering the cross-correlation of RFs, the spatial estimation of a RF is improved. Numerous examples exist either between continuous or categorical variables, such as the relationship between the groundwater flow transport and hydraulic conductivity or the correlation between the facies, respectively. As is explained in previous sections, the categorical variables can be also represented by Gaussian RFs. In spatial statistics, the best linear unbiased estimator (**BLUE**) taking into account two or more spatial RFs is referred to as cokriging estimator, which is based on the extension of kriging estimation. The term kriging and cokriging is reserved for linear regression using data on the same and different attributes, respectively. In the next sections, we present the basic outlines of cokriging, the linear coregionalization model, and cosimulation, a simulation procedure that produces realizations taking into account the correlation between the RFs.

2.5.2 Cokriging

Let $\mathbf{Z}_1 = (\mathbf{z}_{11}, \dots, \mathbf{z}_{1\alpha})$ and $\mathbf{Z}_2 = (\mathbf{z}_{21}, \dots, \mathbf{z}_{2\beta})$ two spatial RFs at locations $\mathbf{X}_1 = (\mathrm{x}_{11}, \dots, \mathrm{x}_{1\alpha})$ and $\mathbf{X}_2 = (\mathrm{x}_{21}, \dots, \mathrm{x}_{2\beta})$ respectively. The estimation of dependent variable $\boldsymbol{z}_1$ according to the explanatory (independent) variable $\boldsymbol{z}_2$ at a query point x_q is considered a linear combination of both RFs at the sample points:

$$\hat{\mathbf{z}}_{1q} = \sum_{i=1}^{\alpha} \lambda_{1i} \mathbf{z}_{1i} + \sum_{j=1}^{\beta} \lambda_{2j} \mathbf{z}_{2j} \tag{2.10}$$

where λ_{1i} and λ_{2j} are the cokriging weights that must be defined and assigned to $\mathbf{z}_{1i}$ and $\mathbf{z}_{2j}$ respectively. Then, the error of estimation can be defined by the difference between the estimated and the unknown random variables at query points. Even if the multivariate distribution of the error of estimation is unknown, it is possible to describe certain moments of the linear combination of equation 2.10, such as the mean and the variance, by knowing certain moments of the random variables involved in the combination. In order the estimation to be unbiased, two conditions must be satisfied: The sum of λ_{1i} equals to 1 and the sum of λ_{2j} equals to 0. Moreover, the weights must be such that the error variance is the smallest possible. The minimization of error variance in cokriging estimation can be faced as a classical optimization problem with two constrains of unbiassedness. The optimization problem of the cokriging system requires the definition of $\mathbf{Z}_1$-auto-covariance, $\mathbf{Z}_2$-auto-covariance and the cross-covariance between the RFs $\mathbf{Z}_1$ and $\mathbf{Z}_2$.

2.5.3 Linear Model of Coregionalization

In cokriging estimation, it is clear that the variograms of the two random variables should be known (be modeled) and the cross-variogram should be defined. The auto- and cross-variograms may be defined by the linear model of coregionalization (**LCM**), a method in which all the direct and cross-variograms are derived from linear combinations of *m* basic direct variograms. In contrast to other coregionalization models, the LCM is often used in practice (Chiles and Delfiner, 1999; Wackernagel, 2003) due to its simplicity and versatility.

In order to define the LCM, the individual variograms must be constructed using the same basic models as follows:

$$\begin{aligned}\gamma_{z_1}(\mathbf{h}) &= u_0\gamma_0(\mathbf{h}) + u_1\gamma_1(\mathbf{h}) + \cdots + u_m\gamma_m(\mathbf{h})\\ \gamma_{z_2}(\mathbf{h}) &= v_0\gamma_0(\mathbf{h}) + v_1\gamma_1(\mathbf{h}) + \cdots + v_m\gamma_m(\mathbf{h})\\ \gamma_{z_1z_2}(\mathbf{h}) &= w_0\gamma_0(\mathbf{h}) + w_1\gamma_1(\mathbf{h}) + \cdots + w_m\gamma_m(\mathbf{h})\end{aligned} \tag{2.11}$$

where *u*, *v* and *w* are coefficients, possible negative and *m* is the number of variograms used to define the coregionalization model.

In matrix notation, the above equation can be rewritten as:

$$\begin{bmatrix}\gamma_{z_1} & \gamma_{z_1z_2}\\ \gamma_{z_1z_2} & \gamma_{z_2}\end{bmatrix} = \begin{bmatrix}u_0 & w_0\\ w_0 & v_0\end{bmatrix}\gamma_0(\mathbf{h}) + \begin{bmatrix}u_1 & w_1\\ w_1 & v_1\end{bmatrix}\gamma_1(\mathbf{h}) + \cdots + \begin{bmatrix}u_m & w_m\\ w_m & v_m\end{bmatrix}\gamma_m(\mathbf{h}) \tag{2.12}$$

The coefficients must be chosen so that all the matrices in the above equation are positive definite in order to ensure the linear model of coregionalization is also positive definite, thus: $u_j > 0$,$v_j > 0$, $u_j v_j > w_j w_j$ for all *j* from 0 to *m*.

Due to the non-uniqueness of the solution (different coefficients may result in the same variogram model), it is very difficult to design a completely automated process. Thus, usually the basic models are first chosen by the practitioner and then only the coregionalization matrixes need to be determined based on the experimental data (Emery, 2010). More particularly, a trial and error procedure can be applied to conclude infer the coefficients of LCM: First, calculate the sample cross-variogram between the two variables and propose the coregionalization coefficients. Then, test if the model of coregionalization fits satisfactory with the sample cross-variogram.

Another important thing is that in order to calculate the experimental cross-variogram, it is necessary to have information on both variables at the same locations. As a result, in a heterotopic data configuration, as is the usual case in practical applications, it is not possible to infer the cross-variogram. Instead, in such a situation, one could krige the unknown locations and then use this set of data to compute the experimental cross-variograms.

2.5.4 Cosimulation

Cosimulation is a simulation process that allows the integration of different types of data. Verly (1993) proposes the Sequential Gaussian Cosimulation (**SGC**), a cosimulation method that overcomes the difficulties of previous approaches to incorporate correlations between variables, by the generation of independent realizations of these interest variables. SGC is a

simple extension of SGS that combines sequential simulation and cokriging to generate realizations of several continuous variables simultaneously.

The implementation of SGC requires the assumption of multinormality of RFs. Therefore, the RFs are transformed to normal scores and it is supposed that they are jointly Gaussian. Next, the auto- and cross-variograms of normal scores are calculated and a coregionalization model is fitted. At the final step, cosimulation uses cokriging to compute an estimate of the primary variable at a query point from surrounding actual and simulated values and also using nearby values of the secondary variable. In this way, different realizations are generated using various random paths, keeping the correlation between the RFs. The pdf of cosimulation can be written as:

$$
\begin{aligned}
&f_v\left(\mathbf{z}_{1,1}, \mathbf{z}_{1,2}, \dots, \mathbf{z}_{1,\alpha} | \mathbf{z}_{2,1}, \mathbf{z}_{2,2}, \dots, \mathbf{z}_{2,\beta}\right) = \\
&= f_1\left(\mathbf{z}_{1,1} | \mathbf{z}_{2,1}, \mathbf{z}_{2,2}, \dots, \mathbf{z}_{2,\beta}\right) f_{1|1}\left(\mathbf{z}_{1,2} \middle| \mathbf{z}_{1,1}, \mathbf{z}_{2,1}, \mathbf{z}_{2,2}, \dots, \mathbf{z}_{2,\beta}\right) \cdot \\
&\cdot f_{1|2}\left(\mathbf{z}_{1,3} \middle| \mathbf{z}_{1,1}, \mathbf{z}_{1,2}, \mathbf{z}_{2,1}, \mathbf{z}_{2,2}, \dots, \mathbf{z}_{2,\beta}\right) \cdots \cdot f_{1|\alpha-1}\left(\mathbf{z}_{1,\alpha} \middle| \mathbf{z}_{1,1}, . . , \mathbf{z}_{1,\alpha-1}, \mathbf{z}_{2,1}, \mathbf{z}_{2,2}, \dots, \mathbf{z}_{2,\beta}\right)
\end{aligned}
\tag{2.13}
$$

where $\mathbf{z}_{1,i}$ is the dependent variable and $\mathbf{z}_{2,j}$ the explanatory variable while the α and β indicate the number of query points for variables $\mathbf{z}_1$ and the number of known points (or nearby estimated points) $\mathbf{z}_2$.

Alternative algorithms of SGC have been proposed in order to overcome difficulties of SGS such the modeling of cross-variograms (Almeida and Frykman, 1994), the assumption of stationarity (Chambers et al., 1994) and the transformation of RFs to normal scores (Soares, 2001).

Emery and Silva (2009) present a hybrid model to cosimulate continuous and categorical variables. Their proposal faces the limitations of previous studies (Freulon et al., 1990; Dowd, 1994 and 1997; Bahar and Kelkar, 2000) to incorporate the information between the continuous and categorical variables in realizations.

3 The General Discrete Inverse Problem

3.1 Introduction

The study of a physical system can be divided into (i) the parameterization of the system, (ii) the forward modeling and (iii) the inverse modeling. Parameterization of a system means the definition of the parameters that describe the physical system completely, while the forward modeling is the solution of a physical law that governs the system and allows to make predictions of the physical variable for given parameter values. Inverse modeling uses the response of the physical variable to infer the actual values of the model parameters.

The discrete inverse problem involves a finite number n of nested subunits and data at p points allowing the analysis of the system. In mathematical notation, the model of the physical system under study **g** with boundary conditions **u** can be described by the parameters of the system such as a parameter set **m** = {m_1, …, m_n} and their spatial coordinates **x** over a time *t*. Then, the forward solution of the system is:

$$\mathbf{d} = \mathbf{g}(\mathbf{x}, t, \mathbf{u}; \mathbf{m}) \tag{3.1}$$

In geosciences, the objective of inverse problem is to estimate the spatial distribution of the parameters **m** and their moments over the time, using all the available data such as the physical law, moments and possible measurements of parameters as soft data, but also measurements of state variables (e.g. head measurements) as hard data.

For simplicity, given a set of observations of a physical variable **d** = {d_1**, ..,** d_p}, the equation 3.1 can be written:

$$\mathbf{d} = \mathbf{g}(\mathbf{m}_1, \dots, \mathbf{m}_n) \tag{3.2}$$

The function **g** maps the n parameters to the p measurements of the physical variable. This way, the discretized model space $\mathcal{M}$ is mapped into the space of observations $\mathcal{D}$. The simplest way to achieve this is to consider that $\mathcal{M}$ is mapped in $\mathcal{D}$ linearly. Different criteria may be used for the fitting of the estimated parameter set to the physical system. In the next section, we focus on the least squares criterion, however, it is noted that an alternative one is

the least absolute difference of error which is more robust in outliers but more difficult to work with it (Woodbury et al., 1987).

3.2 The Least Squares Criterion

In the linear case, equation 3.1 can be written as:

$$\mathbf{d}_{\text{obs}} = \mathbf{G}\mathbf{m} \tag{3.3}$$

where $\mathbf{G}$ is a p x n known linear operator:

$$\mathbf{G} = \begin{bmatrix} G_{11} & \cdots & G_{1n} \\ \vdots & \ddots & \vdots \\ G_{\text{p}1} & \cdots & G_{pn} \end{bmatrix} \tag{3.4}$$

Solving the equation 3.3, the perfect case is to find a parameter set that fits the data exactly. This is achieved when the number of observations equals with the number of unknown parameters ($p = n$) and the parameters are independent (e.g. any column of $\mathbf{G}$ cannot be written as a linear combination of the other columns). Mathematically, it means that the number of observations must be equal to the rank of matrix $\mathbf{G}$. Then, a unique solution to the system of linear equations 3.3 can be found, which fits the observations exactly. However, this is impossible in most real applications and practitioners must estimate a parameter set that fits the observations adequately. In the general case, the expression of estimated parameter set is:

$$\widehat{\mathbf{m}} = \mathbf{G}^{-\text{g}}\mathbf{d}_{\text{obs}} \tag{3.5}$$

where $\mathbf{G}^{-\text{g}}$ the is the generalized inverse matrix of $\mathbf{G}$ which maps the observations to the space of $\mathcal{M}$. The simplest method to estimate the parameter set is the ordinary least squares (**OLS**) regression. OLS is a method to evaluate the optimum parameter set by minimizing the sum of square differences between the measurements and the response of the system.

3.2.1 Overdetermined Optimization Problem

In classical linear optimization problems, the number of observations is larger than the rank of $\mathbf{G}$ matrix ($p > r > n$), or, larger than the number of parameters ($p > n$) in the case where these parameters are independent. In this case, the system of linear equations is inconsistent

and is referred as overdetermined. More specifically, $\mathbf{d}_{\text{obs}}$ is not a linear combination of the columns of $\mathbf{G}$, hence the equality sign of equation 3.3 is violated and it is used to be written as:

$$\mathbf{d}_{\text{obs}} \cong \mathbf{G}\mathbf{m} \tag{3.6}$$

Then, we are looking for an estimation of parameter $\widehat{\mathbf{m}}$ which has the minimum least square error. The equation 3.6 is written as:

$$\mathbf{d}_{\text{obs}} = \mathbf{G}\mathbf{m} + \mathbf{e} \tag{3.7}$$

where $\mathbf{e}$ is random noise known as residual and $\mathbf{m}$ is a deterministic but unknown parameter set. The objective function to be minimized is:

$$\mathbf{S}(\mathbf{m}) = (\mathbf{d}_{\text{obs}} - \mathbf{G}\widehat{\mathbf{m}})^T(\mathbf{d}_{\text{obs}} - \mathbf{G}\widehat{\mathbf{m}}) \tag{3.8}$$

Setting the expanded derivative of the objective function equal to zero, we have (see Appendix A.4):

$$\widehat{\mathbf{m}} = (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}_{\text{obs}} \tag{3.9}$$

Moreover, it should be shown that the second order derivative of the objective function (Hessian) is positive definite ($\mathbf{G}^{\mathrm{T}}\mathbf{G} > 0$) in order to prove the estimated $\widehat{\mathbf{m}}$ is the minimum of the objective function.

The problem is explained geometrically in Figure 3.1. The plane is an n-dimensional subspace of $\mathcal{D}$ and it represents all possible vectors $\mathbf{Gm}$, which constitute the range $\mathcal{R}(\mathbf{G})$ of $\mathbf{G}$, or the column space of $\mathbf{G}$. The vector $\mathbf{G}\widehat{\mathbf{m}}$ is the orthogonal projection of $\mathbf{d}_{\text{obs}}$ in the column space of $\mathbf{G}$ and the residual $\mathbf{e} = (\mathbf{d}_{\text{obs}} - \mathbf{G}\widehat{\mathbf{m}})$ is vertical to $\mathbf{G}\widehat{\mathbf{m}}$. It is noted that projection ensures the minimum distance between two vectors in a Euclidean space. Due to the orthogonality of $\mathbf{G}^{\mathrm{T}}$ and $(\mathbf{d}_{\text{obs}} - \mathbf{G}\widehat{\mathbf{m}})$, the inner product of them equals to zero. Therefore, the best OLS estimation of $\mathbf{m}$ can be written as:

Normal Equation: $$\widehat{\mathbf{m}} = (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}_{\text{obs}} \tag{3.10}$$

It is noted that $\widehat{\mathbf{m}}$ solves exactly the equation 3.3 when the linear system is consistent.

Also, considering the mean of the estimator:`

$$\mathrm{E}\{\widehat{\mathbf{m}}\} = \mathrm{E}\{(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}_{\text{obs}}\} = (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathrm{E}(\mathbf{d}_{\text{obs}}) \tag{3.11}$$

$$= (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathrm{E}\{\mathbf{Gm} + \mathbf{e}\} = (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{Gm} + (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathrm{E}\{\mathbf{e}\} = \mathbf{m}$$

Thus, $\hat{\mathbf{m}}$ is referred as BLUE of $\mathbf{m}$.

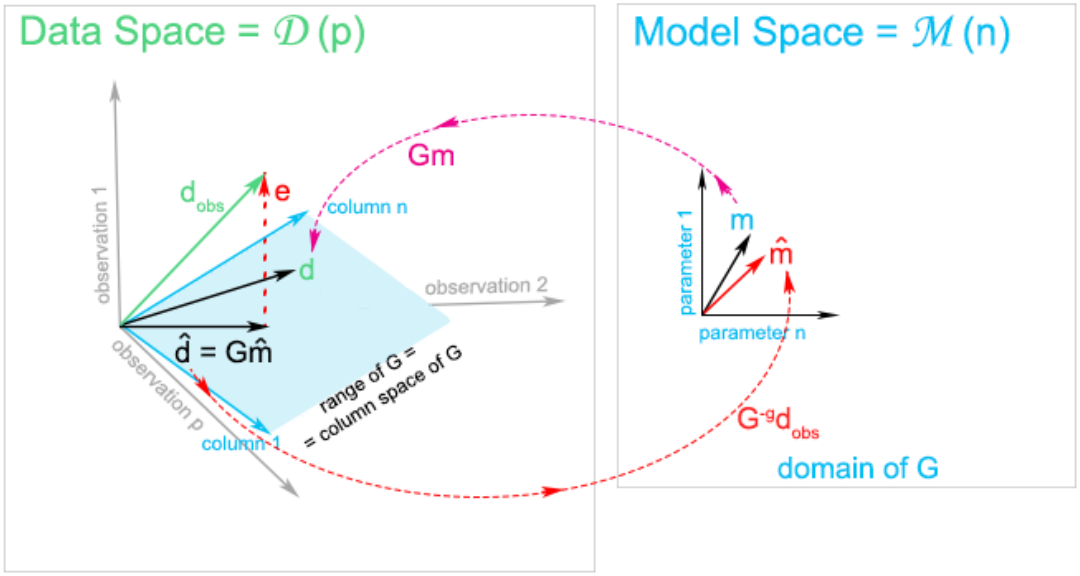


Figure 3.1: The geometrical interpretation of overdetermined least squares estimation. When $n < p$, the range of $\mathbf{G}$ is a subspace of data space $\mathcal{D}$, so we can have no exact solution to $\mathbf{d}_{\text{obs}} = \mathbf{Gm}$. We then solve $\mathbf{G}^{\mathrm{T}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) = 0$ to find $\hat{\mathbf{m}}$ with the smallest error.

The equation 3.10 implies that the matrix $\mathbf{G}^{\mathrm{T}}\mathbf{G}$ is nonsingular. This happens only when the column rank of matrix $\mathbf{G}$ equals the number of parameters to be estimated (all parameters are linearly independent), so we call the problem purely overdetermined. Then, we can find a unique estimator $\hat{\mathbf{m}}$ with the minimum least square error and the generalized inverse matrix of $\mathbf{G}$ is:

$$\mathbf{G}^{-\mathrm{g}} \equiv (\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}} \tag{3.12}$$

3.2.2 Underdetermined Optimization Problems

Unlike the previous case, most geoscientifical optimization problems are underdetermined: the number of linearly independent unknown parameters (column rank of $\mathbf{G}$) is larger than the number of observations ($r > p$).

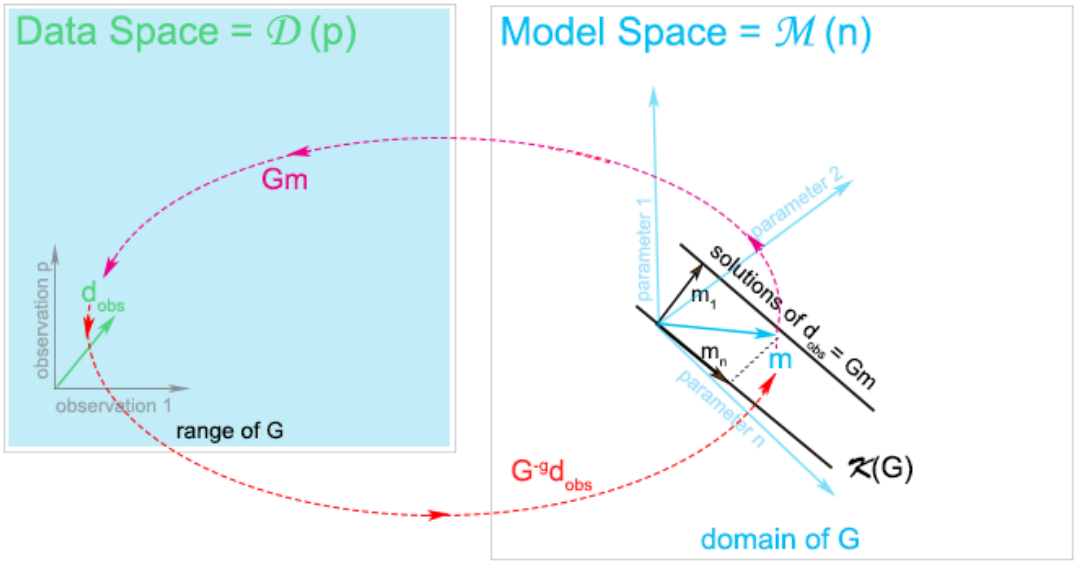


Figure 3.2: The geometrical interpretation of underdetermined least squares. When $n > p$ and $r > p$, the range of $\mathbf{G}$ covers the entire data space, so we can find infinite solutions to $\mathbf{d}_{\text{obs}} = \mathbf{Gm}$. We may, then, reasonably choose the minimum norm solution ($\mathbf{m}_1$).

Thus, the system of linear equations 3.3 has infinite solutions and we should add a-priori information to obtain a unique solution. A possible a-priori information is our expectation that the estimator is the simplest one (Menke, 1984, pp. 49). The Euclidean length $\hat{\mathbf{m}}^{\mathrm{T}}\hat{\mathbf{m}}$ is a quantity that identifies the simplicity of the solution. Among the solutions, we consider that the best solution is the solution with minimum length. The optimization problem now is to find the minimum length of $\mathbf{m}$ with the constrain that it fits the data exactly. It is proven (see Appendix A.5) that the best OLS solution in this case is:

$$\mathbf{m} = \mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{d}_{\text{obs}} \tag{3.13}$$

The generalized inverse matrix of $\mathbf{G}$ of the solution $\mathbf{m}$ is now:

$$\mathbf{G}^{-g} \equiv \mathbf{G}^{T}(\mathbf{G}\mathbf{G}^{T})^{-1} \tag{3.14}$$

3.2.3 Ill-conditioned, Ill-posed systems and regularization

Ill-conditioned linear systems are defined as those which are unsolvable due to limits on computational precision. In practice, the generalized matrix $\mathbf{G}^{-g}$ cannot be computed because $\mathbf{G}\mathbf{G}^{T}$ is a singular or nearly singular matrix. This happens when the condition number of $\mathbf{G}$, which is the ratio of its largest and smallest singular values, is very large. Intuitively, the solution of the linear system will magnify any noise in the data. The system is thus affected significantly by small changes in the observations, which cause large variations in the estimator of the parameter set.

Well-posed systems are defined as those which the solution exists, it is unique and it depends continuously on data after the definition of Hadamard (1923). Problems that are not well-posed are termed ill-posed. A well-posed problem may be ill-conditioned due to precision or a well-conditioned problem and vice versa.

The objective in these problems is not only to find an unbiased estimator, but also one with a stable behavior (e.g. the estimator will be a good one even if a small change occurs on the observations). In other words, the estimator should have reduced variance. A common practice to do this is by shrinking its coefficients (model parameters) by a constant factor. This kind of regularization introduces bias in the estimator in the form of extra information and is added to the objective function as a penalty term. The interpretation of the above regularization is based on the acceptance that the more complex the estimator, the more is specialized on the particular measurements. But if the result of estimation is overfitting, then this effect should be penalized. The regularized objective function is expressed as:

$$\mathbf{S}(\mathbf{m}) = (\mathbf{G}\hat{\mathbf{m}} - \mathbf{d}_{\text{obs}})^2 + \lambda\hat{\mathbf{m}}^2 \tag{3.15}$$

where λ controls the amount of shrinkage of the parameter length. After the minimization of the objective function, the resulting estimator, also known as ridge estimator (Hoerl and Kennard, 1968; 1970a; 1970b), is written as:

$$\hat{\mathbf{m}}_{\text{ridge}} = (\mathbf{G}^{T}\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^{T}\mathbf{d}_{\text{obs}} \tag{3.16}$$

where $\mathbf{I}$ is p x p identity matrix. In practice, the inclusion of a shrinkage parameter makes the problem non-singular. In the above equation, bias is introduced by the term $\lambda\mathbf{I}$: the estimator is not penalized when λ is zero, while, with large values of λ, the parameters are heavily constrained.

The regularized objective function can also be formulated into the following constrained optimization problem:

$$\min(\mathbf{G}\hat{\mathbf{m}} - \mathbf{d}_{\text{obs}})^2$$
$$\text{subject to } \hat{\mathbf{m}}^2 < c\,, \qquad (3.17)$$
where c is a positive value.

The two optimization problems have the same solution when $c = \hat{\mathbf{m}}^2_{\text{ridge}}$. Then, the value of c is the radius of the circle in Figure 3.3 as explained below.

In a more statistical view, the regularized objective function (3.15) expresses the sources of errors in the model prediction. The errors can be divided into errors due to bias and errors due to the model (parameter) variance, while λ is the adjustment parameter of bias-variance tradeoff. The geometric interpretation of this dilemma is depicted in Figure 3.3. The estimators into the inner ellipses have a smaller residual sum of squares (**RSS**) and RSS is minimized at OLS estimate in the center. The circle in the middle indicates the penalty term. The optimal point is one which is a common point between ellipse and circle and gives a minimum value for the above function as well.

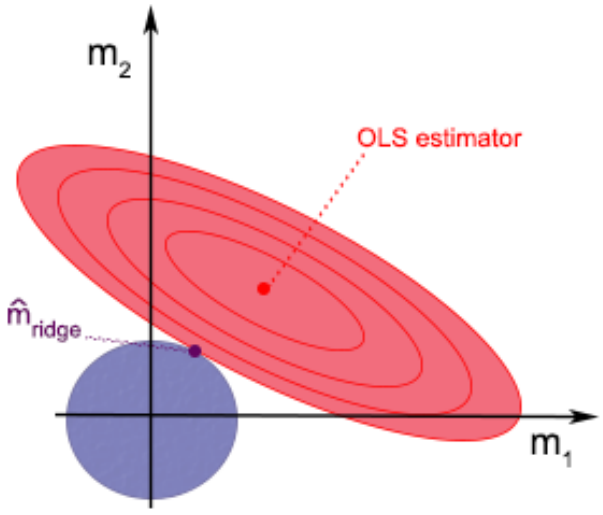


Figure 3.3: Geometric interpretation of bias-variance tradeoff.

3.3 Statistical Aspects of Least squares

3.3.1 Weighted Least Squares and Maximum Likelihood Estimation

If we consider that the observations $\mathbf{d}_{\text{obs}}$ are not perfect measurements but rather they possess white noise (i.e. random errors $\mathbf{e}_j$ independent of each other with zero mean and variance σ_j^2), it is reasonable to weight the squared observation errors according to their variance. A squared error with small variance indicates a more accurate observation and thus it is weighted heavier than others.

In this case, the error weighting matrix is:

$$\mathbf{W}_{\text{D}} = \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_p \end{bmatrix} \tag{3.18}$$

If $\mathbf{C}_{\text{D}}$ is the positive definite and symmetric matrix $\mathbf{W}_{\text{D}}^{\text{T}}\mathbf{W}_{\text{D}}$ used to weight the squared errors, then the objective function to be minimized in the overdetermined case becomes:

$$\mathbf{S}(\mathbf{m}) = (\mathbf{d}_{\text{obs}} - \mathbf{G}\widehat{\mathbf{m}})^T \mathbf{C}_{\text{D}}^{-1} (\mathbf{d}_{\text{obs}} - \mathbf{G}\widehat{\mathbf{m}}) \tag{3.19}$$

Then, the OLS estimator is given by the expression (see Appendix A.6):

$$\widehat{\mathbf{m}}_{\text{w}} = (\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{d}_{\text{obs}} \tag{3.20}$$

where

$$\mathbf{C}_{\text{D}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_{\text{p}}^2} \end{bmatrix} \tag{3.21}$$

But, adding random errors to the observations, transforms them to random variables and their pdf's are used to define the joint distribution that characterizes the solution space. Working on a sample of size $p\,(\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_p)$, each random variable $\mathbf{d}_j$ belongs to $f(\mathbf{d}_{\text{obs}}\,|\mathbf{m}_{\text{w}})$ and now the objective is to find the parameter set that maximizes the equation:

$$L(\mathbf{m}_w; \mathbf{d}_{\text{obs}}) = \prod_{\text{i}=1}^{\text{p}} f(\mathbf{d}_j | \mathbf{m}_{\text{w}}) \tag{3.22}$$

In this case, L is known as the likelihood function, because it measures the most likely model as parameterized by $\mathbf{m}$ and $\mathbf{d}_{\text{obs}}$. Under the assumption that the errors are independently distributed and $\mathbf{e}_j \sim \boldsymbol{N}(0, \sigma_j^2)$, substitution leads to:

$$\begin{aligned} f(\mathbf{e}) &= \frac{1}{\sqrt{2\pi^p |\mathbf{C}_{\text{D}}|}} \exp\left(-\frac{1}{2} (\mathbf{e} - \mathbf{0})^T \mathbf{C}_{\text{D}}^{-1} (\mathbf{e} - \mathbf{0}) \right) \\ &= \frac{1}{\sqrt{2\pi^p |\mathbf{C}_{\text{D}}|}} \exp\left(-\frac{1}{2} (\mathbf{d}_{\text{obs}} - \mathbf{Gm})^T \mathbf{C}_{\text{D}}^{-1} (\mathbf{d}_{\text{obs}} - \mathbf{Gm}) \right) \end{aligned} \tag{3.23}$$

where $|\mathbf{C}_{\text{D}}|$ is the determinant of $\mathbf{C}_{\text{D}}$.

Thus, the distribution of $\mathbf{d}_{\text{obs}}$ parameterized by $\mathbf{m}$ is $\boldsymbol{N}(\mathbf{Gm}, \mathbf{C}_{\text{D}})$.

The objective now is to maximize the equation:

$$L(\mathbf{m}; \mathbf{d}_{\text{obs}}) = \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\left(\mathbf{d}_j - (\mathbf{Gm})_j\right)^2}{2\sigma_j^2} \right) \tag{3.24}$$

The maximization of the likelihood is equivalent to the minimization of the exponent $(\mathbf{d}_{\text{obs}} - \mathbf{Gm})^T \mathbf{C}_{\text{D}}^{-1} (\mathbf{d}_{\text{obs}} - \mathbf{Gm})$, which is the objective function of the weighted least squares criterion in 3.19 above. That is, the maximum likelihood estimate of the model parameters is simply the weighted least squares solution, where the weighting matrix is the inverse of the covariance matrix of the data. If the data are uncorrelated with equal variance, then the maximum likelihood solution coincides to the simple least squares solution.

If the linear problem is underdetermined, there may be infinite solutions to the least squares inverse, as seen in section 3.2.2. To solve this underdetermined problem, a priori information can be added that causes the distribution of the data to have a well-defined peak. In this case, the solution is given by:

$$\mathbf{m}_{\text{w}} = \mathbf{G}^{\text{T}} \mathbf{C}_{\text{D}}^{-1} (\mathbf{G} \mathbf{C}_{\text{D}}^{-1} \mathbf{G}^{\text{T}})^{-\mathbf{1}} \mathbf{d}_{\text{obs}} \tag{3.25}$$

3.3.2 A-priori information in regularization

Prior information on the observations in the form of weights may also be used in ill-conditioned systems (3.2.3). Weighting the observations, the ridge estimator is written as:

$$\hat{\mathbf{m}}_{\mathrm{w}} = (\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G} + \lambda\mathbf{I})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{d}_{\mathrm{obs}} \tag{3.26}$$

In a similar way, weights on parameters ($\mathbf{W}_{\mathrm{M}}$) could also be used if prior information exists about them. In this case, the regularized objective function is written:

$$\mathbf{S}_{\mathrm{d,m}} = (\mathbf{d}_{\mathrm{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}(\mathbf{d}_{\mathrm{obs}} - \mathbf{G}\hat{\mathbf{m}}) + \lambda\hat{\mathbf{m}}^{\mathrm{T}}\mathbf{C}_{\mathrm{M}}^{-1}\hat{\mathbf{m}} \tag{3.27}$$

and the regularized estimator resulting from minimizing the above objective function is:

$$\hat{\mathbf{m}}_{\mathrm{w}} = (\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G} + \lambda\mathbf{C}_{\mathrm{M}}^{-1})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{d}_{\mathrm{obs}} \tag{3.28}$$

where $\mathbf{C}_{\mathrm{M}}$ is the positive definite and symmetrical matrix $\mathbf{W}_{\mathrm{M}}^{\mathrm{T}}\mathbf{W}_{\mathrm{M}}$.

Furthermore, we have seen in underdetermined problems (section 3.2.2), that the objective is to find the solution with the minimum length; that is to say, we use the information that the length of parameters should be close to zero. However, possible prior knowledge of the length of parameters could be used instead, even if the resulting estimator will not have the minimum length parameter set. For example, an a-priori set of parameters $\mathbf{m}^{\mathrm{pr}}$ can be used to express a desirable length instead of the one close to zero. This practice is adopted mainly in ill-conditioned systems when a-priori information is crucial in order to solve complex problems. In this case, the regularized objective function can be written:

$$\mathbf{S}_{\mathrm{d,m}} = (\mathbf{d}_{\mathrm{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}(\mathbf{d}_{\mathrm{obs}} - \mathbf{G}\hat{\mathbf{m}}) + \lambda(\mathbf{m}^{\mathrm{pr}} - \hat{\mathbf{m}})^{\mathrm{T}}\mathbf{C}_{\mathrm{M}}^{-1}(\mathbf{m}^{\mathrm{pr}} - \hat{\mathbf{m}}) \tag{3.29}$$

The regularized estimator is now (see Appendix A.7):

$$\hat{\mathbf{m}}_{\mathrm{w}} = \mathbf{m}^{\mathrm{pr}} + (\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G} + \lambda\mathbf{C}_{\mathrm{M}}^{-1})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}(\mathbf{d}_{\mathrm{obs}} - \mathbf{G}\mathbf{m}^{\mathrm{pr}}) \tag{3.30}$$

We note that the equation 3.30 can also be written as:

$$\hat{\mathbf{m}}_{\mathrm{w}} = \mathbf{m}^{\mathrm{pr}} + \mathbf{C}_{\mathrm{M}}\mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{C}_{\mathrm{M}}\mathbf{G}^{\mathrm{T}} + \mathbf{C}_{\mathrm{D}})^{-\mathbf{1}}(\mathbf{d}_{\mathrm{obs}} - \mathbf{G}\mathbf{m}^{\mathrm{pr}}) \tag{3.31}$$

after the matrix inversion identity (see Appendix A.3):

$$\mathbf{C}_{\mathrm{M}}\mathbf{G}^{\mathrm{T}}\left(\mathbf{G}\mathbf{C}_{\mathrm{M}}\mathbf{G}^{\mathrm{T}}+\mathbf{C}_{\mathrm{D}}\right)^{-\mathbf{1}}=\left(\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G}+\lambda\mathbf{C}_{\mathrm{M}}^{-1}\right)^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1} \quad (3.32)$$

Due to the computation efficiency, the equation 3.31 is preferable when the number of observations p is smaller than the number of parameter set n, since the demanding computation of $\mathbf{C}_{\mathrm{M}}^{-1}$ is avoided.

Technically, the existence of prior knowledge about the measurements and the parameter set is helpful in order to derive the generalized inverse matrix. In a deterministic view of least squares, a conceptual problem is how to define the weights, since observations and parameters are not treated as random variables. However, equations hold anyway and weighting can be seen as an implementation of our expertise on the system, without the assumption of randomness.

3.4 Nonlinear least squares

In nonlinear problems, the estimator or the solution is obtained by an iterative process, using either a deterministic or a stochastic method. In deterministic methods, the objective function is linearized (weakly nonlinear case) around a parameter set $\mathbf{m}^*$ and a gradient method is chosen to produce a sequence of iterates $\mathbf{m}^i$. The minimization of the objective function requires such a sequence of iterates $\mathbf{m}^i$, for which the corresponding sequence of objective function values is monotonically decreasing. A positive definite second order derivative of the objective function (Hessian) guarantees the condition of decreasing of objective function values between two neighborhood parameter sets. Among the gradient methods, the Newton method is the only well-defined method because it uses directly the Hessian to produce the sequence of iterates $\mathbf{m}^i$. The Newton method approximates locally the objective function with a quadratic form at each $\mathbf{m}^i$ and the minimization of the approximated objective function is done at that point. All the other gradient algorithms aim to emulate the local convergence properties of Newton's method.

The gradient methods work by following the steps described in the flow chart of Figure 3.4. The first partial derivative of the objective function at a given parameter set $\mathbf{m}^i$ indicates the gradient or the slope of the objective function at $\mathbf{m}^i$. In order to find the next parameter set closer to the global minimum of the objective function, an opposite direction from the gradient should be followed defining the search direction $\mathbf{s_d}$. An important component of the

gradient methods is the step size $\mathbf{a}^i$ to be taken in the indicated direction. Different algorithms to choose the step size have been proposed producing various modified gradient methods. A small step will increase the convergence time, while a large step won't guarantee that the objective function, given the new parameter set, will be closer to the global minimum.

The number of iterations also depends on the stopping criteria of the algorithm is used. Using a threshold as tolerance, the algorithm stops if the tolerance is crossed. A very small tolerance increases the computation cost of the estimation while a large tolerance produces a large error in the estimation. Usually, the tolerance is the value that the iterations converge or the difference between successive estimations is not significant.

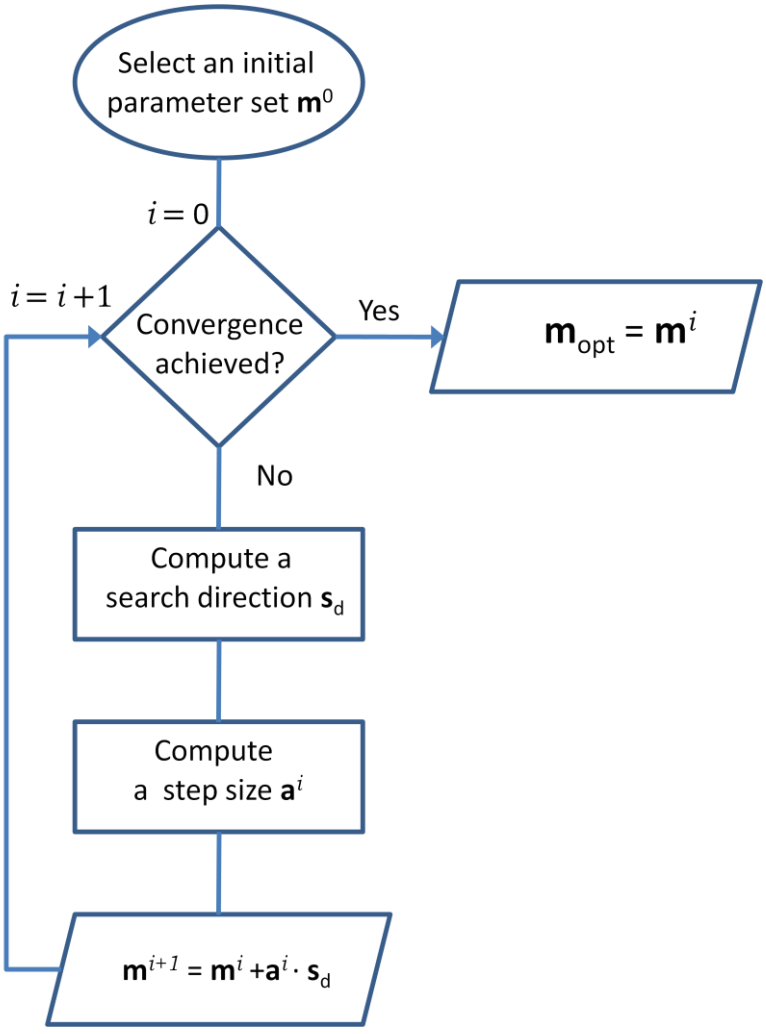


Figure 3.4: Generic diagram of gradient methods.

The computation of Jacobian or sensitivity matrix is necessary for gradient methods and useful due to the fact that it involves important information about the reliability of both state

variables and model parameters. Jacobian can be obtained by the direct derivation of physical model with respect to the model parameters, or by adjoint state method and finite differences. An analytical discussion and comparison of these methods can be found in Carrera et al. (1990) and Carrera and Medina (1994). Generally, the choice of the appropriate method depends on the nature of the physical law and the available information on variables involved. It should be noticed that the direct or finite difference method requires the solution of the forward problem $n+1$ times, while the adjoint method requires $p+1$ forward runs. Therefore, adjoint state method is preferred when the number of observations (p) is smaller than the model parameters (n).

More particularly, in the case of non-linear least squares problems, the objective of the iterative process is to find a parameter set in the hyperplane which is tangential to the surface of error (Figure 3.5). The slope of tangent plane is described by the Jacobian $\mathbf{G}_i$ of error function $\mathbf{r}(\mathbf{m}) = \mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m})$, where $\mathbf{g}(\mathbf{m}^i)$ is the solution of the forward model using the parameter set $\mathbf{m}^i$. The new parameter set $\mathbf{m}^{i+1}$ in the tangent plane should correspond to the point of the data space that is closest to the origin. This is a generalization of the linear case (Figure 3.1) where the image of $\mathbf{g}$ was a hyperplane embedded in the data space. Using Taylor series, a first order approximation of $\mathbf{r}(\mathbf{m})$ around $\mathbf{m}^i$ is written as:

$$\mathbf{r}(\mathbf{m}^i + \boldsymbol{\delta}_{\text{t}}) \approx \mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i \boldsymbol{\delta}_{\mathbf{t}} \tag{3.33}$$

The desirable parameter set $\mathbf{m}^{i+1}$ in the tangent plane can be found by the projection of the origin to the tangent plane. Then, the product $\mathbf{G}_i^{\text{T}} \boldsymbol{\delta}_{\mathbf{t}} [\mathbf{r}(\mathbf{m}^i) + \mathbf{G}_i \boldsymbol{\delta}_{\text{t}}]$ equals to zero.

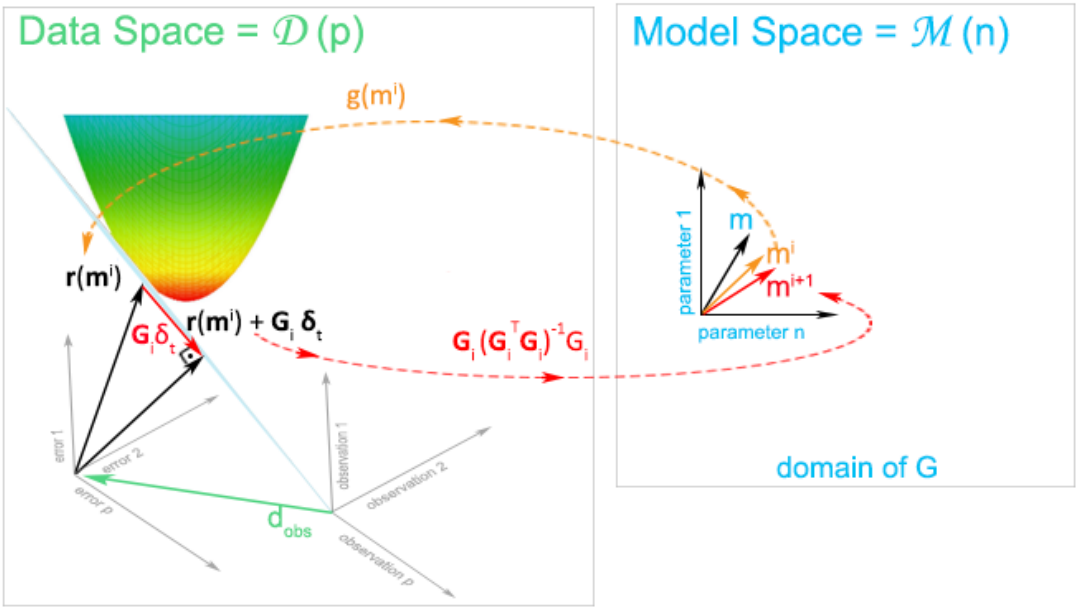


Figure 3.5: Geometric interpretation of an un-weighted ($\mathbf{C}_\mathrm{D} = \mathbf{I}$) nonlinear least squares problem. The error function $\mathbf{r}(\mathbf{m})$ defines a surface in $\mathcal{R}^p$ (here $p = 3$). The solution to the least squares problem is the point on the surface that is closest to the origin **O**. At the point $\mathbf{r}(\mathbf{m}^i)$ the surface is locally approximated by the tangent plane spanned by the columns of $\mathbf{G}_i$.The point on the tangent plane closest to the origin is found by the orthogonal projection of the negative error $\mathbf{r}(\mathbf{m}^i)$: $\mathbf{G}_i\boldsymbol{\delta}_\mathrm{t} = \mathbf{G}_i\left(\mathbf{G}_i^\mathrm{T}\mathbf{G}_i\right)^{-1}\mathbf{G}_i^\mathrm{T}\left[-\mathbf{r}(\mathbf{m}^i)\right]$.

It results that:

$$\mathbf{m}^{i+1} = \mathbf{m}^i + \left(\mathbf{G}_i^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{G}_i\right)^{-1}\mathbf{G}_i^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\left(\mathbf{d}_\mathrm{obs} - \mathbf{g}(\mathbf{m}^i)\right) \tag{3.34}$$

Where $\mathbf{G}_i$ is a matrix with the derivatives of $\mathbf{g}$ with respect to $\mathbf{m}$ at ith iteration (Jacobian or Sensitivity matrix):

$$\mathbf{G}_i = \begin{bmatrix} \frac{\partial \mathbf{g}_1(\mathbf{m})}{\partial \mathbf{m}_1^i} & \cdots & \frac{\partial \mathbf{g}_1(\mathbf{m})}{\partial \mathbf{m}_\mathrm{n}^i} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{g}_p(\mathbf{m})}{\partial \mathbf{m}_1^i} & \cdots & \frac{\partial \mathbf{g}_p(\mathbf{m})}{\partial \mathbf{m}_\mathrm{n}^i} \end{bmatrix} \tag{3.35}$$

The above method is known as Gauss-Newton method and the same result can be derived by linearly approximating the vector of functions $\mathbf{r}$, using Taylor's theorem (see Appendix A.8).

The Gauss-Newton method can be considered as an approximation of the Newton method, which uses the second derivative of forward model:

$$\mathbf{m}^{i+1} = \mathbf{m}^{i} + \left(\mathbf{G}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G}_i - \mathbf{H}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\mathbf{m}^{i})\right)\right)^{-1}\mathbf{G}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\mathbf{m}^{i})\right) \tag{3.36}$$

where $\mathbf{H}_i$ is the Hessian matrix n x n or the second-order partial derivatives of forward model:

$$\mathbf{H}_i = \begin{bmatrix} \dfrac{\partial^2\mathbf{g}(\mathbf{m})}{\partial {\mathbf{m}_1^i}^2} & \cdots & \dfrac{\partial^2\mathbf{g}(\mathbf{m})}{\partial\mathbf{m}_1^i\,\partial\mathbf{m}_n^i} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2\mathbf{g}(\mathbf{m})}{\partial\mathbf{m}_n^i\,\partial\mathbf{m}_1^i} & \cdots & \dfrac{\partial^2\mathbf{g}(\mathbf{m})}{\partial {\mathbf{m}_n^i}^2} \end{bmatrix} \tag{3.37}$$

In the first parenthesis of the equation 3.36, the first term is more important than the second one, which is close to zero either because the error is very small, or the matrix $\mathbf{H}_i$ has very small values due to the quasi-linearity of the forward model. Assuming that the second term equals to zero, the method is an approximation of Newton method as mentioned before.

By the same process, the optimal solution of the objective function $\mathbf{S}_{\mathrm{d,m}}$ in the case of an ill-conditioned system is (see Appendix A.9):

$$\mathbf{m}^{i+1} = \mathbf{m}^{i} + \left(\mathbf{G}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G}_i + \lambda\mathbf{C}_{\mathrm{M}}^{-1}\right)^{-1}\mathbf{G}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\mathbf{m}^{i}) - \mathbf{G}_i\mathbf{m}^{i}\right) \tag{3.38}$$

But, while in the linear case one step is needed to obtain the ridge estimator and the parameter λ is defined once, in nonlinear estimation Levenberg (1944) and Marquardt (1963) propose the parameter λ could be modified according to the improvement of estimation. Small values of λ result in a Gauss-Newton method while large values of λ result in a gradient descent update.

In strongly nonlinear problems the gradient methods usually get stuck at local optima depending on the starting parameter set $\mathbf{m}^0$, therefore the choice of $\mathbf{m}^0$ is an important factor for the convergence of the sequence of iterates. Gradient descent methods have better convergence than the Gauss-Newton method when the initial estimation is far than the optimum while the Gauss-Newton method is more efficient close to the optimum. Therefore, the Levenberg–Marquardt algorithm is preferable. The Newton method is more effective than both gradient descent and Gauss-Newton methods in complicated forward problems. However, the computation of Hessian matrix may be a demanding task with high computational cost and therefore the quasi-Newton methods are preferred, where the Hessian is approximated in

terms of the first derivative of forward model and an appropriate step size is chosen to derive the new estimation parameter set (e.g. Fletcher, 1987; Bonnans et al., 2006). Carrera and Neuman (1986) propose a combination of quasi-Newton and conjugate gradient methods. Their algorithm has much smaller computation cost than other gradient methods, however, they conclude that the Levenberg-Marquardt method is the most robust one.

In order to overcome the above problems of gradient methods, alternative stochastic methods have been proposed such as the simulated annealing, McMC and genetic algorithms **GAs**. All these methods are also iterative but perform a stochastic search for the optimal solution instead.

Simulated annealing is an optimization method which allows to a certain degree the move from a better point of the objective function to a worse one in order to escape from local optima. The probability of accepting a move decreases as the values of objective function increase; therefore it can be seen as a time inhomogeneous Markov chain. However, simulated annealing can also be modeled as time-homogeneous Markov chain. In this case, the number of iterates is not accounted in the transition probability from one state to another. McMC algorithms use a probabilistic rule to select the next candidate model in a chain. As explained in the next sections, apart from searching for an optimal solution, McMC methods are also appropriate for sampling the posterior distribution.

Finally, heuristic methods have been developed such as GAs in order to search for global optima. GAs mimic the process of natural evolution (Holland, 1975; Goldberg, 1989). More specific, a random population of candidate solutions is evolved toward better solutions iteratively. In each iterate, a new population is generated by the previous and it is called as generation. The mechanic to produce a new generation based on the evaluation (using an objective function) of the individuals of the previous generation known as chromosomes. The chromosomes of the previous generation with the best fitness are chosen to produce the next generation by applying genetic operations (mutation and crossover).The fundamental difference between GAs and other stochastic methods is that GAs work with a population of candidate solutions while the other methods try to optimize an initial estimation. Furthermore, McMC methods use probabilistic rules to sample the posterior distribution while the objective of GAs is to find global optima.

4 Bayesian framework for inversion

4.1 Bayes' rule in a physical system

In a probabilistic formulation, the observations, the response variable $\mathbf{d} = \mathbf{g}(\mathbf{m})$ and the parameter set of a system are considered as random variables and symbolized as $\mathbf{d}_{\text{obs}}$, $\mathbf{d}|\mathbf{m}$ and $\mathbf{m}$ respectively. The vector of $\mathbf{d}_{\text{obs}}$ is considered as the response from a particular parameter set, which has to be estimated. As mentioned in section 3.3.1, the distribution of $\mathbf{d}$ given $\mathbf{m}$ can be defined by assuming a known error distribution of the observations. Now, we also consider that the parameter set has a known prior distribution $f(\mathbf{m})$. Then, according to Bayes' rule, which is based on the definition of conditional density, the a-posteriori distribution of the parameters is defined as:

$$f(\mathbf{m}|\mathbf{d}) = \frac{f(\mathbf{d}|\mathbf{m})f(\mathbf{m})}{f(\mathbf{d})} \tag{4.1}$$

where $f(\mathbf{d}|\mathbf{m})$ is the likelihood of data and $f(\mathbf{m})$ is the prior distribution of the parameter set. Repeating the calculation of equation 4.1 for different parameter sets coming from their prior pdf, the result is the construction of the a-posteriori distribution of $\mathbf{m}$. The denominator has a constant value for different parameter sets and ensures that the posterior density adds up to one and therefore is called normalization constant. Since the denominator is constant, equation 4.1 can be written as:

$$f(\mathbf{m}|\mathbf{d}) \propto f(\mathbf{d}|\mathbf{m})f(\mathbf{m}) \tag{4.2}$$

Then, the best estimator of $\mathbf{m}$ is the estimator that maximizes the a-posteriori probability $f(\mathbf{m}|\mathbf{d})$. Therefore this method is called Maximum A-posteriori Probability (**MAP**).

The Bayesian approach to be applied differs according the available prior information and the objective of the research. In optimization problems, researchers often use strong informative priors, while weakly informative priors are used to generate realizations in order to sample the posterior distribution.

Generally, in the Bayesian framework for inversion, the proposed parameter sets are evaluated by both their prior pdf and their likelihood, while in the maximum likelihood method of section 3.3.1, all the proposed parameter sets are considered equiprobable.

4.1.1 Linear theory with Gaussian prior

Assuming that the response variable is related with the parameter set with a linear equation: $\mathbf{d} = \mathbf{G}\mathbf{m} + \mathbf{e}$, where $\mathbf{G}$ is a known linear operator and $\mathbf{e}$ is a random variable following a Gaussian distribution with zero mean and covariance $\mathbf{C}_{\mathrm{D}}$. In the Bayesian formulation, the vector of $\mathbf{d}_{\mathrm{obs}}$ is considered as the response from a particular realization of a parameter set to be estimated. Thus, we assume that $\mathbf{d} \sim \boldsymbol{N}(\mathbf{d}_{\mathrm{obs}}, \mathbf{C}_{\mathrm{D}})$. Since we know the prior distribution $f(\mathbf{m})$ we can generate a realization of $\mathbf{m}$ and the corresponding response values $\mathbf{d} = (d_1, d_2, \dots, d_p)$ where each random variable d_j belongs to $f(d_j; \mathbf{m})$.

Theoretically, the prior and the likelihood could have any form. However, it is a common practice to choose a prior for which the posterior has the same algebraic form as the prior. This practice results in a more convenient calculation of the posterior distribution and this prior is called a conjugate prior. Assuming that the prior distribution of $\mathbf{m}$ is Gaussian with mean $\mathbf{m}^{\mathrm{pr}}$ and covariance $\mathbf{C}_{\mathrm{M}}$, we can write:

$$\mathbf{m} = \mathbf{m}^{\mathrm{pr}} + \mathbf{v} \tag{4.3}$$

with $\mathbf{v}$ the error of the prior $\mathbf{m}$, independently distributed with $\mathbf{v} \sim \boldsymbol{N}(0, \mathbf{C}_{\mathrm{M}})$ and $\mathbf{C}_{\mathrm{M}}$ a $n \times n$ known spatial covariance matrix of $\mathbf{m}$ (i.e., expected variabilities given by the variogram under the geostatistical approach). Thus, the posterior probability is proportional to:

$$\begin{aligned} & f(\mathbf{m}|\mathbf{d}) \propto f(\mathbf{d}|\mathbf{m}) f(\mathbf{m}) \\ & \propto \exp\Big((\mathbf{G}\mathbf{m} - \mathbf{d}_{\mathrm{obs}})^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} (\mathbf{G}\mathbf{m} - \mathbf{d}_{\mathrm{obs}})\Big) \cdot \exp\Big((\mathbf{m} - \mathbf{m}^{\mathbf{pr}})^{\mathrm{T}} \mathbf{C}_{\mathrm{M}}^{-1} (\mathbf{m} - \mathbf{m}^{\mathbf{pr}})\Big) \end{aligned} \tag{4.4}$$

It is obvious that the posterior distribution has a Gaussian form. The maximization of posterior is equivalent to the minimization of the exponent of equation 4.4. it results that:

$$\widehat{\mathbf{m}} = \mathbf{m}^{\mathrm{pr}} + \mathbf{C}_{\mathrm{M}} \mathbf{G}^{\mathrm{T}} (\mathbf{G}\mathbf{C}_{\mathrm{M}} \mathbf{G}^{\mathrm{T}} + \mathbf{C}_{\mathrm{D}})^{-\mathbf{1}} (\mathbf{d}_{\mathrm{obs}} - \mathbf{G}\mathbf{m}^{\mathrm{pr}}) \tag{4.5}$$

A basic difference from the formulation of section 3.3.2 is that the regularized term (the second term of the objective function) in equation 3.30 acts as a correction term in the estimation of least squares, while in the Bayesian framework the response of the system is

evaluated in a probabilistic form. Therefore, the ridge estimation method (regularized least square method) and MAP method do not necessarily yield the same estimates for the parameter set.

4.1.2 Non-linear theory with Gaussian prior

In non-linear cases where the likelihood can be considered locally Gaussian, the model function is linearized around the $\mathbf{m}^{\mathrm{pr}}$ parameter set. The posterior distribution is approximately Gaussian and its maximization point is given by:

$$\widehat{\mathbf{m}} = \mathbf{m}^{\mathrm{pr}} + (\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G} + \mathbf{C}_{\mathrm{M}}^{-1})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\mathbf{m}^{\mathrm{pr}})\right) \tag{4.6}$$

In the above estimation we assumed that the distance between the mean of the posterior and the mean of the prior distribution is sort enough so as the objective function between these points can be considered linear, however, this is not often the case. Under these circumstances, the estimation should be done iteratively. In a Bayesian updating scheme, the previous estimation that maximizes the posterior, together with its covariance, becomes the new mean and covariance of the prior distribution. Then we can replace the usual Bayes' rule

$$posterior \propto \mathrm{prior} \cdot likelihood$$

with

$$revised \propto current \cdot new\ likelihood$$

The $i+1$ estimation of the iterative process is:

$$\widehat{\mathbf{m}}^{i+1} = \widehat{\mathbf{m}}^{i} - \left(\mathbf{G}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G}_i + \mathbf{C}_{\mathrm{M}}^{-1}\right)^{-1}\mathbf{G}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}\left(\widehat{\mathbf{m}}^{i}\right)\right) \tag{4.7}$$

where $\mathbf{g}\left(\widehat{\mathbf{m}}^{i}\right)$ is the response of $\widehat{\mathbf{m}}^{i}$ parameter set.

However, if the linearization between $\widehat{\mathbf{m}}^{i+1}$ and $\widehat{\mathbf{m}}^{i}$ is not acceptable but the objective function is still quasi-linear, the right strategy is to obtain an intermediate parameter set $\widetilde{\mathbf{m}}^{i+1}$ for which the linearization will still be acceptable. A possible $\widetilde{\mathbf{m}}^{i+1}$ may be found by the maximization of the likelihood $L(\mathbf{d}_{\mathrm{i}}; \mathbf{m})$ alone (Tarantola, 1994, pp.75). The posterior distribution can be written:

$$f(\widehat{\mathbf{m}}^{i+1}|\mathbf{d}) \propto \exp\Big(\big(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}(\widehat{\mathbf{m}}^{i} - \widetilde{\mathbf{m}}^{i+1}) - \mathbf{d}_{\mathrm{obs}}\big)^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\big(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}(\widehat{\mathbf{m}}^{i} - \widetilde{\mathbf{m}}^{i+1}) - \mathbf{d}_{\mathrm{obs}}\big)\Big) \cdot \exp\Big(\big(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\big)^{\mathrm{T}}\mathbf{C}_{\mathrm{M}}^{-1}\big(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\big)\Big) \tag{4.8}$$

The next estimation can then be found by the minimization of the exponent (see Appendix A.10):

$$\widehat{\mathbf{m}}^{i+1} = \widehat{\mathbf{m}}^{i} + \big(\widetilde{\mathbf{G}}_{i}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}} + \mathbf{C}_{\mathrm{M}}^{-1}\big)\widetilde{\mathbf{G}}_{i}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1} + \Big(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) - \widetilde{\mathbf{G}}_{i}\big(\widehat{\mathbf{m}}^{i} - \widetilde{\mathbf{m}}^{i+1}\big)\Big) \tag{4.9}$$

The above equation is also written (see Appendix A.10):

$$\widehat{\mathbf{m}}^{i+1} = \widetilde{\mathbf{m}}^{i+1} + \big(\widetilde{\mathbf{G}}_{\mathrm{i}}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}} + \mathbf{C}_{\mathrm{M}}^{-1}\big)^{-1}\Big(\widetilde{\mathbf{G}}_{\mathrm{i}}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\Big(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}\big(\widehat{\mathbf{m}}^{i}\big)\Big) + \mathbf{C}_{\mathrm{M}}^{-1}(\widehat{\mathbf{m}}^{i} - \widetilde{\mathbf{m}}^{i+1})\Big) \tag{4.10}$$

4.2 Sampling the Posterior distribution

4.2.1 Monte Carlo Methods

In most cases, the probability distribution in the model space is not so simple and analytic techniques cannot be used to characterize it. Then, the more general way for solving the inverse problem would be the systematic exploration of the model space comparing all possible combinations of parameters, either when the objective is to sample the model space of posterior distribution or to find the set of parameters optimizing the forward solution. But generally, this practice is not applicable due to the computing and time costs, especially in high dimensional problems. Even if this was feasible, other problems would arise due to the non-uniqueness of the solution. Other consequences of ill-posedness could be that a solution would not exist or would be instable with regard to small variations in the input data (Carrera and Neuman, 1986).

Instead of taking all possible parameter combinations, we can sample the model space randomly. These practices are generally called Monte Carlo methods. The use of Monte Carlo methods for the solution of inverse problems was initiated by Keilis-Borok and Yanovskaya (1967) and Press (1968 and 1971) and concerned only uniform sampling of a model space, without taking a Bayesian approach (i.e. the conditional distribution between the previous and the current realization is not taken into account). In an optimization context, Burton and

Kennett (1972), Worthington et al. (1972 and 1974), Biswas and Knopoff (1974), Goncz and Cleary (1976), Burton (1977), Mills and Fitch (1977), Jones and Hutton (1979), Ricard et al. (1989), Jestin et al. (1994) and Kennett (1998) used uniform random search techniques in geophysical inversion, again without taking into consideration the Bayesian principles. In general, the use of Monte Carlo methods in well-designed random explorations allows avoiding entrapment in local maxima and thus solving complex optimization problems.

Yet, in complex prior model distributions with large-dimensional spaces, uniform sampling is insufficient because of their "emptiness". In these cases, the process can be improved by importance sampling, rejection sampling, or McMC methods like the Metropolis algorithm and Gibbs sampling.

4.2.2 McMC Methods

As previously mentioned, large-dimensional spaces may be under-sampled due to their emptiness. If not appropriately modified, Monte Carlo methods sparsely sample the model space and, thus, the local maxima of the posterior distribution (Mosegaard and Tarantola 1995). McMC techniques preferentially visit the model space where the posterior density is high. The basic idea behind these methods is to perform a random walk that normally would sample some prior probability distribution and then, using a probabilistic rule to modify the walk by accepting or rejecting samples in such a way that the produced samples are representative of the target distribution.

In inversion problems, McMC techniques generate candidate models from the prior distribution and use an acceptance criterion to reject or not the candidate model, under the consideration of a likelihood function. A candidate model $\mathbf{m}^*$ in a Markov Chain is generated by modifying the previous model $\mathbf{m}^i$, after addition of a random perturbation. The asymptotic behavior of a Markov Chain is governed by its transition kernel, that is, the probability density function of the transition from a model $\mathbf{m}^i$ to a new model $\mathbf{m}^*$.

4.2.2.1 Metropolis- Hastings algorithm

The Metropolis algorithm (Metropolis et al., 1953) is the mostly used McMC method that allows the analysis of nonlinear inverse problems with complex prior information and data with an arbitrary noise distribution. In the Metropolis algorithm, the transition probability from a

model $\mathbf{m}^i$ to a new model $\mathbf{m}^{i+1}$ is symmetric $\left(f\left(\mathbf{m}^{i+1}\middle|\mathbf{m}^i\right)=f\left(\mathbf{m}^i\middle|\mathbf{m}^{i+1}\right)\right)$ resulting to a simpler acceptance ratio.

Hastings (1970) proposed a more general form of Metropolis algorithm where the transition probability does not have to be symmetric. In the Metropolis-Hastings algorithm, the distribution of the conditional variable $\mathbf{m}^{i+1}|\mathbf{m}^i$ is considered known or it is feasible to generate a realization of that distribution. Early examples of determination of posterior probabilities for an inverse problem by means of Metropolis sampler are given by Boiden Pedersen and Knudsen (1990) and Koren et al. (1991). Using the Metropolis-Hastings algorithm in high dimensional spaces, it is difficult to draw samples from a suitable proposal distribution and, thus, a large proportion of the proposed samples are being rejected. The acceptance ratio to approve a proposed sample $\mathbf{m}^*$ as the next member $\mathbf{m}^{i+1}$ of a Markov chain is:

$$P(\mathbf{m}^*)=\frac{f\left(\mathbf{m}^i\middle|\mathbf{m}^*\right)f(\mathbf{m}^*)}{f(\mathbf{m}^*|\mathbf{m}^i)f(\mathbf{m}^i)} \tag{4.11}$$

This ratio is compared to a random number *u* from Uniform distribution ***U***(0,1) and if the ratio is larger than *u*, the proposed sample becomes the next member of Markov chain, otherwise, the transition does not take place. Accepting a candidate in a Markov chain means that this parameter set is a realization of the unknown posterior distribution

In an inverse problem where the solution of a forward problem takes place, the acceptance ratio to approve a proposed sample $\mathbf{m}^*$ to be the next member $\mathbf{m}^{i+1}$ of a Markov chain can be rewritten as:

$$\begin{aligned} P(\mathbf{m}^*) &= \frac{f\left(\mathbf{m}^*\middle|\mathbf{m}^i,\mathbf{d}\right)}{f(\mathbf{m}^i|\mathbf{m}^*,\mathbf{d})}=\frac{f\left(\mathbf{m}^i,\mathbf{d}\middle|\mathbf{m}^*\right)f(\mathbf{m}^*)}{f(\mathbf{m}^*,\mathbf{d}|\mathbf{m}^i)f(\mathbf{m}^i)} \\ &= \frac{f\left(\mathbf{m}^i\middle|\mathbf{m}^*\right)f(\mathbf{d}|\mathbf{m}^*)f(\mathbf{m}^*)}{f(\mathbf{m}^*|\mathbf{m}^i)f(\mathbf{d}|\mathbf{m}^i)f(\mathbf{m}^i)}=\frac{f\left(\mathbf{m}^i\middle|\mathbf{m}^*\right)f(\mathbf{m}^*|\mathbf{d})}{f(\mathbf{m}^*|\mathbf{m}^i)f(\mathbf{m}^i|\mathbf{d})} \end{aligned} \tag{4.12}$$

Mosegaard and Tarantola (1995) propose that if the candidates can be generated from the prior distribution directly without necessarily evaluating $f\left(\mathbf{m}^i\middle|\mathbf{m}^*\right)$ anywhere, the acceptance ratio can be written:

$$P(\mathbf{m}_*)=\frac{f(\mathbf{d}|\mathbf{m}^*)}{f(\mathbf{d}|\mathbf{m}^i)}=\frac{L(\mathbf{m}^*)}{L(\mathbf{m}^i)} \tag{4.13}$$

The above ratio allows always the transition from $\mathbf{m}^i$ to $\mathbf{m}^*$ when the likelihood of $\mathbf{m}^*$ is larger than the likelihood of $\mathbf{m}^i$, in a different case the candidate sample is accepted with probability $P(\mathbf{m}^*)$ as a member of the Markov chain. They also use probabilistic rules in order to gradually perturb a current model to another one, while the actual shape of prior distribution need not be known. However, the incorporation of geostatistical information of the parameter set helps to reproduce sequential candidates from the prior distribution and, thus, to create a McMC without evaluating the probability of transition.

4.2.2.2 Gibbs Sampler

Geman and Geman (1984) developed the Gibbs sampler in image reconstruction, another popular McMC sampling technique which avoids proposing samples that are likely to be rejected, while at the same time is better in means of computational cost. In Gibbs sampler, the proposed sample is generated using parameter information from current and previous states.

Gibbs sampler is often used in geostatistical simulation to reproduce realizations that preserve the spatial structure of a RF. For example, it is used in Plurigaussian simulation as it was referred in section 2.4.3. In this manner, the iterative algorithm of Gibbs sampler is written as:

$$
\begin{aligned}
&\mathbf{m}_1^{i+1} \sim f\left(\mathbf{m}_1^{i+1} | \mathbf{m}_2^i, \dots, \mathbf{m}_n^i\right) \\
&\mathbf{m}_2^{i+1} \sim f\left(\mathbf{m}_2^{i+1} | \mathbf{m}_1^{i+1}, \mathbf{m}_3^i, \dots, \mathbf{m}_n^i\right) \\
&\vdots \\
&\mathbf{m}_j^{i+1} \sim f\left(\mathbf{m}_j^{i+1} | \mathbf{m}_1^{i+1}, \dots, \mathbf{m}_{j-1}^{i+1}, \mathbf{m}_{j+1}^i, \dots, \mathbf{m}_n^i\right) \\
&\vdots \\
&\mathbf{m}_n^{i+1} \sim f\left(\mathbf{m}_n^{i+1} | \mathbf{m}_1^{i+1}, \dots, \mathbf{m}_{n-1}^{i+1}\right)
\end{aligned}
\tag{4.14}
$$

where the index symbolizes the number of iteration (state) and the exponent identifies the discretized parameters. Following a random path of subunits, different realizations are reproduced.

4.2.2.3 Dependent samples in McMC

A disadvantage of McMC is that the samples are correlated and they do not correctly reflect the posterior distribution, therefore the convergence of the chain must be reached before performing any sampling to reproduce samples of the posterior distribution. Another option to create independent samples is to make large jumps from one to another member of the Markov chain. In the usual case where the number of model parameters and thus the number of manifold dimensions is high, a systematic exploration of the prior model space is not possible because the number of required points grows too rapidly with the dimension of the space, while most of these models will have near zero probability. Therefore, sampling has to be done by making small jumps, using a random walk. However, this kind of sampling does not produce independent samples. Then, if the samples of the prior distribution presented to the Metropolis algorithm are not independent, the samples of the posterior distribution produced will not be independent too (Tarantola, 2005, pp. 52). To overcome this problem, Mosegaard and Tarantola (1995) suggest keeping only one model every μ samples as it is explained in section 5.1.2.

5 Using informative priors in facies inversion: The case of C-ISR method.

5.1 The McMC structure in C-ISR method

5.1.1 Kernel Transition: Iterative spatial resampling

Different transition kernels have been proposed in previous studies to generate Markov chains applied to spatially dependent variables. Oliver et al. (1997) create a McMC walk by updating one grid node of a simulated realization at each step. Fu and Gomez-Hernandez (2008) improve the efficiency of the method by updating many grid nodes at the same time, introducing the Blocking McMC (**BMcMC**) method that induces local perturbations by successively re-simulating a whole block of the realization.

Mariethoz et al. (2010a) propose a similar algorithm to Gibbs sampler and they call it Iterative Spatial Resampling (**ISR**). Hansen et al. (2012) present the theoretical background for using the method and argue that ISR is a special type of Sequential Gibbs sampler. The idea behind the ISR is to condition the j subunit in $i+1$ iteration only with a particular random sample of subunits $(1,\dots,v)$ of previous state i. Thus, the j query value of subunit using ISR algorithm is:

$$\mathbf{m}_j^{i+1} \sim f\left(\mathbf{m}_j^{i+1} | \mathbf{m}_1^i, \dots, \mathbf{m}_v^i\right) \tag{5.1}$$

Hansen et al. (2008) and Hansen et al. (2012) use a more general approach which can be expressed as:

$$\mathbf{m}_j^{i+1} \sim f\left(\mathbf{m}_j^{i+1} |, \mathbf{m}_1^{i+1}, \dots, \mathbf{m}_{j-1}^{i+1}, \mathbf{m}_1^i, \dots, \mathbf{m}_v^i\right) \tag{5.2}$$

In fact, in sampling the posterior distribution, the defined transition probability of Gibbs sampler can be used in a chain to generate candidates from the previous accepted member, while the sequential simulation is used. Hansen et al. (2008), Mariethoz et al. (2010a) and Hansen et al. (2012) agree to perturb the next state model by using a subset of model parameters r of the current state in order to generate the next candidate state. In case of ISR, the transition probability $f\left(\mathbf{m}^{i+1} \middle| \mathbf{m}^i\right)$ can be decomposed as the product of a series of one dimensional conditional densities:

$$f\left(\mathbf{m}^{i+1}\middle|\mathbf{m}^{i}\right) = f_{1|\nu}(\mathbf{m}_1^{i+1}|\mathbf{m}_1^{i}, \dots \mathbf{m}_\nu^{i}) \cdots f_{n|\nu}(\mathbf{m}_n^{i+1}|\mathbf{m}_1^{i}, \dots \mathbf{m}_\nu^{i}) \tag{5.3}$$

Also, another type of conditional re-simulation in the framework of McMC algorithm is the work of Irving and Singha (2010) which integrates different sources of information to solve the inverse problem.

In this work, since the aim is to condition on the state measurements, we use ISR to create a chain of dependant realizations by obtaining a random subset r_i of each member model $\mathbf{m}^i$ and impose it as conditioning data to generate the next candidate model $\mathbf{m}^*$. Representing the parameter set **m** by a RF $\mathbf{K}(\mathbf{x})$, the perturbation mechanism $f\left(\mathbf{m}^*\middle|\mathbf{m}^i\right)$ of the chain works as follows:

1. Generate an initial model $\mathbf{m}^1 = \{\mathbf{k}_1^1(\mathbf{x}_1), \dots, \mathbf{k}_n^1(\mathbf{x}_n)\}$ as a realization of the RF $\mathbf{K}(\mathbf{x})$ discretized on a grid with *n* nodes, using a geostatistical simulation algorithm and evaluate its likelihood $L(\mathbf{m}^1)$.
2. Iterate on i:
   a. Select randomly a subset $r_i = \left\{\mathbf{k}_1^i(\mathbf{x}_1), \dots, \mathbf{k}_r^i(\mathbf{x}_\nu)\right\}$ of the previous model, where ν is the number of conditioning data to generate the next candidate model $\mathbf{m}^*$.
   b. Generate a proposal realization $\mathbf{m}^*$ using the same geostatistical simulation and under the conditioning data r_i.
   c. Evaluate $L(\mathbf{m}^*)$.
   d. Accept or reject the candidate model $\mathbf{m}^*$. If accept, set $\mathbf{m}^{i+1} = \mathbf{m}^*$.

The performance of the method depends on the criterion for candidate model selection and the time of chain interruption. The number of conditioning data should be enough to permit a certain dependency between two members of the chain, but this number cannot be too high, in order to avoid artifacts in the simulation. The measurements, if any, are added to the subset r in each iteration.

It is also noted that, even though in this work we apply ISR with two-point statistics, its principal is not associated with a specific simulation method or a certain type of spatial variability (Mariethoz et al. 2010a).

5.1.2 ISR for Optimization

As it is explained in the previous chapter, sampling the posterior using McMC techniques raises two important issues:

a) the criterion to accept a state model as a member of a chain

b) and the criterion to accept a model as a member of the posterior ensemble.

Mosegaard and Tarantola (1995) propose the acceptance ratio of a chain member can be represented by equation 4.12. Concerning the second issue, they suggest keeping only one model every μ samples. In this way, they ensure uniform sampling of posterior distribution and thus their modified algorithm can also be used for finding the optimal solution of an inverse problem. Since there are samples available from the posterior distribution, one can find between the samples, the parameter set that maximizes this distribution. Yet, Hansen et al. (2012) explain that the Sequential Gibbs sampler cannot be used for optimization problems in a Bayesian sense, because the actual prior probability of a given model cannot be computed and evaluated, thus the samples maximize the likelihood and not the posterior distribution.

Mariethoz et al. (2010a) propose an alternative procedure for optimization by applying ISR and using independent Markov chains. Their algorithm accepts a model in a chain only if his likelihood is larger than the likelihood of previous model:

$$L(\mathbf{m}^*) \geq L(\mathbf{m}^i) \tag{5.4}$$

This tactic leads to having samples biased toward high fits and thus coming from a more biased prior $f^+(\mathbf{m})$ in each iteration, which is nevertheless a subset of the prior $f(\mathbf{m})$. As a bias correction, they propose to interrupt a chain in a stochastic criterion inspired by the rejection sampling method (Von Neumann, 1951) with probability

$$P(\mathbf{m}^*) = \frac{L(\mathbf{m}^*)}{L(\mathbf{m})_{\max}} \tag{5.5}$$

where $L(\mathbf{m})_{\max}$ denotes the highest likelihood value (constant). The interruption control takes place in each iteration and the probability to interrupt the chain increases as the likelihood of the candidate model increases. Therefore, a second bias is introduced but in opposite direction.

The equation 5.5 essentially describes an "accept only improvements" process. The resulting chain is a stochastic search for a single calibrated model. Taking independent chains, the global maximum will eventually be reached after an infinite number of iterations.

The main difference between the two approaches is that Mosegaard and Tarantola (1995) propose an algorithm whose purpose is to generate a 'sample' of the posterior distribution. Mariethoz et al. (2010a) propose a method that generates a selection of models that fit the data and the prior information. Then, this collection of models provides a good approximation to the posterior. In an optimization context, this mechanism performs an importance sampling effect, since the objective is to find a realization that maximizes the posterior distribution and not to sample it.

5.1.3 Likelihood Function

When sampling the posterior or searching for an optimal solution, it is necessary to define a criterion of goodness of fit, either to accept a model in a chain or as a stopping mechanism for search algorithms. The criterion of goodness of fit may be the likelihood function as it is expressed in equation 3.24 (L_2 norm). In groundwater literature, other norms have also been used. For example, errors can be quantified as the absolute value of the difference between measured and computed values (L_1 norm). Furthermore, a regularized function (see chapter 2) can be used.

In this work, we use the L_2 norm as a measure of goodness of fit which can be also expressed as:

$$L(\mathbf{m}) \propto \exp\left(\frac{1}{p\sigma^2}\sum_{j=1}^{p}\left(\mathbf{g}_j(\mathbf{m}) - \mathbf{d}_j\right)^2\right) \tag{5.6}$$

where p is the number of measurements of the response variable, $\mathbf{g}_j(\mathbf{m})$ is the solution of forward model at the points of measurements, σ^2 is the total "noise" variance and corresponds to the sum of epistemic and measurements errors. The variance is assumed to be the same for all p data values.

5.2 The C-ISR Method

5.2.1 The act of cosimulation

As seen in the previous sections, when inverting under the McMC framework, the measurements of a physical variable (e.g. hydraulic measurements) can only be used as indirect data to evaluate the prior models and drive the search path. This inconvenience is due to the nonlinearity relation between the measurements of physical variable and unknown parameters. On the other hand, well-known cokriging and cosimulation rely on a linear predictor approach and use covariance and cross-covariance functions derived from a first-order approximation, therefore they often result in unacceptable solutions when the multivariate distribution of integrated variables is not Gaussian (Yeh et al., 1995).

Under the framework of the C-ISR method, although the problem is not linear, we use cosimulation with the reference data as the auxiliary variable, in order to improve the search path in the Markov chains (Valakas and Modis 2015; Valakas and Modis 2016). More specifically, each proposed model $\mathbf{m}^*$ is produced by cosimulating the underlying Gaussian variable $\mathbf{Z}$, which represents the facies distribution, with the normal scores transformation $\mathbf{N}$ of the reference variable. The realizations so produced belong to a more confined set than in the original prior, that is, the prior described solely by the underlying Gaussian variogram. Our method relies on the approximation that, temporally, the $p \times 1$ vector $\mathbf{N}$ of normal scores transformed hydrologic measurements is related to $\mathbf{Z}$ assuming a linear coregionalization model. Thus, a narrower and more informative prior to that of equation 5.3 is created here by utilization of the reference variable, as seen in equation 5.7.

$$f\left(\mathbf{m}^{i+1}\middle|\mathbf{m}^{i},\mathbf{d}_{\text{obs}}\right) = \\ = f_{1|r}(\mathbf{m}_1^{i+1}|\mathbf{m}_1^{i},\dots\mathbf{m}_r^{i},\mathrm{d}_1,\dots,\mathrm{d}_p)\cdots f_{n|r}(\mathbf{m}_n^{i+1}|\mathbf{m}_1^{i},\dots\mathbf{m}_{\mathrm{r}}^{i},,\mathrm{d}_1,\dots,\mathrm{d}_p) \tag{5.7}$$

Moreover, the addition of cosimulation to the ISR algorithm gives the ability to relate the prior distribution to the reference variables, thus enforcing the generation of more probable models. This, in a way, enables accounting for the probability of a model. Considering the limitations of ISR algorithm in section 5.1.2 (Hansen et al. 2012), the C-ISR method with its narrowing effect can actually be used as a Bayesian optimization tool.

5.2.2 The C-ISR algorithm

In order to obtain one sample by an interrupted Markov chain, we design an ever-improving Markov chain that accepts new members under condition 5.4. The chain should be interrupted following the stochastic stopping criterion 5.5. Symbolizing with $\boldsymbol{\xi}(\cdot)$ the function returns the discretized variable (facies), the C-ISR algorithm is accomplished in the following steps:

1. Generate an initial model $\mathbf{m}^1 = \left\{\boldsymbol{\xi}\left(\mathbf{z}_1^1(\mathbf{x}_1)\right), \dots, \boldsymbol{\xi}\left(\mathbf{z}_n^1(\mathbf{x}_n)\right)\right\}$ as a realization of the RF **Z**(**x**), discretized on a grid with *n* nodes, using a geostatistical cosimulation with the normal scores transformed reference variable and evaluate its likelihood $L(\mathbf{m}_1)$.
2. Iterate on i:
   a. Select randomly a subset $r_i = \left\{\boldsymbol{\xi}\left(\mathbf{z}_1^1(\mathbf{x}_1)\right), \dots, \boldsymbol{\xi}\left(\mathbf{z}_r^1(\mathbf{x}_r)\right)\right\}$ of the previous model, where r is the number of conditioning data to generate the next candidate model $\mathbf{m}^*$.
   b. Generate a proposal realization $\mathbf{m}^*$ using the same geostatistical cosimulation and under the conditioning data r_i.
   c. Evaluate $L(\mathbf{m}^*)$.
   d. If $L(\mathbf{m}^*) \geq L\left(\mathbf{m}^i\right)$ accept the candidate model $\mathbf{m}^*$ and set $\mathbf{m}^{i+1} = \mathbf{m}^*$.
   e. Decide whether or not to interrupt the chain:
      i. Compute $P(\mathbf{m}^*) = L(\mathbf{m}^*)/L(\mathbf{m})_{\max}$
      ii. Draw *u* in ***U***[0,1]
      iii. If u ≤ $P(\mathbf{m}^*)$ interrupt the chain, else continue the chain using $\mathbf{m}^{i+1}$

Hence, the above means of successive linearizations is used to transfer from one model to another at each step of the optimization procedure in every Markov chain. The approach works by creating an importance sampling effect (Stordal and Elsheikh, 2015) that steers the process to selected areas of the prior and thus improving convergence. Therefore, by the incorporation of indirect data to narrow the prior distribution, our approach promises to allow the full utilization of measurements in achieving the best possible site characterization. The effectiveness of the proposed formulation is shown next, using a synthetic example.

5.3 Validation of the C-ISR algorithm

5.3.1 Materials and methods

The effectiveness of cosimulation is demonstrated by means of a synthetic aquifer example using a pump test. A simple two-dimensional synthetic flow system is set up for that purpose and finite element software package COMSOL™ 3.4 (COMSOL Multiphysics User's Guide, 2005) being controlled via a MATLAB™ 7.8 script, has been employed. We consider a square zone of side 100 m in a confined 2D aquifer, with a given spatial distribution of four distinct facies discretized in 100x100 nodes, as shown in Figure 5.1a. The hydraulic conductivity values of $10^{-3.5}$ (m/s), $10^{-4.5}$ (m/s), $10^{-5.5}$ (m/s) and $10^{-6.5}$ (m/s) are assigned to facies A, B, C and D, respectively. Using this spatial structure of facies as the reference, we can produce the hydraulic head observations required in our example. A pumping well injecting 0.001 (m^3/s) is set at the left lower corner of the field and another pumping well extracting 0.001 (m^3/s) at the right upper corner of the field (Figure 5.1b).

The hydraulic potential governing the flow through the aquifer zone and the surrounding area can be represented by the 2D pressure head $\mathbf{H}(\mathbf{K}(\mathbf{x}))$ distribution, which obeys Darcy's Law:

$\nabla(-\mathbf{K}(\mathbf{x})\nabla H) = Q \Rightarrow \nabla(-\xi(\mathbf{Z}(\mathbf{x}))\nabla H) = Q$ and $H = b_u(\mathbf{x})$ on model boundary **u**, where $\mathbf{K}(\mathbf{x})$ is the spatial hydraulic conductivity, $\mathbf{Z}(\mathbf{x})$ is a standard Gaussian RF representing facies distribution and Q (m^3/s) represents aquifer recharge or discharge. Considering the change in the head due to pumping and applying superposition principle, we derive the following equation for δH:

$\nabla(-\boldsymbol{\xi}(\mathbf{Z}(\mathbf{x}))\nabla\mathbf{H}) = q$ and $\delta H = 0$, on model boundary **u**,

where q represents sources and sinks due to pump test. The modeling domain's boundary **u** is assumed to be a square of side 12,000 m, where any effects of pumping are negligible. Also, the area outside the square zone is considered to have roughly constant hydraulic conductivity 10^{-5} (m/s). We used different mesh sizes inside and outside the central zone, COMSOL's normal size triangular mesh for the interior where the calculations need to be more accurate and coarse size triangular mesh for the exterior.

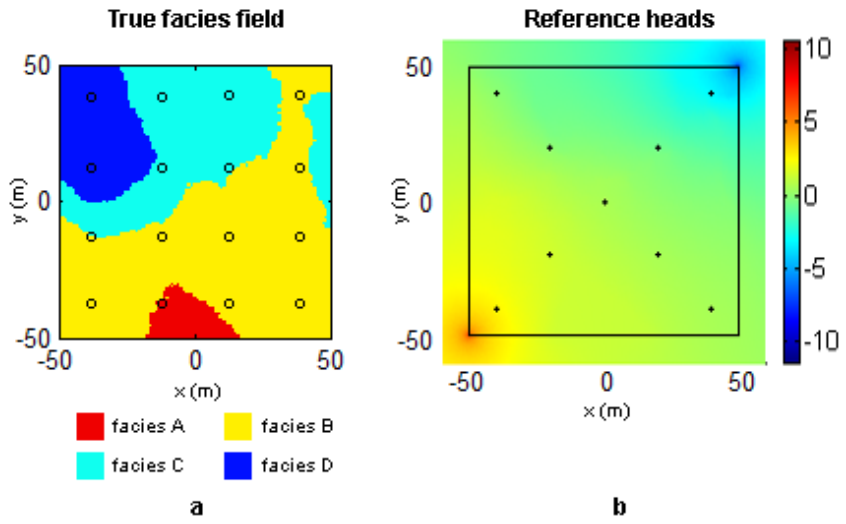


Figure 5.1: True facies distribution in aquifer zone and the facies observations with circles (left). The reference heads and head observations with black crosses (right).

Prior information on the aquifer structure consists of a set of 16 facies observations. Using this information together with the 9 head measurements as shown in Figure 5.1, we explored the performance of the C-ISR in comparison ISR using a classical geostatistical simulation, such as SIS. Firstly, we run 150 independent Markov chains and the results presented in Valakas and Modis (2015). We demonstrated that our method is faster, more reliable and explores more effectively the prior space avoiding entrapment in local maxima. Consequently, we compared our method with the ISR using TGS, a geostatistical simulation method more effective than SIS. This thesis presents only the results of the last comparison. For the TGS and cosimulation we use the algorithms by Emery (2007) and Emery and Silva (2009), respectively.

For both the C-ISR and ISR approaches, we ran 250 independent Markov chains. We used the likelihood function of equation 5.6 with σ = 0.05 m, which can reasonably correspond to the total head measurements error. The supremum value of L is set to 0.607, which, according to the above equation, corresponds to a RMSE of 0.05. Larger likelihood values might cause undue delays and seriously affect the performance of the process. The percentage of resampled nodes to generate the next candidate model $\mathbf{m}^*$ is set to 1% ($r = 10$).

5.3.2 Variography

For the needs of TGS, a standardized Gaussian variable **Z** with the characteristics of Figure 5.2 was defined. Using the proportions of each facies which were calculated based on the observations and the truncation rule which is assumed to be known a priori, an exponential model with a sill of 1 and isotropic range of 49 m was adopted for **Z** according to the procedure described in section 2.4.3.

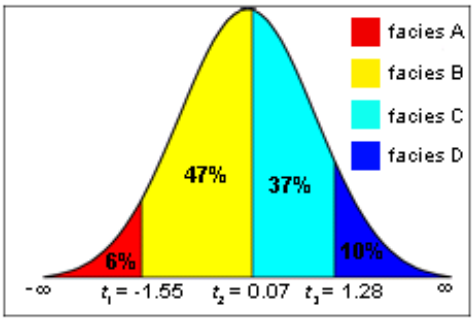


Figure 5.2: Truncation rule, showing contact relationship, proportions and Gaussian thresholds associated with the facies.

Considering the state variable, the head measurements were first transformed into Normal scores on which variogram analysis would be performed. A Gaussian model with a sill of 1 and isotropic range of 41.5 m was used to represent the transformed data.

The last step is to define a linear coregionalization model between the underlying Gaussian variable **Z** and Normal scores of head observations **N**. Since, as seen in Figure 5.1, we have an heterotopic data configuration, we krige the unknown locations of **N** and then we use the new dense set of data to infer the experimental cross-variogram. Finally, according to section 2.5.3 we obtain a coregionalization model of the form:

$$\begin{bmatrix} \gamma_{ZZ} & \gamma_{ZN} \\ \gamma_{ZN} & \gamma_{NN} \end{bmatrix} = \begin{bmatrix} 0.6949 & 0.1311 \\ 0.1311 & 0.0248 \end{bmatrix} \exp(49\ \mathrm{m}) + \begin{bmatrix} 0.3051 & -0.5451 \\ -0.5451 & 0.9752 \end{bmatrix} \mathrm{gaus}(41.5\ \mathrm{m}) \quad (5.8)$$

The experimental, model and cross-variogram of the underlying Gaussian variable and the transformed heads are displayed in Figure 5.3. A graphical comparison between true experimental variograms and variograms derived from the underlying Gaussian and the coregionalization models respectively is shown in Figure 5.4. It is apparent that the fitting is satisfactory.

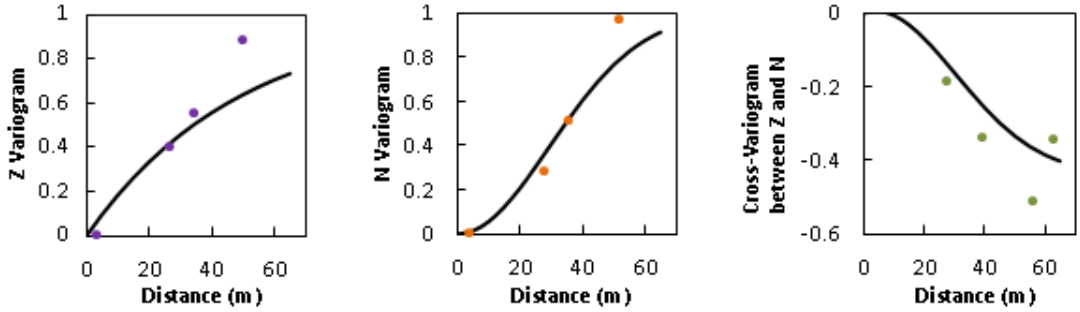


Figure 5.3: Sample (points) and linear model of coregionalization (solid lines) between underlying Gaussian variable **Z** and Normal scores transformed reference data **N**.

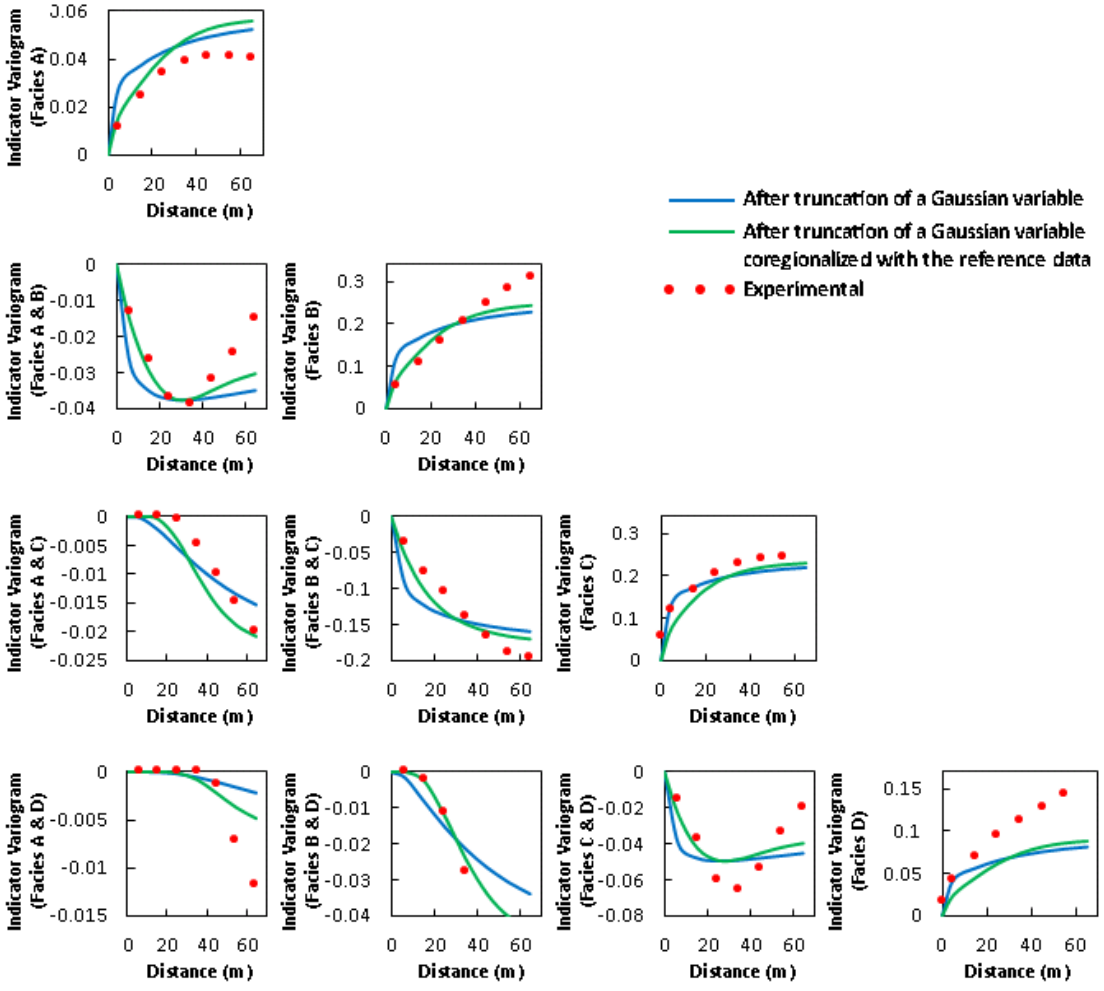


Figure 5.4: Experimental and model variograms of facies indicators.

5.3.3 Results and Discussion

The main advantage of C-ISR method is the reduction of computational and timing costs in order to approximate the a-posteriori distribution. More specifically, the forward model is solved 28914 times using C-ISR against 48390 runs using ISR. Figure 5.5 shows the evolution of 50 randomly selected optimizations under the two approaches. The average slope of the curves on the left is higher, showing that C-ISR converges faster. The black line indicates the convergence of the chain for which the optimal solution is achieved among the 250 independent chains for each method. Generally, the average number of evaluations (solution of forward model) for each chain is 116 for C-ISR versus 194 for ISR, while the average number of accepted models in a chain is 9.9 and 8.3 for C-ISR and ISR respectively. Furthermore, the likelihood of the posterior models produced by C-ISR is better, as seen from the ending points of the ensemble lines in Figure 5.5a, compared to those of Figure 5.5b.

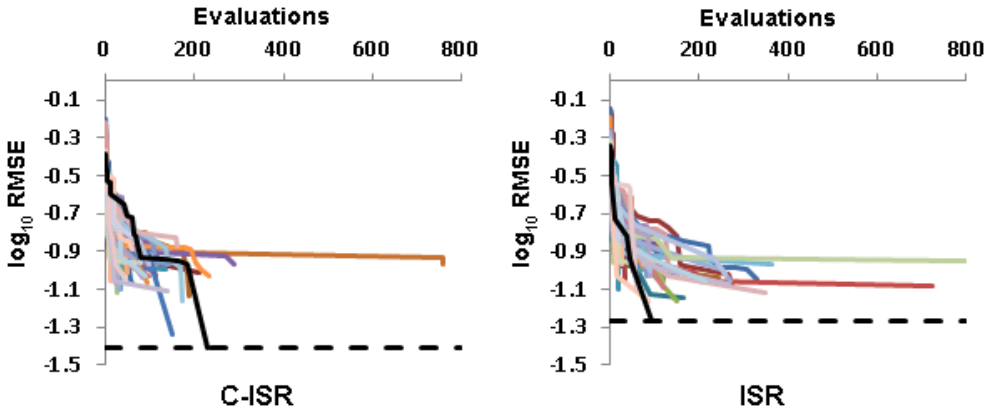


Figure 5.5: Convergence of 50 randomly selected individual optimizations for C-ISR (a) and ISR (b).

In order to demonstrate the precision of hydraulic head predictions, we set up Figure 5.6 showing the average RMSE (250 chains) for the two methods. The RMSE is calculated by the solution of forward model at the center of the blocks and the true spatial distribution of hydraulic head. It is apparent that C-ISR method produces more accurate hydraulic head predictions across the spatial field, while the RMSE average is higher at top-right corner for both methods. The distribution in Figure 5.7 shows that even if the two approaches produce on average similar RMSE, the incorporation of cosimulation in C-ISR helps to avoid entrapment to

local maxima, as seen by the right tail of ISR distribution. Also, the left tail of C-ISR distribution is longer, indicating that this method produces more possible solutions of forward problem.

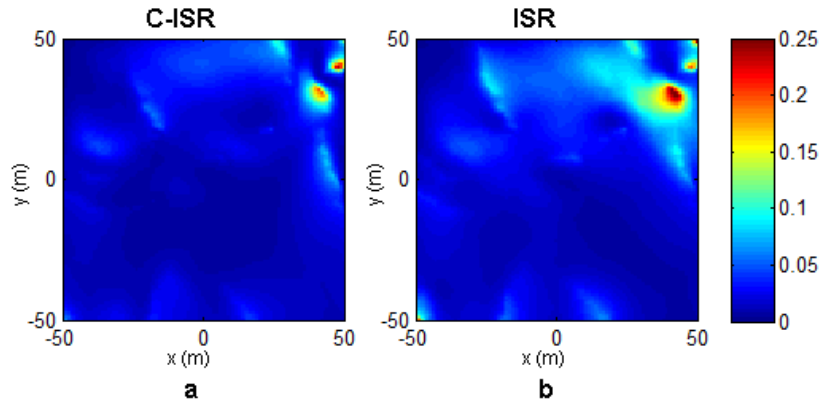


Figure 5.6: Average RMSE map for C-ISR (a) and ISR (b).

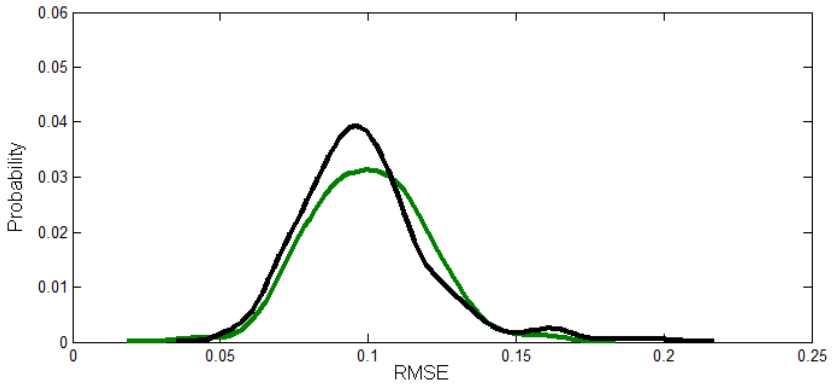


Figure 5.7: RMSE distribution of posterior samples for C-ISR (green) and ISR (black).

Both methods produced accepted realizations while keeping the spatial structure of reality for each facies, as shown by the experimental variograms of a random subset of the posterior ensembles in Figure 5.8. The reduced variability of models in the case of C-ISR reveals a narrower prior due to conditioning to heads. A post-processing of the posterior ensembles is shown in Figure 5.9. Comparing to Figure 5.1a, it is apparent that the probability of occurrence of each facies resembles more the reality in the case C-ISR method. This is an indication that the posterior space is properly covered.

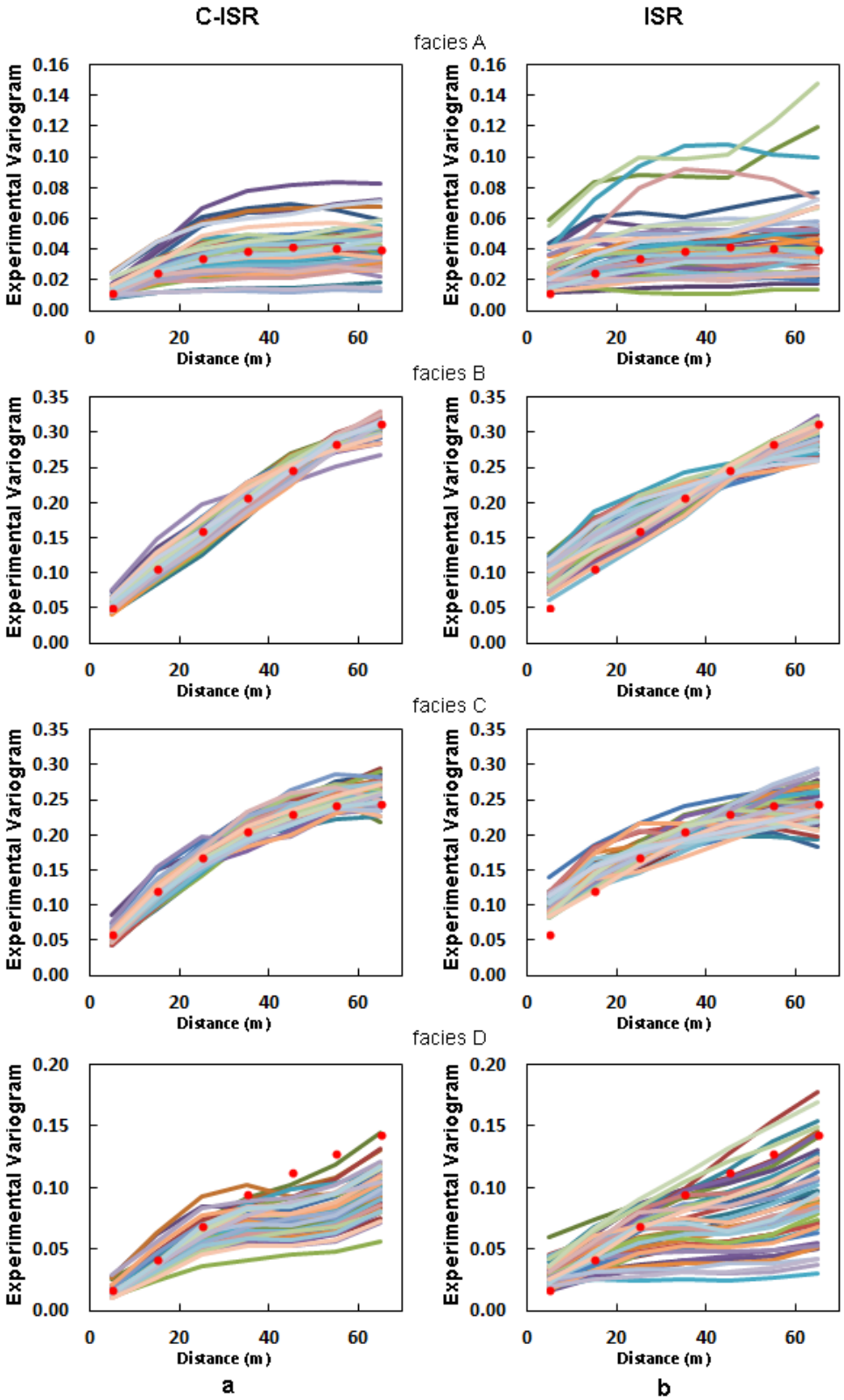


Figure 5.8: Experimental variogram of 50 accepted realizations for each facies in C-ISR (a) and ISR (b). The red dots show the experimental variogram of reality.

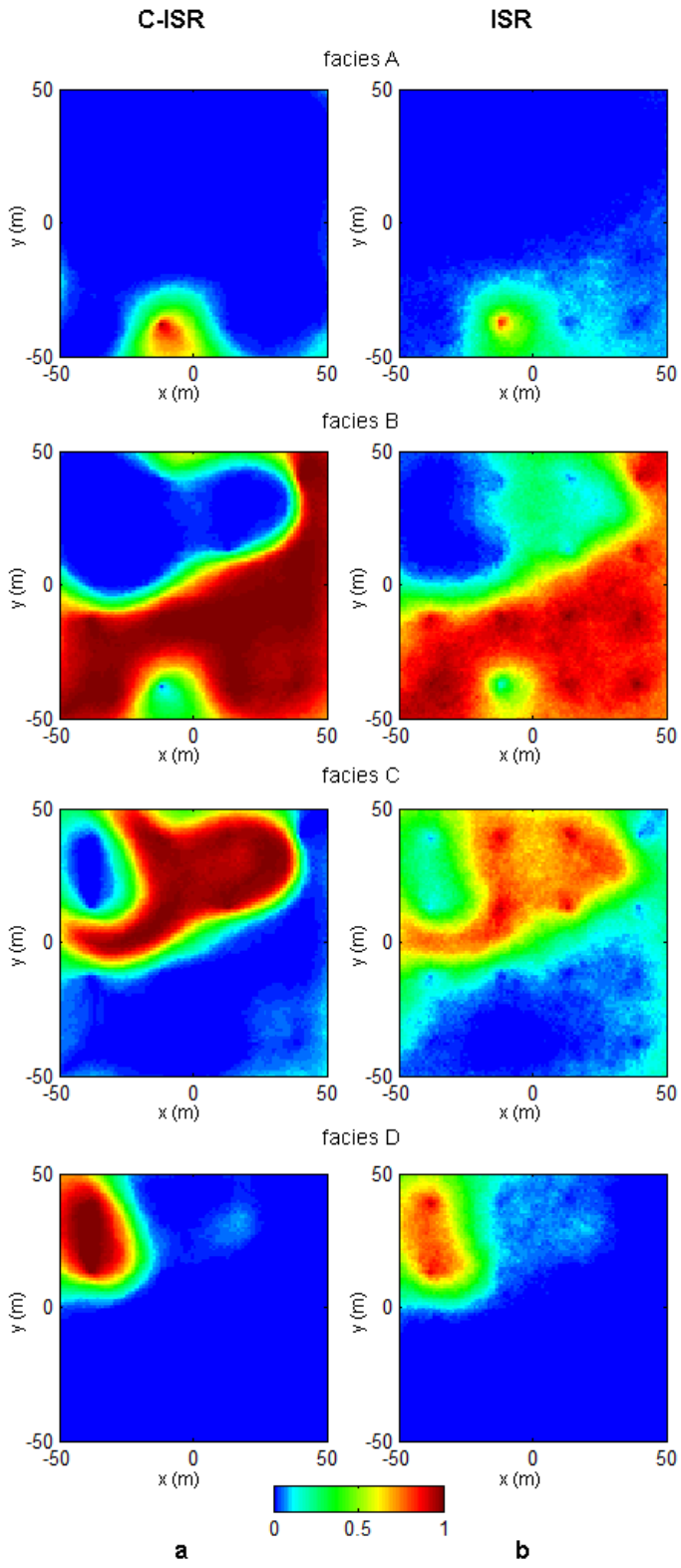


Figure 5.9: Probability of occurrence of the four facies in C-ISR (a) and ISR (b).

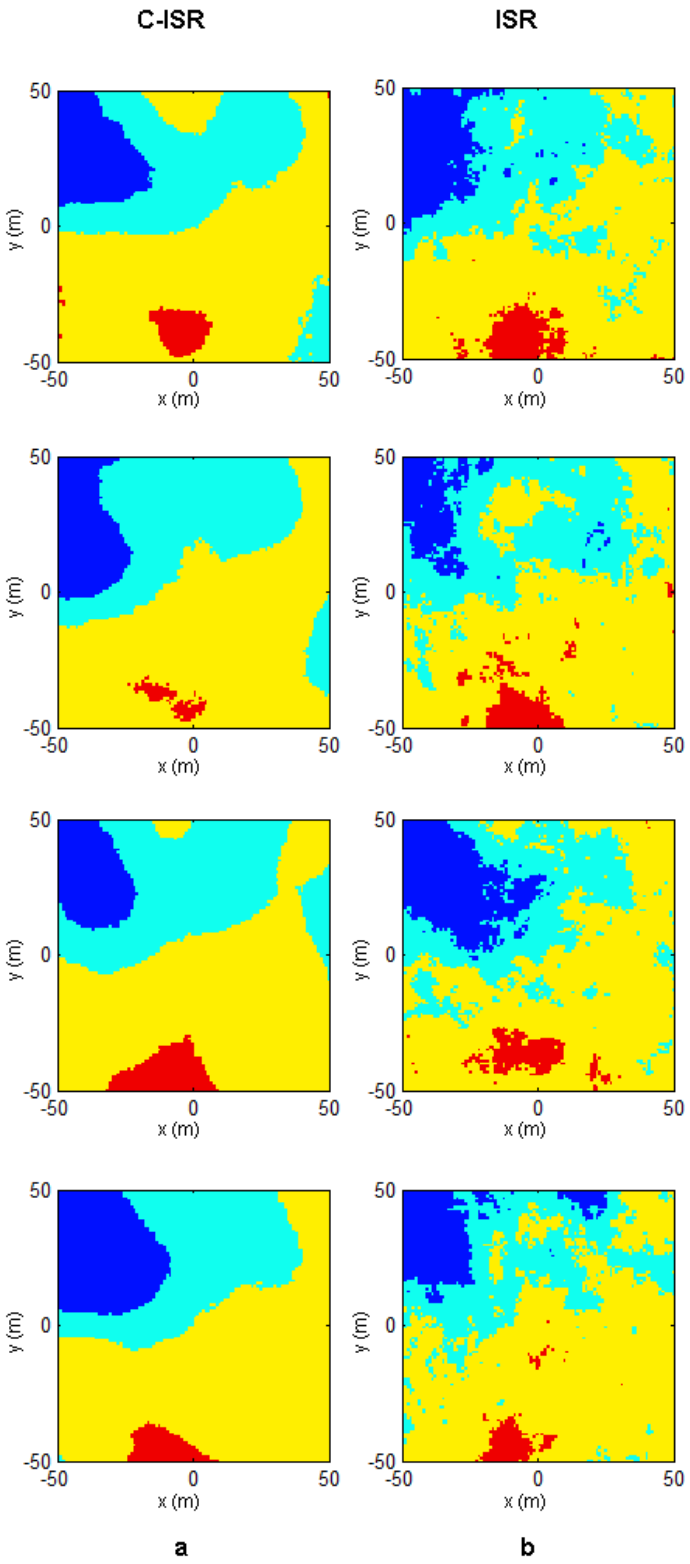


Figure 5.10: Accepted realizations for C-ISR (a) and ISR (b).

A set of realizations in both cases can be visualized in Figure 5.10. According to Figures 5.9 and 5.10, the C-ISR method produced posterior members which reveal the spatial homogeneity of reality in contrast with ISR. In order to evaluate the reliability of the two methods, we calculated the probability of positive hit for each block of spatial field. From 250 a-posteriori members, it results that C-ISR has more positive hits and it is more efficient in preserving the shapes of existing formations inside the investigated area, as it is depicted in Figure 5.11. The intense green color indicates probability close to one, while intense red color probability close to zero. C-ISR fails to indicate the facies C (cyan color in Figure 5.1a) at top-right corner with precision, but the probability of occurrence is not zero, as it is detected in Figure 5.9a for facies C. An explanation could be that the positive hits of ISR at the blocks where C-ISR fails, occur due to the excessive spatial heterogeneity induced by this method and thus its weakness to preserve the shapes of existing formations.

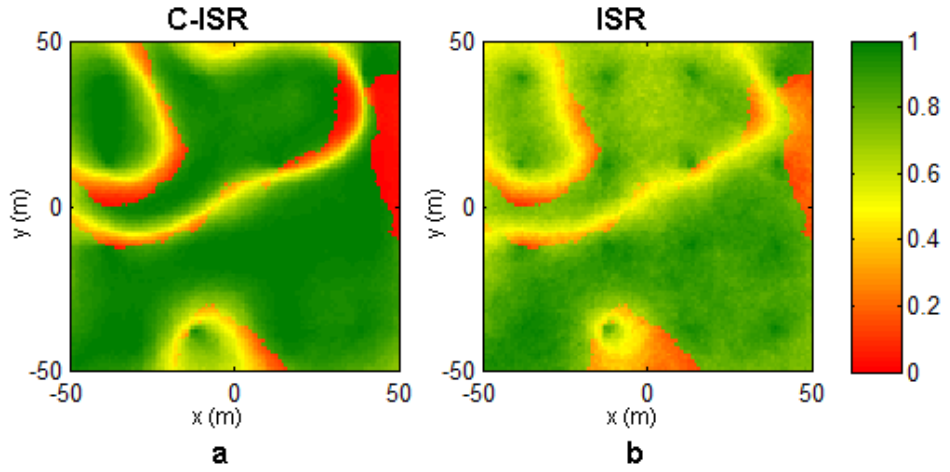


Figure 5.11: Probability of positive hit for C-ISR (a) and ISR (b). C-ISR preserves shapes better.

The main objective of the synthetic example was to demonstrate the properties of C-ISR and find the optimal solution of determined inverse problem. Figure 5.12 shows the best resulting models of facies distribution reached by C-ISR and ISR respectively, after the 250 individual optimizations. Compared to the true facies field (Figure 5.1a), C-ISR results in a much better approximation (85.69% similarity) than ISR (81.76% similarity). A brief summary of the two algorithms performance is shown in Table 5.1. Taking into consideration the evidence of the results, we summarize that the C-ISR method improves the search path of optimal solution

reducing the computational and timing costs, leads to more precise estimations of parameters, produces realizations which preserve the shapes of existing formations and results in an optimal solution closer to the reality.

	C-ISR	ISR
Min RMSE	0.0383	0.0537
Similarity ratio when min RMSE	0.8569	0.8176
Average similarity ratio	0.8028	0.7104
Total number of evaluations	28914	48390
Average number of evaluations for each chain	115.656	193.56
Average number of accepted models for each chain	9.9	8.3

Table 5.1: Summary performance of the two algorithms.

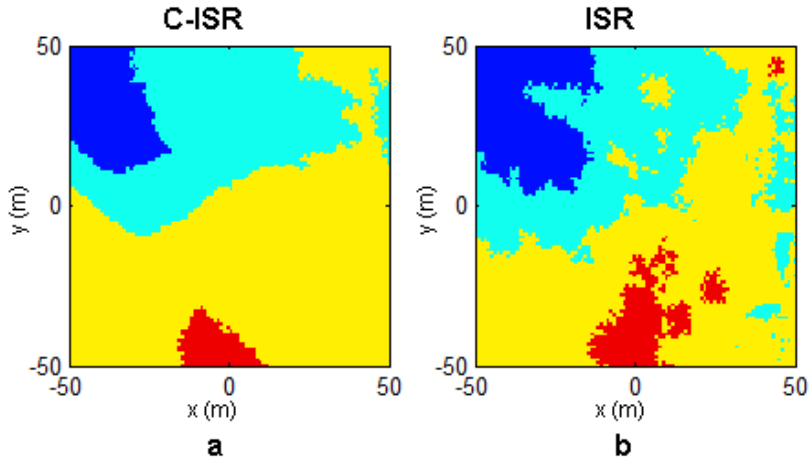


Figure 5.12: Optimal results for facies prediction from C-ISR (a) and ISR (b) after 250 individual optimizations. Similarity to true facies is 85.69% and 81.76% respectively.

6 Conclusions and future research

6.1 Conclusions

Aquifer characterization is a demanding task and requires modeling of physical properties such as saturation, porosity and permeability. In addition, geological formations (e.g. facies identification) should be taken into account before any physical forecasting. Deterministic and stochastic methods using a two-step approach in which, first, the geological facies are modeled, and second, they are populated with heterogeneous hydraulic parameters, have lower performance in general. Inverse methods, on the other hand, use more information on the reservoir behavior, such as the governing physical law and the system response variables.

Solving the inverse problem for hydrofacies is the right strategy before the aquifer characterization; however inverse modeling encounters important difficulties in implementation due to the ill-posedness and the complex spatial heterogeneity of hydrological systems. Also, practical issues arise either due to discretization or non-linearity when applying deterministic methods (e.g. gradient methods), or due to CPU time and speed performance when applying stochastic methods.

In this thesis, we developed a stochastic method (C-ISR) to solve the inverse problem and overcome the previously mentioned difficulties. Our method narrows the prior distribution of parameters and produces improved proposal realizations in McMC optimization under the Bayesian framework for inversion. This is achieved by using cosimulation with the system response data. The algorithm of C-ISR is based on ISR kernel and relies on the approximation that, temporally, the vector of normal scores transformed hydrologic measurements are related to the underlying Gaussian variable of the facies**,** assuming a linear coregionalization model. This process of successive linearizations acts as an importance sampling effect and speeds up convergence of the Markov chains.

The effectiveness of cosimulation is illustrated by a synthetic aquifer inversion example, using a pump test. We apply C-ISR vs. ISR based solely on TGS. The results of 250 individual optimizations for both approaches show that C-ISR needs less forward model runs and, as a

result, is faster. Also, it is more reliable since it produces smaller RMSE and explores more effectively the prior space avoiding entrapment in local maxima.

6.2 Research limitations

The performance of the C-ISR method has been evaluated in a realistic synthetic example supposing different degrees of prior information. Although the effectiveness of C-ISR has been demonstrated, more synthetic examples should be examined using different formulations. For instance, the C-ISR performance should be evaluated using more complicated types of contacts between facies. Moreover, it should be examined how the data availability (sampled facies or head measurements) and their spatial distribution affect the method.

The benefit of using synthetic data is that every possible geological structure could be represented. Nevertheless, synthetic examples cannot mimic reality. Though practically difficult, C-ISR should be evaluated in a real case study in order to investigate other advantages or disadvantages of the method.

Moreover, C-ISR was basically evaluated in an optimization context, assuming that it could be used also to sample the a-posteriori distribution of parameters in order to investigate more parameters properties space. However, this should be more thoroughly examined by applying C-ISR in a more focused algorithm, using for example the acceptance ratio of candidate models and keeping only one model every μ samples, as it is proposed by Mosegaard and Tarantola (1995) and it has been discussed in previous sections.

The limitations of the present research associate to the high computational and time cost that is demanded by the solution of forward model. The use of independent chains in order to find the optimal solution, gives the opportunity to solve the optimization problem in parallel computing. Therefore, we suggest the implementation of the C-ISR algorithm to be done in an open-source software environment, such as the R Project for Statistical Computing. The use of open-source software gives the ability to run the simulations on cloud computing platforms, such as Amazon's Web Services. Cloud computing offers flexible capacity, speed, agility and increases the computing performance, as the iterations can be run in parallel processors.

6.3 Suggestions for future research

Future research might focus on extending the concept of C-ISR to the direction of multiple point statistics (**MPS**). It has been shown (Mariethoz et al. 2010a) that the ISR kernel is not associated with a specific simulation method or a certain type of spatial variability. Hence, a natural extension would be to investigate the integration of cosimulation with the reference variable, in MPS methodology.

The MPS method has been recently developed mainly to simulate hydrofacies, without taking into account the solution of the inverse problem (e.g. Chugunova and Hu 2008; Feyen and Caers 2006; Huysmans and Dassargues 2009; Michael et al. 2010; Renard 2007). Most applications focus on the effectiveness of the method in relation to traditional techniques such as Indicator Kriging and Plurigaussian simulation. It has also been applied to solve the inverse problem (Alcolea and Renard 2010; Caers and Hoffmann 2006; Ronayne et al. 2008). Moreover, Mariethoz et al. (2010b) highlighted the benefits of the ISR method by solving the inverse problem in MPS. In this context, it is the first time where MPS is proposed for the solution of the inverse problem by taking into account the covariance between the parameters and the reference variable, as in the case of the C-ISR algorithm. This way, the Markov chain convergence rate is expected to increase, reducing time and computer cost, while the spatial distribution of the underground structures is expected to be more accurately represented than in all previous methods.

References

Alabert, F., 1987. The practice of fast conditional simulations through the LU decomposition of the covariance matrix. Mathematical Geology 19, 369-386. doi:10.1007/bf00897191

Alcolea, A., Renard, P., 2010. The blocking moving window sampler: Conditioning multiple point simulations to hydrogeological data. Water Resources Research 46, W08511. doi:10.1029/2009WR007943

Almeida, A. S., Frykman, P., 1994. Geostatistical modeling of chalk reservoir properties in the Dan Field, Danish North Sea, in Yarus, J. M., Chambers, R. L., Stochastic modeling and geostatistics. American Association of Petroleum Geologists, Tulsa, OK, pp. 273-286.

Armstrong, M., Galli, A., Beucher, H., Loc'h, G., Renard, D., Doligez, B., Eschard, R., Geffroy, F., 2011. Plurigaussian simulations in geosciences. Springer, New York.

Bahar, A., Kelkar, M., 2000. Journey From Well Logs/Cores to Integrated Geological and Petrophysical Properties Simulation: A Methodology and Application. SPE Reservoir Evaluation & Engineering 3, 444-456. doi:10.2118/66284-pa

Biswas, N., Knopoff, L., 1974. The Structure of the Upper Mantle under the United States from the Dispersion of Rayleigh Waves. Geophysical Journal International 36, 515-539. doi:10.1111/j.1365-246x.1974.tb00612.x

Boiden Pedersen, J., Knudsen, O., 1990. Variability of estimated binding parameters. Biophysical Chemistry 36, 167-176. doi:10.1016/0301-4622(90)85020-7

Bonnans, J., Gilbert, J., Lemaréchal, C., Sagastizábal, C., 2006. Numerical optimization. Springer, Berlin.

Borgman, L., Taheri, M., Hagen, R., 1984. Three-dimensional, frequency domain simulations of geological variables, in Verly, G., David, M.,Journel, A., Marechal, A., Geostatistics for Natural Resources Characterization. Springer, Dordrecht, Netherlands, pp. 517-541.

Burton, P., 1977. Inversions of high frequency Q-1 (f). Geophysical Journal International 48, 29-51. doi:10.1111/j.1365-246x.1977.tb01283.x

Burton, P., Kennet, B., 1972. Upper Mantle Zone of Low Q. Nature Physical Science 238, 87-90. doi:10.1038/physci238087a0

Caers, J., Hoffman, T., 2006. The probability perturbation method: A new look at Bayesian inverse modeling. Mathematical Geology 38, 81-100. doi:10.1007/s11004-005-9005-9

Cardiff, M., Kitanidis, P., 2009. Bayesian inversion for facies detection: An extensible level set framework. Water Resources Research 45. doi:10.1029/2008wr007675

Carrera, J., Alcolea, A., Medina, A., Hidalgo, J., Slooten, L., 2005. Inverse problem in hydrogeology. Hydrogeology Journal 13, 206-222. doi:10.1007/s10040-004-0404-7

Carrera, J., Medina, A., 1994. An improved form of adjoint-state equations for transient problems, in Peters, A., Wittum, G., Herrling, B., Meissner, U., Brebbia, C.A., Gray, W.G., Pinder, G.F., Computational Methods in Water Resources X. Springer, Dordrecht, Netherlands, pp.199-206.

Carrera, J., Navarrina, F., Vives, L., Heredia, J., Medina, A., 1990. Computational aspects of the inverse problem, in Computational Methods in Subsurface Hydrology: Proceedings of the Eighth International Conference on Computational Methods in Water Resources, Venice, Italy, pp. 513-522.

Carrera, J., Neuman, S., 1986. Estimation of Aquifer Parameters Under Transient and Steady State Conditions: 2. Uniqueness, Stability, and Solution Algorithms. Water Resources Research 22, 211-227. doi:10.1029/wr022i002p00211

Chambers, R.L., Zinger, M.A., Kelly, M.C., 1994.Constraining geostatistical reservoir descriptions with 3-D seismic data to reduce uncertainty, in Yarus, J. M., Chambers, R. L., Stochastic modeling and geostatistics. American Association of Petroleum Geologists, Tulsa, OK, pp. 143-157.

Chang, H.C, Chen, H. C., Fang, J. H., 1997. Lithology determination from well logs with fuzzy associative memory neural network. IEEE Transactions on Geoscience and Remote Sensing 35, 773-780. doi:10.1109/36.582000

Chautru, J.M., 1989. The Use of Boolean Random Functions in Geostatistics,in Armstrong, M., Geostatistics. Quantitative Geology and Geostatistics, vol 4. Springer, Dordrecht, Netherlands, pp.201-212.

Chen, J., Rubin, Y., 2003. An effective Bayesian model for lithofacies estimation using geophysical data. Water Resources Research 39. doi:10.1029/2002wr001666

Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Advances in Water Resources 29, 1107-1122. doi:10.1016/j.advwatres.2005.09.007

Chilès, J., Delfiner, P., 1999. Geostatistics: Modelling spatial uncertainty. Wiley-Interscience, New York.

Chugunova, T., Hu, L., 2008. Multiple point simulations constrained by continuous auxiliary data. Mathematical Geosciences 40, 133-146. doi:10.1007/s11004-007-9142-4

COMSOL Multiphysics User's Guide, 2005. COMSOLAB, Stockholm (Sweden).

Christakos, G., 1990. A Bayesian/maximum-entropy view to the spatial estimation problem. Mathematical Geology 22, 763-777. doi:10.1007/bf00890661

Davis, M.W., 1987. Generating large stochastic simulations via the LU triangular decomposition of the covariance matrix. Mathematical Geology 19, 91-98. doi:10.1007/BF00898190.

Deutsch, V., Joumel, A., 1992. Annealing techniques applied to the integration of geological and engineering data, in Report 5, Stanford Center for Reservoir Forecasting, Stanford, CA.

Doveton, J. H., 1986. Log Analysis of Subsurface Geology Concepts and Computer Methods. John Wiley, New York.

Dowd, P.A., 1994. Geological controls in the geostatistical simulation of hydrocarbon reservoirs. Arabian Journal for Science and Engineering 19 (2B), 237-247.

Dowd, P.A., 1997. Structural controls in the geostatistical simulation of mineral deposits, in Baafi, E.Y., Schofield, N.A., Geostatistics Wollongong'96. Kluwer Academic, Dordrecht, pp. 647–657.

Emery, X., 2010. Iterative algorithms for fitting a linear model of coregionalization. Computers & Geosciences 36, 1150-1160. doi:10.1016/j.cageo.2009.10.007

Emery, X., 2007. Simulation of geological domains using the plurigaussian model: New developments and computer programs. Computers & Geosciences 33, 1189-1201. doi:10.1016/j.cageo.2007.01.006

Emery, X., Silva, D., 2009. Conditional co-simulation of continuous and categorical variables for geostatistical applications. Computers & Geosciences 35, 1234-1246. doi:10.1016/j.cageo.2008.07.005

Emsellem, Y., De Marsily, G., 1971. An Automatic Solution for the Inverse Problem. Water Resources Research 7, 1264-1283. doi:10.1029/wr007i005p01264

Feyen, L., Caers, J., 2006. Quantifying geological uncertainty for flow and transport modelling in multimodal heterogeneous formations. Advances in Water Resources 29, 912-929. doi: 10.1016/j.advwatres.2005.08.002

Fienen, M., Clemo, T., Kitanidis, P., 2008. An interactive Bayesian geostatistical inverse protocol for hydraulic tomography. Water Resources Research 44. doi:10.1029/2007wr006730

Fletcher, R., 1987. Practical Methods of Optimization, 2nd edn. John Wiley & Son, New York.

Frind, E., Pinder, G., 1973. Galerkin solution of the inverse problem for aquifer transmissivity. Water Resources Research 9, 1397-1410. doi:10.1029/wr009i005p01397

Fu, J., Gómez-Hernández, J., 2008. Preserving spatial structure for inverse stochastic simulation using blocking Markov chain Monte Carlo method. Inverse Problems in Science and Engineering 16, 865-884. doi:10.1080/17415970802015781

Galli, A., Beucher, H., Le Loc'h, G., Doligez, B., the Heresim Group, 1994. The pros and cons of the truncated Gaussian method, in Armstrong, M., Dowd, P., Geostatistical Simulations: Proceedings of the Geostatistical Workshop, Fontainebleau, France. Springer, Dordrecht, Netherlands, pp. 217-233.

Gavalas, G., Shah, P., Seinfeld, J., 1976. Reservoir History Matching by Bayesian Estimation. Society of Petroleum Engineers Journal 16, 337-350. doi:10.2118/5740-pa

Geman, S., Geman, D., 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6, 721-741. doi:10.1109/tpami.1984.4767596

Giudici, M., Morossi, G., Parravicini, G., Ponzini, G., 1995. A New Method for the Identification of Distributed Transmissivities. Water Resources Research 31, 1969-1988. doi:10.1029/95wr01205

Gómez-Hernánez, J., Sahuquillo, A., Capilla, J., 1997. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data—I. Theory. Journal of Hydrology 203, 162-174. doi:10.1016/s0022-1694(97)00098-x

Goldberg, D., 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Boston.

Goncz, J., Cleary, J., 1976. Variations in the Structure of the Upper Mantle Beneath Australia, from Rayleigh Wave Observations. Geophysical Journal International 44, 507-516. doi:10.1111/j.1365-246x.1976.tb03670.x

Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press, Paris, France.

Guadagnini, L., Guadagnini, A., Tartakovsky, D., 2004. A geostatistical model for distribution of facies in highly heterogeneous aquifers, in Sanchez-Vila, X., Carrera, J., Gomez-Hernandez, J., GeoENV IV – Geostatistics for environmental applications. Springer, Netherlands, pp. 211-222.

Gutjahr, A., 1989. Fast fourier transforms for random field generation: project report for Los Alamos Grant to New Mexico Tech. New Mexico Institute of Mining and Technology, Socorro, NM.

Hadamard, J., 1923. Lectures on Cauchy's problem in linear partial differential equations. Yale University Press, New Haven.

Hansen, T., Cordua, K., Mosegaard, K., 2012. Inverse problems with non-trivial priors: efficient solution through sequential Gibbs sampling. Computational Geosciences 16, 593-611. doi:10.1007/s10596-011-9271-1

Hansen, T., Mosegaard, K., Cordua, K.S., 2008. Using geostatistics to describe complex a priori information for inverse problems, in Ortiz, J.M., Emery, X., GEOSTATS 2008: Proceedings of the Eighth International Geostatistics Congress. Santiago, pp. 329-338.

Hantush, M., Mariño, M., 1997. Estimation of Spatially Variable Aquifer Hydraulic Properties Using Kalman Filtering. Journal of Hydraulic Engineering 123, 1027-1035. doi:10.1061/(asce)0733-9429(1997)123:11(1027)

Harp, D., Dai, Z., Wolfsberg, A., Vrugt, J., Robinson, B., Vesselinov, V., 2008. Aquifer structure identification using stochastic inversion. Geophysical Research Letters 35. doi:10.1029/2008gl033585

Hastings, W., 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika 57, 97. doi:10.2307/2334940

Hendricks Franssen, H., Kinzelbach, W., 2008. Real-time groundwater flow modeling with the Ensemble Kalman Filter: Joint estimation of states and parameters and the filter inbreeding problem. Water Resources Research 44. doi:10.1029/2007wr006505

Hewett, T., Behrens, R., 1990. Conditional Simulation of Reservoir Heterogeneity With Fractals. SPE Formation Evaluation 5, 217-225. doi:10.2118/18326-pa

Hoeksema, R., Kitanidis, P., 1985. Analysis of the Spatial Structure of Properties of Selected Aquifers. Water Resources Research 21, 563-572. doi:10.1029/wr021i004p00563

Hoerl, A. E., Kennard, R. W. 1968. On Regression Analysis and Biased Estimation. Technometrics 10, 422–423.

Hoerl, A., Kennard, R., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 55-67. doi:10.1080/00401706.1970.10488634

Hoerl, A., Kennard, R., 1970. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics 12, 69-82. doi:10.1080/00401706.1970.10488635

Holland, J., 1975. Adaptation in natural and artificial systems. University of Michigan Press.

Huysmans, M., Dassargues, A., 2009. Application of multiple point geostatistics on modelling groundwater flow and transport in a crossbedded aquifer (Belgium). Hydrogeology Journal 17, 1901-1911. doi: 10.1007/s10040-009-0495-2

Ibáñez, J. J., Saldaña, A., 2008. The continuum dilemma in pedometrics and pedology, in Krasilnikov, P., Carré, F., Montanarella, L., Geostatistics and soil geography. European Communities, Luxembourg, pp. 129-147.

Irving, J., Singha, K., 2010. Stochastic inversion of tracer test and electrical geophysical data to estimate hydraulic conductivities. Water Resources Research 46. doi:10.1029/2009wr008340

Jaynes, E., 1985. Highly informative priors, in Bernardo, J., DeGroot, M., Lindley, D., Smith, A., Bayesian Statistics, vol. 2., Elsevier, Amsterdam, pp. 329-360.

Jaynes, E., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge.

Jestin, F., Huchon, P., Gaulier, J., 1994. The Somalia plate and the East African Rift System: present-day kinematics. Geophysical Journal International 116, 637-654. doi:10.1111/j.1365-246x.1994.tb03286.x

Jeulin, D., 1987. Anisotropic rough surface modeling by random morphological function. Acta Stereologica 6, 183–189.

Jones, A., Hutton, R., 1979. A multi-station magnetotelluric study in southern Scotland -- II. Monte-Carlo inversion of the data and its geophysical and tectonic implications. Geophysical Journal International 56, 351-368. doi:10.1111/j.1365-246x.1979.tb00169.x

Journel, A., 1974. Geostatistics for Conditional Simulation of Ore Bodies. Economic Geology 69, 673-687. doi:10.2113/gsecongeo.69.5.673

Journel, A., 1983. Nonparametric estimation of spatial distributions. Journal of the International Association for Mathematical Geology 15, 445-468. doi:10.1007/bf01031292

Journel, A., 1989. Fundamentals of Geostatistics in Five Lessons. American Geophysical Union, Washington.

Journal, A., Alabert, F., 1989. Non-Gaussian data expansion in the Earth Sciences. Terra Nova 1, 123-134. doi:10.1111/j.1365-3121.1989.tb00344.x

Journel, A., Alabert, F., 1990. New Method for Reservoir Mapping. Journal of Petroleum Technology 42, 212-218. doi:10.2118/18324-pa

Keilis-Borok, V., Yanovskaja, T., 1967. Inverse Problems of Seismology (Structural Review). Geophysical Journal International 13, 223-234. doi:10.1111/j.1365-246x.1967.tb02156.x

Kennett, B., 1998. On the density distribution within the Earth. Geophysical Journal International 132, 374-382. doi:10.1046/j.1365-246x.1998.00451.x

Kerrou, J., Renard, P., Hendricks Franssen, H., Lunati, I., 2008. Issues in characterizing heterogeneity and connectivity in non-multiGaussian media. Advances in Water Resources 31, 147-159. doi:10.1016/j.advwatres.2007.07.002

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P. , 1983. Optimization by simulated annealing. Science 22O, 671–680. doi:10.1126/science.220.4598.671

Kitanidis, P., 1995. Quasi-Linear Geostatistical Theory for Inversing. Water Resources Research 31, 2411-2419. doi:10.1029/95wr01945

Kitanidis, P. K., 2007. On stochastic inverse modeling, in Hyndman, D. W., Day-Lewis, F. D., Singha, K., Subsurface Hydrology Data Integration for Properties and Processes. AGU, Washington, pp. 19-30.

Kitanidis, P., Vomvoris, E., 1983. A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. Water Resources Research 19, 677-690. doi:10.1029/wr019i003p00677

Koren, Z., Mosegaard, K., Landa, E., Thore, P., Tarantola, A., 1991. Monte Carlo estimation and resolution analysis of seismic background velocities. Journal of Geophysical Research: Solid Earth 96, 20289-20299. doi:10.1029/91jb02278

Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics 2, 164-168. doi:10.1090/qam/10666

Liu, N., 2005. Automatic History Matching of Geologic Facies (PhD). University of Oklahoma.

Lu, Z., Robinson, B., 2006. Parameter identification using the level set method. Geophysical Research Letters 33. doi:10.1029/2005gl025541

Mariethoz, G., Renard, P., Caers, J., 2010a. Bayesian inverse problem and optimization with iterative spatial resampling. Water Resources Research 46. doi:10.1029/2010wr009274

Mariethoz, G., Renard, P., Straubhaar, J., 2010b. The Direct Sampling method to perform multiple-point geostatistical simulations. Water Resources Research 46, W11536. doi:10.1029/2008WR007621

Mariethoz, G., Renard, P., Cornaton, F., Jaquet, O., 2009. Truncated Plurigaussian Simulations to Characterize Aquifer Heterogeneity. Ground Water 47, 13-24. doi:10.1111/j.1745-6584.2008.00489.x

Marquardt, D.W., 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. SIAM Journal on Applied Mathematics 11 (2), 431-441. doi:10.1137/0111030

Matheron, G. 1967. Elements pour une Theorie des Milieux Poreux. Masson & Cie, Paris.

Matheron, G., Beucher, H., Fouquet, C, Galli, A., Guerrillot, D., Ravenne, C., 1987. Conditional simulation of the geometry of fluvio-deltaic reservoirs, SPE 62nd Annual Conference, Dallas, Texas, September 27–30, SPE 16753, 571–599.

Menke, W., 1984. Geophysical data analysis: Discrete inverse theory. Academic Press, San Diego.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E., 1953. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics 21, 1087-1092. doi:10.1063/1.1699114

Michael, H., Boucher, A., Sun, T., Caers, J., Gorelick, S., 2010. Combining geologic process models and geostatistics for conditional simulation of 3D subsurface heterogeneity. Water Resources Research 46, W05527. doi:10.1029/2009WR008414

Mills, J., Fitch, T., 1977. Thrust faulting and crust--upper mantle structure in East Australia. Geophysical Journal International 48, 351-384. doi:10.1111/j.1365-246x.1977.tb03677.x

Modis, K., Sideri, D., 2013. Geostatistical Simulation of Hydrofacies Heterogeneity of the West Thessaly Aquifer Systems in Greece. Natural Resources Research 22, 123-138. doi:10.1007/s11053-013-9200-1

Mosegaard, K., 2011. Quest for consistency, symmetry, and simplicity — The legacy of Albert Tarantola. Geophysics 76, W51-W61. doi:10.1190/geo2010-0328.1

Mosegaard, K., Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. Journal of Geophysical Research: Solid Earth 100, 12431-12447. doi:10.1029/94jb03097

Neuman, S., 1973. Calibration of distributed parameter groundwater flow models viewed as a multiple-objective decision process under uncertainty. Water Resources Research 9, 1006-1021. doi:10.1029/wr009i004p01006

Oliver, D., Cunha, L., Reynolds, A., 1997. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. Mathematical Geology 29, 61-91. doi:10.1007/bf02769620

Press, F., 1968. Earth models obtained by Monte Carlo Inversion. Journal of Geophysical Research 73, 5223-5234. doi:10.1029/jb073i016p05223

Press, F., 1971. An introduction to Earth structure and seismotectonics, in Coulomb, J., Caputo, M., Mantle and Core in Planetary Physics (Proceedings of the International School of Physics “Enrico Fermi”). Academic Press, New York, pp. 209-241.

Renard, P., 2007. Stochastic Hydrogeology: What Professionals Really Need?. Ground Water 45, 531-541. doi:10.1111/j.1745-6584.2007.00340.x

Renard, P., Gomez-Hernandez, J., Ezzedine, S., 2005. Characterization of porous and fractured media, Encyclopedia of Hydrological Sciences 13, 147. doi: 10.1002/0470848944.hsa154

Ricard, Y., Vigny, C., Froidevaux, C., 1989. Mantle heterogeneities, geoid, and plate motion: A Monte Carlo inversion. Journal of Geophysical Research: Solid Earth 94, 13739-13754. doi:10.1029/jb094ib10p13739

Rogers, S. J., Fang, J., Karr, C., Stanley, D., 1992. Determination of lithology from well logs using a neural network. AAPG Bulletin 76, 731–739.

Ronayne, M., Gorelick, S., Caers, J., 2008. Identifying discrete geologic structures that produce anomalous hydraulic response: An inverse modeling approach. Water Resources Research 44, W08426. doi:10.1029/2007WR006635

Scales, J., Sneider, R., 1997. To Bayes or not to Bayes? Geophysics 62, 1045-1046. doi: 10.1190/1.6241045.1

Serra, J., 1982. Image Analysis and Mathematical Morphology. Academic Press, London.

Soares, A., 2001. Direct Sequential Co-Simulation: A New Stochastic Modelling for Environmental Applications, in Monestiez, P., Allard, D., Froidevaux, R., geoENV III — Geostatistics for Environmental Applications. Quantitative Geology and Geostatistics, vol 11. Springer, Dordrecht, Netherlands, pp. 381-391.

Stordal, A., Elsheikh, A., 2015. Iterative ensemble smoothers in the annealed importance sampling framework. Advances in Water Resources 86, 231-239. doi:10.1016/j.advwatres.2015.09.030

Tarantola, A., 1994. Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation, Elsevier, Amsterdam.

Tarantola, A., 2005. Inverse Problem Theory and Methods for Model Parameter Estimation, Siam, Philadelphia.

Tikhonov, A. N., Arsenin, V. Y., 1977. Solutions of Ill Posed Problems. Winston and Sons, Washington DC.

Tikhonov, A., I., Leonov, A. N., Tikhonov, A. G., Yagola, A. S., Leonov, 1997. Non-linear ill-posed problems, CRC Press, Boca Raton, FL.

Valakas, G., Modis, K., 2015. Bayesian facies inversion, using spatial resampling and cosimulation with model response data, in Helmut Schaeben, Raimon Tolosana Delgado, K. Gerald van den Boogaart, Regina van den Boogaart, Proceedings of IAMG 2015 Freiberg, September 5-13, 2015, the 17th Annual Conference of the International Association for Mathematical Geosciences, 1372 pages, ISBN 978-3-00-050337-5 (DVD).

Valakas, G., Modis, K., 2016. Using informative priors in facies inversion: The case of C-ISR method. Advances in Water Resources 94, 23-30. doi:10.1016/j.advwatres.2016.04.019

Verly, G, 1993. Sequential Gaussian Cosimulation: A Simulation Method Integrating Several Types of Information, in Soares, A., Geostatistics Tróia '92, vol 1, Kluwer, Dordrecht, Netherlands, pp. 543-554.

Von Neumann, J., 1951. Various techniques used in connection with random digits. Applied Math Series, Notes by G. E. Forsythe, in National Bureau of Standards, 12, 36-38.

Wackernagel, H., 2003. Multivariate Geostatistics - An Introduction with Applications. Springer,Berlin.

Whitley, D., 1994. A genetic algorithm tutorial. Statistics and Computing 4. doi:10.1007/bf00175354

Williamson, J., 2010. Review of Bruno di Finetti's "Philosophical Lectures on Probability". Philosophia Mathematica 18, pp. 130-135. doi:10.1093/philmat/nkp019

Winter, C., Tartakovsky, D., 2000. Mean Flow in composite porous media. Geophysical Research Letters 27, 1759-1762. doi:10.1029/1999gl011030

Winter, C., Tartakovsky, D., 2002. Groundwater flow in heterogeneous composite aquifers. Water Resources Research 38, 23-1-23-11. doi:10.1029/2001wr000450

Woodbury, A., Smith, L., Dunbar, W., 1987. Simultaneous inversion of hydrogeologic and thermal data: 1. Theory and application using hydraulic head data. Water Resources Research 23, 1586-1606. doi:10.1029/wr023i008p01586

Worthington, M., Cleary, J., Anderssen, R., 1972. Density Modelling by Monte Carlo Inversion--II Comparison of Recent Earth Models. Geophysical Journal International 29, 445-457. doi:10.1111/j.1365-246x.1972.tb06170.x

Worthington, M., Cleary, J., Anderssen, R., 1974. Upper and Lower Mantle Shear Velocity Modelling by Monte Carlo Inversion. Geophysical Journal International 36, 91-103. doi:10.1111/j.1365-246x.1974.tb03627.x

Ye, M., Neuman, S., Meyer, P., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resources Research 40. doi:10.1029/2003wr002557

Yeh, J., Gutjahr, A., Jin, M., 1995. An iterative cokriging-like technique for groundwater flow modeling. Groundwater 331, 33–41. doi10.1111/ j.1745-6584.1995.tb00260.x

Zhou, H., Gómez-Hernández, J., Li, L., 2014. Inverse methods in hydrogeology: Evolution and recent trends. Advances in Water Resources 63, 22-37. doi:10.1016/j.advwatres.2013.10.014

Zinn, B., Harvey, C., 2003. When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields. Water Resources Research 39. doi:10.1029/2001wr001146

Appendix A: Mathematical Notes

A.1 General Rules of Matrices

A.1.1: Let **A** is p x n, **C** n x n and **B** n x q matrices, then

$$(\mathbf{ABC})^{\mathrm{T}} = \mathbf{C}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}$$

A.1.2: Let **C** is a symmetric matrix then

$$\mathbf{C} = \mathbf{C}^{\mathrm{T}}$$

A.1.3: Let **y** a *p x 1* vector, **b** a vector *n x 1* and **A** a *p x n* matrix then

$$\mathbf{y}^{\mathrm{T}}\mathbf{Ab} = (\mathbf{y}^{\mathrm{T}}\mathbf{Ab})^{\mathrm{T}} = \mathbf{b}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{y}$$

A.2 Derivatives of Matrices and Vectors

A.2.1: $$\frac{\partial}{\partial \mathbf{b}}\mathbf{y}^{\mathrm{T}}\mathbf{Ab} = \mathbf{A}^{\mathrm{T}}\mathbf{y}$$

A.2.2: $$\frac{\partial}{\partial \mathbf{b}}\mathbf{b}^{\mathrm{T}}\mathbf{Ab} = (\mathbf{A} + \mathbf{A}^{\mathrm{T}})\mathbf{b}$$

A.2.3: $$\frac{\partial}{\partial \mathbf{b}}(\mathbf{y} - \mathbf{Ab})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{Ab}) = -2\mathbf{A}^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{Ab})$$

A.3 Matrix Inversion Identity

Let $\mathbf{C}_\mathrm{M}$ and $\mathbf{C}_\mathrm{D}$, *n x n* and *p x p* positive-definite matrices. Also, let **G** is *p x n* matrix. Then

$$\mathbf{C}_\mathrm{M}\mathbf{G}^\mathbf{T}(\mathbf{G}\mathbf{C}_\mathbf{M}\mathbf{G}^\mathbf{T}+\mathbf{C}_\mathbf{D})^{-1}=\left(\mathbf{G}^\mathbf{T}\mathbf{C}_\mathrm{D}^{-\mathbf{1}}\mathbf{G}+\mathbf{C}_\mathrm{M}^{-\mathbf{1}}\right)^{-\mathbf{1}}\mathbf{G}^\mathbf{T}\mathbf{C}_\mathrm{D}^{-\mathbf{1}}$$

Proof

Beginning from the expression $(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1})\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{G}^\mathrm{T}$ and multiplying the second term with $\mathbf{C}_\mathrm{D}^{-1}\mathbf{C}_\mathrm{D}$ we have:

$$\begin{aligned}(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1})\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{G}^\mathrm{T}&=(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1})\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{C}_\mathrm{D}\\&=\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}(\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{C}_\mathrm{D})\end{aligned}$$

Similarly, multiplying the second term with $\mathbf{C}_\mathrm{M}^{-1}\mathbf{C}_\mathrm{M}$ we have:

$$\begin{aligned}(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1})\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{G}^\mathrm{T}&=(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1})\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{C}_\mathrm{M}^{-1}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}\\&=(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{G}+\mathbf{C}_\mathrm{M}^{-1})\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}\end{aligned}$$

Since the left-hand sides of the above equations are identical, then:

$$\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}(\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{C}_\mathrm{D})=(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{G}+\mathbf{C}_\mathrm{M}^{-1})\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}$$

Since $\mathbf{C}_\mathrm{D}$ and $\mathbf{C}_\mathrm{M}$ are positive-definite matrices, it follows that $(\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{C}_\mathrm{D})$ and $(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{G}+\mathbf{C}_\mathrm{M}^{-1})$ are both nonsingular positive-definite matrices. Multiplying the two hands of above equation by $(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{G}+\mathbf{C}_\mathrm{M}^{-1})^{-1}$, we have:

$$(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{G}+\mathbf{C}_\mathrm{M}^{-1})^{-1}\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}(\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{C}_\mathrm{D})=\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}$$

Multiplying the two hands of above equation by $(\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{C}_\mathrm{D})^{-1}$, we have:

$$(\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}\mathbf{G}+\mathbf{C}_\mathrm{M}^{-1})^{-1}\mathbf{G}^\mathrm{T}\mathbf{C}_\mathrm{D}^{-1}=\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}(\mathbf{G}\mathbf{C}_\mathrm{M}\mathbf{G}^\mathrm{T}+\mathbf{C}_\mathrm{D})^{-1}$$

A.4 Least Squared Estimation in Overdetermined Problems:

The minimization of objective function

$$\mathbf{S}(\hat{\mathbf{m}}) = (\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\text{T}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})$$

is given by

$$\hat{\mathbf{m}} = (\mathbf{G}^{\text{T}}\mathbf{G})^{-1}\mathbf{G}^{\text{T}}\mathbf{d}_{\text{obs}}$$

Proof

Setting the expanded derivative of the objective function equal to zero and using the A.2.3 property, we have:

$$\frac{\partial}{\partial \hat{\mathbf{m}}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\text{T}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) = 0$$

$$-2\mathbf{G}^{\text{T}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) = 0 \Rightarrow$$

$$\hat{\mathbf{m}} = (\mathbf{G}^{\text{T}}\mathbf{G})^{-1}\mathbf{G}^{\text{T}}\mathbf{d}_{\text{obs}}$$

In order to prove the estimated $\hat{\mathbf{m}}$ is the minimum, it should be shown that the second order derivative of objective function (Hessian) is positive definite:

$$\frac{\partial^2}{\partial \hat{\mathbf{m}}^2}\mathbf{S}(\hat{\mathbf{m}}) = 2\mathbf{G}^{\mathbf{T}}\mathbf{G}$$

This is a positive definite matrix, since **G** has full rank. It is noted that the above solution use the fact that the inverse of $\mathbf{G}^{\mathbf{T}}\mathbf{G}$ exists.

A.5 Least Squared Estimation in Underdetermined Problems

The solution of constrained optimization problem

$$\min \mathbf{m}^{\mathrm{T}}\mathbf{m}$$

$$\text{subject to } \mathbf{d}_{\text{obs}} - \mathbf{G}\mathbf{m} = \mathbf{0}$$

is

$$\mathbf{m} = \mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{d}_{\text{obs}}$$

Proof

Minimization with constraints can be done with Lagrange multipliers $\boldsymbol{\lambda}$.

$$\boldsymbol{S}(\mathbf{m};\boldsymbol{\lambda}) = \mathbf{m}^{\mathrm{T}}\mathbf{m} - \boldsymbol{\lambda}^{\mathrm{T}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\mathbf{m})$$

Taking the derivative of the objective function:

$$\frac{\partial}{\partial \mathbf{m}}\boldsymbol{S}(\mathbf{m};\boldsymbol{\lambda}) = 2\mathbf{m} - \mathbf{G}^{\mathrm{T}}\boldsymbol{\lambda}$$

The first term results from A.2.2 property and the second term from A.2.1 property.

Also,

$$\frac{\partial}{\partial \boldsymbol{\lambda}}\boldsymbol{S}(\mathbf{m};\boldsymbol{\lambda}) = \mathbf{d}_{\text{obs}} - \mathbf{G}\mathbf{m}$$

Setting the derivatives to zero we get:

$$\mathbf{m} = \frac{1}{2}\mathbf{G}^{\mathrm{T}}\boldsymbol{\lambda} \quad \text{and} \quad \mathbf{d}_{\text{obs}} = \mathbf{G}\mathbf{m}$$

Then

$$\mathbf{d}_{\text{obs}} = \frac{1}{2}\mathbf{G}\mathbf{G}^{\mathrm{T}}\boldsymbol{\lambda} \Rightarrow \quad \boldsymbol{\lambda} = \mathbf{2}(\mathbf{G}\mathbf{G}^{\mathrm{T}})^{-\mathbf{1}}\mathbf{d}_{\text{obs}}$$

Therefore $\mathbf{m} = \mathbf{G}^{\mathrm{T}}(\mathbf{G}\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{d}_{\text{obs}}$

We note that the above solution use the fact that the inverse of $\mathbf{G}^{\mathbf{T}}\mathbf{G}$ exists.

A.6 Weighed Least Square Estimation in Overdetermined Problems:

The minimization of objective function

$$\mathbf{S}(\hat{\mathbf{m}}) = (\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})$$

is given by

$$\hat{\mathbf{m}} = (\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\text{T}}\mathbf{d}_{\text{obs}}$$

Proof

Setting the expanded derivative of the objective function equal to zero and using the A.2.3 property, we have:

$$\frac{\partial}{\partial \hat{\mathbf{m}}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) = 0$$

$$-2\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) = 0 \Rightarrow$$

$$\hat{\mathbf{m}} = (\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G})^{-1}\mathbf{G}^{\text{T}}\mathbf{d}_{\text{obs}}$$

In order to prove the estimated $\hat{\mathbf{m}}$ is the minimum, it should be shown that the second order derivative of objective function (Hessian) is positive definite:

$$\frac{\partial^2}{\partial \hat{\mathbf{m}}^2}\mathbf{S}(\hat{\mathbf{m}}) = 2\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G}$$

This is a positive definite matrix, since $\mathbf{G}$ has full rank and $\mathbf{C}_{\text{D}}$ is positive-definite matrix. It is noted that the above solution use the fact that the inverse of $\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G}$ exists.

A.7 Regularized Weighted Least Square Estimation in Overdetermined Problems:

The minimization of objective function:

$$\mathbf{S}(\hat{\mathbf{m}}) = (\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) + \lambda(\mathbf{m}^{\text{pr}} - \hat{\mathbf{m}})^{\text{T}}\mathbf{C}_{\text{M}}^{-1}(\mathbf{m}^{\text{pr}} - \hat{\mathbf{m}})$$

is given by

$$\hat{\mathbf{m}} = \mathbf{m}^{\text{pr}} + (\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G} + \boldsymbol{\lambda}\mathbf{C}_{\text{M}}^{-1})^{-1}\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\mathbf{m}^{\text{pr}})$$

Proof

Setting the expanded derivative of the objective function equal to zero and using the A.2.3 property, we have:

$$\frac{\partial}{\partial\hat{\mathbf{m}}}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}})^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) + \lambda(\mathbf{m}^{\text{pr}} - \hat{\mathbf{m}})^{\text{T}}\mathbf{C}_{\text{M}}^{-1}(\mathbf{m}^{\text{pr}} - \hat{\mathbf{m}}) = 0 \Rightarrow$$

$$-2\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\hat{\mathbf{m}}) - 2\lambda\mathbf{C}_{\text{M}}^{-1}(\mathbf{m}^{\text{pr}} - \hat{\mathbf{m}}) = 0 \Rightarrow$$

$$\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G}\hat{\mathbf{m}} + \lambda\mathbf{C}_{\text{M}}^{-1}(\hat{\mathbf{m}} - \mathbf{m}^{\text{pr}}) = \mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{d}_{\text{obs}}$$

Subtracting the term $\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G}\mathbf{m}^{\text{pr}}$ to the left and right hand of the above equation we have:

$$\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G}(\hat{\mathbf{m}} - \mathbf{m}^{\text{pr}}) + \lambda\mathbf{C}_{\text{M}}^{-1}(\hat{\mathbf{m}} - \mathbf{m}^{\text{pr}}) = \mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{d}_{\text{obs}} - \mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G}\mathbf{m}^{\text{pr}} \Rightarrow$$

$$(\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G} + \lambda\mathbf{C}_{\text{M}}^{-1})(\hat{\mathbf{m}} - \mathbf{m}^{\text{pr}}) = \mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\mathbf{m}^{\text{pr}}) \Rightarrow$$

$$\hat{\mathbf{m}} = \mathbf{m}^{\text{pr}} + (\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G} + \lambda\mathbf{C}_{\text{M}}^{-1})^{-1}\mathbf{G}^{\text{T}}\mathbf{C}_{\text{D}}^{-1}(\mathbf{d}_{\text{obs}} - \mathbf{G}\mathbf{m}^{\text{pr}})$$

In order to prove the estimated $\hat{\mathbf{m}}$ is the minimum, it should be shown that the second order derivative of objective function (Hessian) is positive definite:

$$\frac{\partial^2}{\partial\hat{\mathbf{m}}^2}\mathbf{S}(\hat{\mathbf{m}}) = 2\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G} + \lambda\mathbf{C}_{\text{M}}^{-1}$$

This is a positive definite matrix, since $\mathbf{G}$ has full rank and $\mathbf{C}_{\text{D}}$ and $\mathbf{C}_{\text{M}}$ are positive-definite matrices. It is noted that the above solution use the fact that the inverse of $\mathbf{G}^{\mathbf{T}}\mathbf{C}_{\text{D}}^{-1}\mathbf{G} + \lambda\mathbf{C}_{\text{M}}^{-1}$ exists.

A.8 Nonlinear Weighted Least Square Estimation:

The minimization of objective function:

$$\mathbf{S}(\mathbf{m}^{i+1}) = \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right)^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right)$$

is given by

$$\mathbf{m}^{i+1} = \mathbf{m}^i + \left(\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i\right)^{-1} \mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i)\right)$$

Proof

Setting the expanded derivative of the objective function equal to zero and using the A.2.3 property, we have:

$$\frac{\partial}{\partial \mathbf{m}^{i+1}} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right)^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right)$$

$$= 0 \Rightarrow$$

$$-2\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right) = 0 \Rightarrow$$

$$\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i) = \mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i)\right) \Rightarrow$$

$$\mathbf{m}^{i+1} = \mathbf{m}^i + \left(\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i\right)^{-1} \mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i)\right)$$

In order to prove the estimated $\mathbf{m}^{i+1}$ is a minimum, it should be shown that the second order derivative of objective function is positive definite:

$$\frac{\partial}{\partial (\mathbf{m}^{i+1})^2} \mathbf{S}(\mathbf{m}^{i+1}) = 2\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i$$

This is a positive definite matrix, if $\mathbf{G}_i$ has full rank and $\mathbf{C}_{\mathrm{D}}$ and $\mathbf{C}_{\mathrm{M}}$ are positive-definite matrices. It is noted that the above solution use the fact that the inverse of $\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i$ exists. Finally, the $\mathbf{m}^{i+1}$ is a better solution than the previous if $\mathbf{S}(\mathbf{m}^{i+1}) < \mathbf{S}(\mathbf{m}^i)$

A.9 Regularized Weighted Nonlinear Least Square Estimation:

The minimization of objective function:

$$\mathbf{S}(\mathbf{m}^{i+1}) = \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right)^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right) + \lambda \mathbf{m}^{i+1^{\mathrm{T}}} \mathbf{C}_{\mathrm{M}}^{-1} \mathbf{m}^{i+1}$$

is given by

$$\mathbf{m}^{i+1} = \mathbf{m}^i + \left(\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i + \lambda \mathbf{C}_{\mathrm{M}}^{-1}\right)^{-1} \mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i \mathbf{m}^i\right)$$

Proof

Setting the expanded derivative of the objective function equal to zero and using the A.2.3 property, we have:

$$\frac{\partial}{\partial \mathbf{m}^{i+1}} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right)^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right) + \lambda \mathbf{m}^{i+1^{\mathrm{T}}} \mathbf{C}_{\mathrm{M}}^{-1} \mathbf{m}^{i+1} = 0 \Rightarrow$$

$$-2\mathbf{G}_{\mathrm{i}}^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i)\right) + 2\lambda \mathbf{C}_{\mathrm{M}}^{-1} \mathbf{m}^{i+1} = 0 \Rightarrow$$

$$\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i(\mathbf{m}^{i+1} - \mathbf{m}^i) + \lambda \mathbf{C}_{\mathrm{M}}^{-1} \mathbf{m}^{i+1} = \mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i)\right)$$

Subtracting the term $\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i \mathbf{m}^i$ to the left and right hand of the above equation we have:

$$\left(\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i + \lambda \mathbf{C}_{\mathrm{M}}^{-1}\right)(\mathbf{m}^{i+1} - \mathbf{m}^i) = \mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i \mathbf{m}^i\right) \Rightarrow$$

$$\mathbf{m}^{i+1} = \mathbf{m}^i + \left(\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i + \lambda \mathbf{C}_{\mathrm{M}}^{-1}\right)^{-1} \mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \left(\mathbf{d}_{\text{obs}} - \mathbf{g}(\mathbf{m}^i) - \mathbf{G}_i \mathbf{m}^i\right)$$

In order to prove the estimated $\mathbf{m}^{i+1}$ is a minimum, it should be shown that the second order derivative of objective function is positive definite:

$$\frac{\partial}{\partial (\mathbf{m}^{i+1})^2} \mathbf{S}(\mathbf{m}^{i+1}) = 2\mathbf{G}_i^{\mathrm{T}} \mathbf{C}_{\mathrm{D}}^{-1} \mathbf{G}_i + \lambda \mathbf{C}_{\mathrm{M}}^{-1}$$

This is a positive definite matrix, if $\mathbf{G}_i$ has full rank and $\mathbf{C}_{\mathrm{D}}$ and $\mathbf{C}_{\mathrm{M}}$ are positive-definite matrices. It is noted that the above solution use the fact that the inverse of

$\mathbf{G}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\mathbf{G}_i + \lambda\mathbf{C}_{\mathrm{M}}^{-1}$ exists. Finally, the $\mathbf{m}^{i+1}$ is a better solution than the previous if $\mathbf{S}(\mathbf{m}^{i+1}) < \mathbf{S}(\mathbf{m}^{i})$.

A.10 Quasi linear Bayesian Estimation:

The minimization of objective function:

$$S(\widehat{\mathbf{m}}^{i+1}) = \left(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}(\widehat{\mathbf{m}}^{i+1} - \widetilde{\mathbf{m}}^{i+1}) - \mathbf{d}_{\mathrm{obs}}\right)^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}(\widehat{\mathbf{m}}^{i+1} - \widetilde{\mathbf{m}}^{i+1}) - \mathbf{d}_{\mathrm{obs}}\right) + \left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right)^{\mathrm{T}}\mathbf{C}_{\mathrm{M}}^{-1}\left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right)$$

is given by

$$\widehat{\mathbf{m}}^{i+1} = \widehat{\mathbf{m}}^{i} + \left(\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_i + \mathbf{C}_{\mathrm{M}}^{-1}\right)^{-1}\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\widehat{\mathbf{m}}^{i}) + \widetilde{\mathbf{G}}_{\mathrm{i}}(\widehat{\mathbf{m}}^{i} - \widetilde{\mathbf{m}}^{i+1})\right)$$

Proof

Setting the expanded derivative of the objective function equal to zero and using the A.2.3 property, we have:

$$\frac{\partial}{\partial \mathbf{m}^{i+1}}\left(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}(\widehat{\mathbf{m}}^{i+1} - \widetilde{\mathbf{m}}^{i+1}) - \mathbf{d}_{\mathrm{obs}}\right)^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}(\widehat{\mathbf{m}}^{i+1} - \widetilde{\mathbf{m}}^{i+1}) - \mathbf{d}_{\mathrm{obs}}\right) + \left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right)^{\mathrm{T}}\mathbf{C}_{\mathrm{M}}^{-1}\left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right) = 0 \Rightarrow$$

$$2\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}\left(\widehat{\mathbf{m}}^{i+1} - \widetilde{\mathbf{m}}^{i+1}\right) - \mathbf{d}_{\mathrm{obs}}\right) + 2\mathbf{C}_{\mathrm{M}}^{-1}\left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right) = 0 \Rightarrow$$

$$\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) - \widetilde{\mathbf{G}}_{\mathrm{i}}\widetilde{\mathbf{m}}^{i+1} - \mathbf{d}_{\mathrm{obs}}\right) + \widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}\widehat{\mathbf{m}}^{i+1} + \mathbf{C}_{\mathrm{M}}^{-1}\left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right) = 0 \Rightarrow$$

$$\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}\widehat{\mathbf{m}}^{i+1} + \mathbf{C}_{\mathrm{M}}^{-1}\left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right) = \widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}\widetilde{\mathbf{m}}^{i+1}\right)$$

Subtracting the term $\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_i\widehat{\mathbf{m}}^{i}$ to the left and right hand of the above equation we have:

$$\left(\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}} + \mathbf{C}_{\mathrm{M}}^{-1}\right)\left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right) = \widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) + \widetilde{\mathbf{G}}_{\mathrm{i}}\widetilde{\mathbf{m}}^{i+1}\right) - \widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_i\widehat{\mathbf{m}}^{i} \Rightarrow$$

$$\left(\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}} + \mathbf{C}_{\mathrm{M}}^{-1}\right)\left(\widehat{\mathbf{m}}^{i+1} - \widehat{\mathbf{m}}^{i}\right) = \widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) - \widetilde{\mathbf{G}}_i\widehat{\mathbf{m}}^{i} + \widetilde{\mathbf{G}}_{\mathrm{i}}\widetilde{\mathbf{m}}^{i+1}\right) \Rightarrow$$

$$\widehat{\mathbf{m}}^{i+1} = \widehat{\mathbf{m}}^{i} + \left(\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}} + \mathbf{C}_{\mathrm{M}}^{-1}\right)\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}} - \mathbf{g}(\widetilde{\mathbf{m}}^{i+1}) - \widetilde{\mathbf{G}}_i\left(\widehat{\mathbf{m}}^{i} - \widetilde{\mathbf{m}}^{i+1}\right)\right)$$

Also, instead to subtract the term $\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_i\widehat{\mathbf{m}}^{i}$, but extending the above equation, we have:

$$\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}\widehat{\mathbf{m}}^{i+1}+\mathbf{C}_{\mathrm{M}}^{-1}\widehat{\mathbf{m}}^{i+1}-\mathbf{C}_{\mathrm{M}}^{-1}\widehat{\mathbf{m}}^{i}=\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}}-\mathbf{g}\left(\widetilde{\mathbf{m}}^{i+1}\right)\right)+\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}\widetilde{\mathbf{m}}^{i+1}\Rightarrow$$

$$\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}\left(\widehat{\mathbf{m}}^{i+1}-\widetilde{\mathbf{m}}^{i+1}\right)+\mathbf{C}_{\mathrm{M}}^{-1}\widehat{\mathbf{m}}^{i+1}-\mathbf{C}_{\mathrm{M}}^{-1}\widehat{\mathbf{m}}^{i}=\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}}-\mathbf{g}\left(\widetilde{\mathbf{m}}^{i+1}\right)\right)$$

And now, subtracting the term $\mathbf{C}_{\mathrm{M}}^{-1}\widetilde{\mathbf{m}}^{i+1}$ to the left and right hand of the above equation we have:

$$\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}\left(\widehat{\mathbf{m}}^{i+1}-\widetilde{\mathbf{m}}^{i+1}\right)+\mathbf{C}_{\mathrm{M}}^{-1}\widehat{\mathbf{m}}^{i+1}-\mathbf{C}_{\mathrm{M}}^{-1}\widehat{\mathbf{m}}^{i}-\mathbf{C}_{\mathrm{M}}^{-1}\widetilde{\mathbf{m}}^{i+1}$$
$$=\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}}-\mathbf{g}\left(\widetilde{\mathbf{m}}^{i+1}\right)\right)-\mathbf{C}_{\mathrm{M}}^{-1}\widetilde{\mathbf{m}}^{i+1}\Rightarrow$$

$$\left(\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}+\mathbf{C}_{\mathrm{M}}^{-1}\right)\left(\widehat{\mathbf{m}}^{i+1}-\widetilde{\mathbf{m}}^{i+1}\right)=\widetilde{\mathbf{G}}_{\mathrm{i}}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}}-\mathbf{g}\left(\widetilde{\mathbf{m}}^{i+1}\right)\right)+\mathbf{C}_{\mathrm{M}}^{-1}(\widetilde{\mathbf{m}}^{i+1}-\widehat{\mathbf{m}}^{i})\Rightarrow$$

$$\widehat{\mathbf{m}}^{i+1}=\widetilde{\mathbf{m}}^{i+1}+\left(\widetilde{\mathbf{G}}_{\mathrm{i}}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}+\mathbf{C}_{\mathrm{M}}^{-1}\right)^{-1}\left(\widetilde{\mathbf{G}}_{\mathrm{i}}^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\left(\mathbf{d}_{\mathrm{obs}}-\mathbf{g}\left(\widetilde{\mathbf{m}}^{i+1}\right)\right)+\mathbf{C}_{\mathrm{M}}^{-1}(\widetilde{\mathbf{m}}^{i+1}-\widehat{\mathbf{m}}^{i})\right)$$

In order to prove the estimated $\widehat{\mathbf{m}}^{i+1}$ is a minimum, it should be shown that the second order derivative of objective function is positive definite:

$$\frac{\partial}{\partial(\widehat{\mathbf{m}}^{i+1})^2}\mathbf{S}\left(\widehat{\mathbf{m}}^{i+1}\right)=\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}+\mathbf{C}_{\mathrm{M}}^{-1}$$

This is a positive definite matrix, if $\widetilde{\mathbf{G}}_{\mathrm{i}}$ has full rank and $\mathbf{C}_{\mathrm{D}}$ and $\mathbf{C}_{\mathrm{M}}$ are positive-definite matrices. It is noted that the above solution use the fact that the inverse of $\widetilde{\mathbf{G}}_i^{\mathrm{T}}\mathbf{C}_{\mathrm{D}}^{-1}\widetilde{\mathbf{G}}_{\mathrm{i}}+\mathbf{C}_{\mathrm{M}}^{-1}$ exists. Finally, the $\widehat{\mathbf{m}}^{i+1}$ is a better solution than the previous if $\mathbf{S}\left(\widehat{\mathbf{m}}^{i+1}\right)<\mathbf{S}\left(\widehat{\mathbf{m}}^{i}\right)$.

Appendix B: *English – Greek Dictionary of Terms*

A	adjoint method	συζυγής μέθοδος
	ambiguity	αμφισημία, ασάφεια
	a-posteriori distribution / information	μεταγενέστερη ή ύστερη κατανομή / πληροφορία
	a-priori distribution / information	εκ προοιμίου ή πρότερη κατανομή / πληροφορία
	auxiliary variable	βοηθητική μεταβλητή
B	bias	μεροληψία
	bias–variance tradeoff	συμβιβασμός μεταξύ μεροληψίας – διασποράς
C	coefficient	συντελεστής
	computational precision	υπολογιστική ακρίβεια
	conjugate prior distribution	συζυγής πρότερη κατανομή
	consistent	συνεπής
	convergence time	χρόνος σύγκλισης
	coregionalization model	μοντέλο συμμεταβλητότητας
	correlation	συσχέτιση
	cosimulation	από κοινού προσομοίωση
	covariance	συνδιακύμανση
	cross-variogram	συν-βαριογράμμα

	cumulative distribution function	αθροιστική συνάρτηση κατανομής
D	determinant	ορίζουσα
	deterministic solution	αιτιοκρατική επίλυση
	discrete problem	διακριτό πρόβλημα
E	error function	συνάρτηση σφάλματος
	Euclidean length	Ευκλείδειο μήκος
F	finite differences	πεπερασμένες διαφορές
	first order approximation	πρώτης τάξης προσέγγιση
	forward problem	ευθύ πρόβλημα
	full rank	μέγιστη τάξη
G	generalized inverse matrix	γενικευμένος αντίστροφος πίνακας
	genetic algorithm	γενετικοί αλγόριθμοι
	geoscientifical	γεωεπιστημικός
	global minimum	ολικό ελάχιστο
	gradient methods	μέθοδοι κλίσης
	groundwater flow system	σύστημα υπόγειας ροής
H	Hessian matrix	Εσσιανός πίνακας
	heuristic methods	ευρετικές μέθοδοι
	hydrofacies	υδρογεωλογικές φάσεις
I	ill-conditioned system	ασταθές σύστημα (εξισώσεων)
	ill-posed system	ασθενώς τοποθετημένο σύστημα

	importance sampling	δειγματισμός σπουδαιότητας
	inconsistent	ασυνεπής
	independent samples	ανεξάρτητα δείγματα
	infinite solutions	άπειρες λύσεις
	informative prior distribution	πληροφορημένη ή ενημερωμένη πρότερη κατανομή
	inner product	εσωτερικό γινόμενο
	interpolation method	μέθοδος παρεμβολής
	inverse problem	αντίστροφο πρόβλημα
J	Jacobian matrix	Ιακωβιανός πίνακας
	join distribution	από κοινού κατανομή
K	kernel transition	πυρήνας μετάβασης
L	least squares criterion	κριτήριο ελαχίστων τετραγώνων
	likelihood	πιθανοφάνεια
	local optima	τοπικά ακρότατα
M	maximum likelihood estimation	εκτίμηση μέγιστης πιθανοφάνειας
	monotonically decreasing	μονότονα φθίνουσα
	mutation	μετάλλαξη
N	nonlinear linear least squares	μη-γραμμικά ελάχιστα τετράγωνα
	nonsingular	μη ιδιόμορφος
	norm	μέτρο, νόρμα
	Normal distribution	Κανονική κατανομή

	normalization constant	σταθερά κανονικοποίησης
O	objective function	αντικειμενική συνάρτηση
	operator	τελεστής
	optimal solution	βέλτιστη λύση
	optimization	βελτιστοποίηση
	ordinary least squares	ελάχιστα τετράγωνα
	overdetermined problem	υπερ-καθορισμένο πρόβλημα
	overfitting	υπερπροσαρμογή
P	penalty term	παράγοντας ποινής
	perturbation	διατάραξη
	positive definite	θετικά ορισμένος
	probability density function	συνάρτηση πυκνότητας πιθανότητας
	projection	προβολή
Q	quadratic form	τετραγωνική μορφή
	quasi-linear	οιονεί-γραμμικός, σχεδόν γραμμικός
R	random noise	τυχαίος θόρυβος
	realization	υλοποίηση
	regularization	κανονικοποίηση
	resampling	επαναδειγματισμός
	residual	κατάλοιπο
	residual sum of squares	άθροισμα των τετραγώνων των σφαλμάτων

	response	απόκριση
	ridge estimator	αμφικλινής εκτιμητής
S	sensitivity matrix	πίνακας ευαισθησίας
	sequences of iterates	ακολουθίες επαναλήψεων
	shrinking	συρρίκνωση
	simulated annealing method	μέθοδος προσομοιωμένης ανόπτησης
	singular matrix	ιδιόμορφος πίνακας
	span (a space)	παράγω (έναν χώρο)
	state variable	μεταβλητή κατάστασης
	stochastic method /solution	στοχαστική μέθοδος / επίλυση
	stopping criterion	κριτήριο τερματισμού ή διακοπής
	system parameterized	παραμετροποιημένο σύστημα
T	tangent plane	εφαπτόμενο επίπεδο
	threshold	κατώφλι
	tolerance	ανοχή
	truncated Gaussian simulation	αποκομμένη Γκαουσιανή προσομοίωση
U	unbiased estimator	αμερόληπτος εκτιμητής
	underdetermined problem	υπο-καθορισμένο πρόβλημα
V	variance	διακύμανση ή διασπορά
	weighted least squares	ζυγισμένα ελάχιστα τετράγωνα
W	white noise	λευκός θόρυβος