View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Repositório Aberto da Universidade do Porto

Faculdade de Engenharia da Universidade do Porto

Impacts of conforming gross errors when solving State Estimation problems in Power Systems

João Luís Ornelas da Silva

Dissertação realizada no âmbito do
Mestrado Integrado em Engenharia Eletrotécnica e de Computadores
Major Energia

Orientador: Prof. Doutor Vladimiro Henrique Barrosa Pinto de Miranda
Co-orientador: Doutor Leonel de Magalhães Carvalho

Julho de 2018

© João Luís Ornelas da Silva, 2018

Resumo

A metodologia tradicional para o exercício da Estimação de Estados nos sistemas elétricos, que tem como base o critério de Mínimos Quadrados / Erro Quadrático Mínimo, baseia-se em certas suposições sobre os dados. Contudo, há situações excecionais em que há medições que se tornam contaminadas por erros grosseiros, degradando a estimativa obtida. Quando é feita uma análise posterior dos resíduos da estimativa, as medições suspeitas (se existirem) são sinalizadas e removidas de exercícios futuros. Por vezes, medições que não estão afetadas por erros grosseiros são também sinalizadas e, em certos cenários de erros grosseiros conformes, poderão não ser detetadas todas as anomalias (se mesmo uma).

Na primeira metade da tese, é feita uma discussão sobre a praticidade matemática de ambos o critério dos MQ e o critério da Correntropia na abordagem a diversas situações quanto ao perfil de erro dos dados. Vários autores recentes têm provado a utilidade da Correntropia em lidar com *outliers*, gerando vários ótimos no espaço de soluções do problema de otimização dos quais o ótimo global corresponde à solução mais coerente. Além disso, são também discutidos diferentes tipos de métodos de otimização, nomeadamente métodos numéricos e meta-heurísticas, quanto às vantagens de cada um para diferentes estratégias de exploração. Uma breve discussão é feita sobre perfis de erro e o problema particular dos erros grosseiros conformes nos quais a tese coloca um foco especial.

Na segunda metade da tese, são realizadas simulações de vários casos de erros grosseiros para um sistema teste, desde um erro singular até dois erros conformes. Primeiramente, os casos são abordados pela metodologia tradicional de Mínimos Quadrados. Verifica-se que ela é eficaz em lidar com a maioria dos casos mas não é satisfatória para casos de dois erros grosseiros e, em especial, quando os erros são conformes. De seguida, é aplicada a Correntropia e empregado um método numérico, resultando que, enquanto o número de falsos positivos se torna nulo nos casos bem-sucedidos, as anomalias ficam por detetar em alguns casos de erros conformes. O posterior emprego de uma meta-heurística de enxame para auxiliar na inicialização do método numérico prova que existe uma solução coerente em que são propriamente sinalizadas as anomalias, sendo atingida com uma taxa de sucesso moderadamente satisfatória. Verifica-se também que o método híbrido converge frequentemente para um ótimo local notavelmente decetivo. Com o fim de aumentar a taxa de sucesso, são testadas algumas definições de parâmetros do método, mecanismos adicionais e uma extensão do método híbrido, ao que os resultados confirmam a dificuldade em lidar com os casos problemáticos em questão. Por fim, é testada a utilidade da generalização da Correntropia com o objetivo de auxiliar a exploração através de uma melhor delimitação das zonas de convergência.

Abstract

The traditional methodology for the State Estimation exercise in power systems, which adopts the criterion of Least Squares / Minimum Quadratic Error, is based on certain assumptions on the data. However, there are exceptional situations where measurements become contaminated by gross errors, degrading the estimate. When a posterior analysis is done on the residuals of the estimate, suspicious measurements (if any) are flagged and removed from futures exercises. Sometimes, measurements that are not affected by gross errors are also flagged, since gross errors have cross influence on other residuals of the estimate and, in certain scenarios of conforming gross errors, not all the anomalies are be detected (if even one).

For the first half of the thesis, there is a discussion on the mathematical practicality of both the Least Squares criterion and the Correntropy criterion on the approach to varying situations regarding the data error profile. Many recent authors have proven the utility of Correntropy on dealing with outliers, generating many optima in the solutions space of the optimisation problem, of which the global optimum is the most coherent solution. Furthermore, there is a discussion on different kinds of optimisation methods, namely numerical methods and meta-heuristics, regarding each one's advantages for different search strategies. A brief discussion is had on error profiles and the particular problem of conforming gross errors on which the thesis puts a special focus.

For the second half of the thesis, simulations of many cases of gross errors are made for a test system, from a single error up to two conforming errors. Firstly, the cases are tackled with the traditional Least Squares methodology. It is verified that it is effective in dealing with the majority of cases but is not satisfactory for cases of two gross errors and, specially, when the errors are conforming. Afterwards, Correntropy is applied and a numerical method is employed, resulting in that, whereas the number of false positives becomes null in successful cases, the anomalies remain to be detected in some particular cases of conforming errors. The posterior employment of a swarm meta-heuristic for aiding in the initialisation of the numerical method proves that a coherent solution exists at which all anomalies are properly flagged, being achieved with a moderately satisfactory success rate. It is also verified that the hybrid method often converges to a noticeably deceptive local optimum. With the aim of improving the success rate, some experiments are made with parameter settings for the method, additional mechanisms and an extension of the hybrid method, with results confirming the difficulty in dealing with the aforementioned problematic cases. Finally, the utility of the generalisation of Correntropy is tested with the objective of aiding the exploration by better delimiting convergence zones.

Acknowledgements

During the elaboration of this thesis, I was met with difficulties at some moments along the way that I felt were going to compromise the work that had been and was to be done. I would like to thank the people who helped and inspired me to push through for, fortunately, it turned out fine.

I would also like to thank the people who cheered me on during my academic journey of five long years, yet now seeming so (too) short when I look back. I feel this was a very important chapter in my life and is a big part of the person I am today, as it teached me many important things besides engineering. I met many good friends and made good memories. I wish I could go back to day one and experience it all over again. Sadly, every journey must come to an end and we must move on to the next chapter and continue writing our stories.

Namely, to:

- my thesis advisor, whose positivism sometimes contrasted with my realistic perspective but who made the best out of every situation, and my co-advisor, who also offered his full availability,
- my teachers from FEUP, some of whom were sources of inspiration and made me enjoy engineering,
- my good friends from FEUP both who started this journey with me and who I met halfway, with whom I shared thoughts and had great fun both in and out of college,
- my close childhood friends, with whom I hanged out during vacation, which gave me much needed refreshment for each next school year,
- my dear family, who always rooted for me both in good and not so good times, who gave me invaluable support, motivation and understanding, to whom I attribute the most importance of everything in life and without whom I could not possibly have made it to the end,

thank you.

This work integrated the global effort developed in the project POCI-01-0145-FEDER-016731 INFUSE, supported by the ERDF – European Regional Development Fund (COMPETE 2020 Programme) and FCT - *Fundação para a Ciência e a Tecnologia*.

Index

List of figures

List of tables

Abbreviations and symbols

List of abbreviations by alphabetical order:

EMS	Energy Management System
EPSO	Evolutionary Particle Swarm Optimisation
FN	False Negative
FP	False Positive
GA	Genetic Algorithms
GD	Gradient Descent
KDE	Kernel Density Estimation
LS*	Least Squares
MCC	Maximum Correntropy Criterion
MSE	Mean Square Error (same as LS)
PDF	Probability Density Function
PSO	Particle Swarm Optimisation
SE	State Estimation
SCADA	Supervisory Control and Data Acquisition
TN	True Negative
TP	True Positive

**Not to be confused with Local Search, for which this abbreviation is sometimes used in works on the field of optimisation.*

List of symbols:

α	scaling factor (Correntropy function parameter)
ß	shaping factor (Correntropy function parameter)

Chapter 1

Introduction

State Estimation (SE) is an exercise, conducted at power system control centres, that aims at providing system operators with a coherent image of the system state. The need for a SE procedure derives from the fact that the measurement data arriving at a SCADA in the control centre are contaminated by errors, usually small, caused by a variety of sources, and therefore they do not form a coherent set compatible with the Kirchhoff Laws governing power networks.

The thesis' focus is the exploration of conditions and consequences of two simultaneously hard conditions sometimes prevailing in power systems State Estimation: the occurrence of gross errors in data (not small errors) and, in particular, conforming gross errors. This exploration aims at gaining insight on the difficulties of successfully handling such situations and developing methodologies mitigating the consequences of those problems.

These difficulties have been recognized since long, but no consistent effort is known to address them in depth. In fact, in most publications in the field, authors generally claim that their test sets are exempt of conforming errors and that their algorithms, developed to deal with several circumstances, are nevertheless applied to data sets clean of conforming errors. In order to tackle this issue, we must firstly decide on how we define error conformance, since there may be differing opinions. In this thesis, such a definition will be given hopefully with clarity and comprehension – but, in an introduction of the concept, we may say that conforming errors are errors that are mutually coherent / concordant; for instance, a negative error in a power injection and the same error in a near-by load. This error concordance (contrary to a random perturbation of divergent errors) creates deceiving landscapes rendering optimization difficult – in fact, if all measurements would have conforming errors in the same sense, this would bias the state estimation into believing that the system state is something different altogether from the real state.

To explore the features of such difficult problem and recognize the characteristics of the landscape of the optimization problem that represents the SE exercise, we had to search for an adequate technique and an experiment environment: a population-based meta-heuristic. The following is the rationale behind such choice:

1. It is common knowledge that SE based on a Least Squares criterion (also denoted as Minimum Square Error) is not adequate to deal with data sets with gross errors. These contaminate the estimation, introducing large deviations in the

regression surface interpolating the data and causing the erroneous correction to several measurements, suggesting that they had large deviations even if this was not the case.

2. To circumvent this effect, recently some authors have suggested the use of a cost function such as Correntropy, which has properties like an M-estimator and, in theory, is able to deal with outliers, leading to their natural rejection (outliers, since the gross errors would be far out of the regression surface passing through the remaining data).

3. Such function has properties that may lead to an estimation ignoring the gross errors (based on a regression surface that approximates the "healthy" data and leaves out the erroneous data). However, in order to do this, the objective function is found to have local optima – at least, one optimum that respects to the "true" / most accurate / most likely optimum and one corresponding to a regression surface interpolating through the outliers.

4. Furthermore, it seems that conforming errors may induce the presence of deceiving landscapes, where the optimum is found in a very narrow region while the remaining landscape seems to hint that the optimum is elsewhere to be found.

The search for multiple optima excludes the possibility of using the classical Gauss-Newton iterations, which is otherwise traditionally adopted when the regression criterion is the quadratic LS.

A population based meta-heuristic will give more flexibility to the search for multiple optima, thus the choice for Evolutionary Particle Swarm Optimization. However, to speed up the search and make it more efficient, a hybrid meta-heuristic/gradient approach was devised.

Lastly, the thesis explores the possibility of adopting Generalised Correntropy, either instead of or even alongside Classical Correntropy, based on the generalised definition of Gaussian functions, to explore different possibilities by providing the optimization landscape with new shapes, which might favour the optimization progression.

Because the adoption of a meta-heuristic may prove to be heavy in computing terms, and as the thesis is focused on understanding the behaviour of distinct models and the interaction of the cost function landscapes with the algorithmic performances, the thesis adopts the traditional DC model to represent power flows. This renders the problem more manageable in terms of computing effort and allows easier interpretation of the difficulties that are met.

The conclusions obtained with solving a DC SE problem may be extremely helpful if one aims at further extending the techniques to a full AC model.

Chapter 2

State of the art

2.1 – State Estimation in power systems

2.1.1 – Paradigm

In the management of power systems, unpredictability has to be dealt with constantly. It is never possible to follow an exact operation planning, which might have been made ahead for system operation, as some elements of the system are extremely dynamic, measuring devices have their own precision classes, data transmission is not fully reliable and may add noise and data capturing may not be made simultaneously for all measurements, leading to incoherence. Also, exceptional incidents in the grid may occur, such as faults, which must be dealt with as they come, often requiring immediate corrective actions – thus, the need for a precise image of the system state.

In order to deal with all sources of imprecision and uncertainty behaviour and maintain technical safety in the operation, an information system (residing in a SCADA / EMS – Supervisory Control And Data Acquisition / Energy Management System) must monitor the power system. This system is used to periodically gather data from many points in many sub-grids that are part of the whole grid. Usually, the data are transmitted to local management centres where they are processed and then sent to centres at higher levels of management, hierarchically.

Due to certain factors, which may introduce error in the data, either during the measurement process or during the transmission process, affecting data quality before they arrive at the control centres, data gathering units are massively deployed. Usually, there are many more measurements than are minimally needed for assessing the state of the power system at each instant.

Redundancy can help in making a more accurate assessment by compensating for data imperfection or invalidity / unavailability. Since measurements are imperfect (affected by error) and there are more of them available than there are variables to be predicted / estimated, we must develop a methodology for taking all available data into account and

building an estimate which is coherent with the Kirchhoff Laws governing the power network. This is the goal of SE.

The three main steps of a classical power system SE procedure are:

1. topology processing. This refers to the assessment of the current topology / configuration of the grid – which branches are in or out of service; which switches and breakers are closed or open. This will determine the model of the power system that is to be considered for the next steps.
2. state estimation calculation. This is the step at which an estimate of the system state is determined, based on the topology that was obtained in the previous step and on all the available data from the measurement process.
3. bad data processing. Classically, an analysis of the results from the previous step is made with the goal of identifying potential anomalies in the data. Suspicious data are flagged and removed from future procedures until the problem which originated the anomalies is fixed.

In modern days, steps 2 and 3 have been subject to attempts to merge them in a single step, so that bad data processing is not made having as its departure point a result already contaminated by bad data. Furthermore, there have been proposals to fuse steps 1, 2 and 3 in a single procedure, to avoid processing bad data analysis over results obtained with an erroneous network topology.

In this thesis, topology estimation is not addressed. However, the fusion of steps 1 and 2 is in fact addressed through the use of a mathematical modelling that originates this fusion as a natural process.

2.1.2 – Optimisation problem

Mathematically, the problem of SE is formulated as a special case of a generic problem of regression. The idea is to minimise a "cost" which is translated from the "difference" between a set of samples / measurements of the system quantities and our model for the system. The model formulates relations between the system variables or, in other words, formulates outputs as functions of inputs.

With these regards, the problems of general regression and of SE are optimisation problems. Their concept only is justified if we are faced with data redundancy, imperfection and, consequently, conflict. Having many more samples than unknowns to be determined makes it impossible to determine a single set of values for the unknowns that match every possible combination of measurements.

The key differences (from a theoretical perspective) between general regression and SE are as follows:

Regression (in general)	State Estimation (in particular)
Samples usually correspond to states at many experimental instances.	Samples correspond to the power system state in a single instant in time. While measurements are known, the state variables are missing and to be determined.
The model structure is a hypothesis for the system behaviour and its parameters are to be determined.	The model structure is assumed as correct and the state variables (usually voltages/angles) to be estimated correspond to the parameters in a general regression model.

The difference between a certain measurement and the respective estimated value, as a consequence of the estimated state variables, is denoted as a residual. This is different from an error, which is the difference between a measurement and the true value for a quantity (therefore being unknown because the true value is unknown as well).

The quality of any regression process is evaluated by a cost function, establishing some merit order on vectors in the state space or voltages/angles. Also, each measurement may be affected by a factor which expresses a degree of confidence in it value, in terms of how accurate we believe it to be – usually related to the precision of the measuring equipment, but other factors may interfere as well, such as asynchronism in measurement collection, unreliability in data transmission and equipment failures.

The goal function, which can be written as a cost function to be minimised, will be a sum of terms. Each term corresponds to an assumed cost associated with the respective residual. The translation from residual to cost is defined by a criterion.

$$f = \sum_{k} [w_k * cost(r_k)] \quad , \qquad r_k = z_k - h_k(x) \tag{2.1}$$

where:

- f function to be minimised;
- w measurement weight;
- cost function to be used as the criterion;
- r residual;
- z measurement;
- h model / equation of the quantity, as function (Kirchhoff Laws) of the system variables;
- x system variables.

Depending on what criterion we choose, the cost of a residual might change and, therefore, so might the “best” estimate, which is the global optimum of the problem. In typical scenarios where the data are affected only by minor errors (noise), a simple and traditional criterion (the Minimum Square Error or Least Squares criterion) may be employed to produce a fairly accurate estimate. In an upcoming chapter, we will discuss other possible scenarios that could require either additional processing of the data (which is the next step of the SE procedure) or the use of an alternative criterion.

2.1.3 – Model for a power system

Regarding our model for power systems, power flows in branches (and, by extension, power injections in buses) are formulated as functions of voltage drop and angle difference between buses. Each voltage magnitude and angle are system variables to be determined.

The AC model for the behaviour of active and reactive power flows and injections as functions of voltage magnitudes and angles is represented by the following set of equations:

$$\begin{aligned} P_{ij} &= -G_{ij}*u_i^2 + u_i*u_j*(G_{ij}*\cos\theta_{ij} + B_{ij}*\sin\theta_{ij}) \\ P_i &= u_i*\sum_k \left[u_k*(G_{ij}*\cos\theta_{ij} + B_{ij}*\sin\theta_{ij})\right] \\ Q_{ij} &= B_{ij}*u_i^2 + u_i*u_j*(-B_{ij}*\cos\theta_{ij} + G_{ij}*\sin\theta_{ij}) \\ Q_i &= u_i*\sum_k \left[u_k*(-B_{ij}*\cos\theta_{ij} + G_{ij}*\sin\theta_{ij})\right] \end{aligned} \tag{2.2}$$

where:

- P_{ij} active power flow from bus i to bus j;
- P_i active power injection at bus i;
- Q_{ij} reactive power from bus i to bus j;
- P_i reactive power injection at bus i;
- u_i voltage magnitude at bus i;
- θ_i voltage angle at bus i;
- G_{ij} element i j of the AC conductance matrix;
- B_{ij} element i j of the AC susceptance matrix;

Based on our knowledge of the typical operation of power systems in the real world, it is possible to come up with some simplifications / approximations in order to relax the highly non-linear equations of the AC model into linear ones. This relaxation greatly reduces the model's mathematical complexity while still keeping it fairly reliable, considering that the real world system's operation is kept within certain boundaries.

The DC (simplified) model for the behaviour of active power flows and injections as functions of voltage angles is made up of the following set of equations:

$$\begin{aligned} P_{ij} &\approx -B_{ij}*\theta_{ij} \\ P_i &\approx -\sum_k \left[B_{ik}*\theta_k\right] \end{aligned} \tag{2.3}$$

where:

- P_{ij} active power flow from bus i to bus j;
- P_i active power injection at bus i;
- θ_i voltage angle at bus i;
- B_{ij} element i j of the DC susceptance matrix.

2.2 – Regression criteria

2.2.1 – Least Squares

The criterion of Least Squares is the standard, most widely known and most employed rule in the field of data fitting for both linear and non-linear models. This is also the traditional criterion of choice for SE in power systems [1][3].

LS is the rule of translating each residual into a cost by squaring it.

$$cost(r) = r^2 \tag{2.4}$$

With this criterion, the cost of a residual grows at a quadratic / parabolic rate relatively to the growth of the residual. Therefore, the marginal cost (variation of cost by variation of residual) grows at a linear rate.

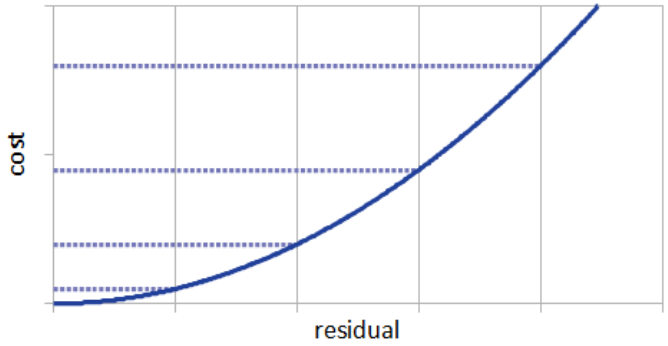


Figure 2.1 – LS cost function.

When employing LS, the estimate to be determined ends up being a weighted quadratic average of the samples. A marginal cost is (approximately) the variation in the total cost by variation of the residual, either adding or subtracting. Since the marginal cost of a larger residual is larger than that of a smaller residual, the tendency is for larger residuals to be brought down and for smaller ones to be amplified, as a trade. A difference between marginal costs indicates that it is possible to reduce the total cost, and such is the goal of an optimisation problem.

LS is a rule of regression by compromise. Every sample in the data set has an unrestrained (despite weighted) impact on the estimate. If samples are very coherent / closely distributed, the obtained estimate will be very similar to them as well. In this case, LS fits the bill perfectly and there is no need for additional data processing or other criteria.

The issue comes when samples are not as coherent as we would like them to be. The presence of a measurement vector component which, by whatever reason, has a very high deviation from a value coherent with the rest of the components (an outlier) might significantly change the estimate that is obtained. Since there would be a residual with a very large marginal cost if the estimate was to be placed among the other measurements and since we are fitting by compromise, the incoherent measurement “pulls” the estimate towards itself. The greater the deviation, the greater the displacement of the estimate.

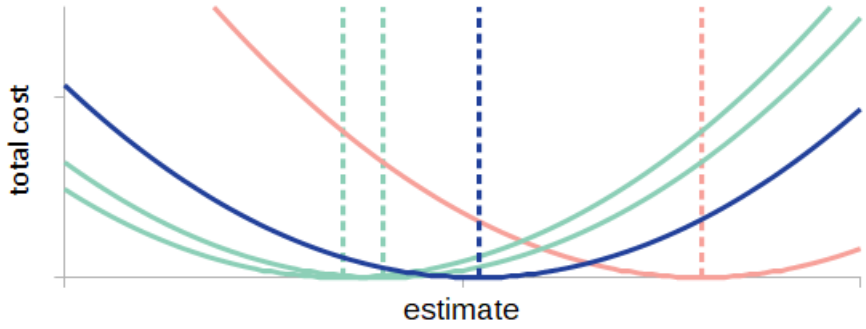


Figure 2.2 – LS estimate.

2.2.2 – Correntropy

Instead of dealing with the variance of the residual distribution, as in the LS method (that actually operates on residuals calculating the summation of quadratic deviations from the mean), we may resort to dealing with an estimation of the underlying density function of the residuals. To do this, we must have an estimate of such function, which can be achieved with Kernel Density Estimation – a concept credited to Emmanuel Parzen [9] and Murray Rosenblatt. In this process, a kernel / window function, which is non-negative and has a unity integral, is applied to a set of samples. Their summation, normalized by the number of samples, gives us an unbiased approximation for the Probability Density Function of a random independent variable, from which only discrete samples are known.

$$f(x) = \frac{1}{n} * \sum_{k} K(r_k - x, \sigma) \tag{2.5}$$

where:

- f residual / error density function estimation;
- n total number of samples;
- K kernel;
- r residual;
- x random variable;
- σ Parzen window width.

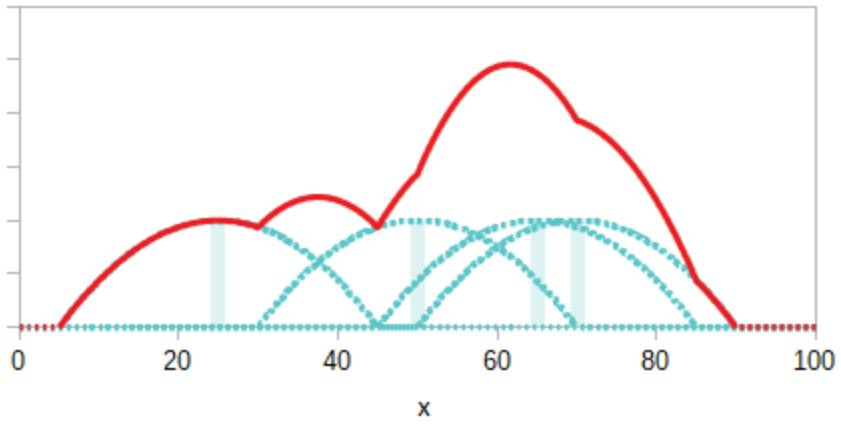


Figure 2.3 - Probability density function approximation of a random variable by KDE (with Epanechnikov kernels).

In SE, we will apply the Gaussian kernel to the residuals of a state estimate so as to determine its likelihood of being the state of operation for the power system.

Correntropy [12][13][18] is an Information Theoretic Learning concept that measures the similarity between two random variables.

$$V_\sigma(X,\ Y) \ = \ E[G(X - Y,\ \sigma^2 I)] \tag{2.6}$$

where:

- V_σ Correntropy with deviation σ;
- E expected value;
- G Gaussian kernel;
- X, Y random variables;
- I Identity matrix (independence among XY sample pairs, or errors).

In practice, the true joint PDF is unknown and, therefore, only an estimate form a sample of size n is possible.

$$\hat{V_\sigma}(X,\ Y) \ = \ \frac{1}{n} * \sum_k G(x_k - y_k,\ \sigma^2 * I) \ = \ \frac{1}{n} * \sum_k G(\varepsilon_k,\ \sigma^2 * I) \tag{2.7}$$

where:

- x, y samples of the random variables;
- ε error.

An Information Theoretic Learning criterion to optimize systems depending on parameters, under supervised learning, is to maximize the Correntropy of the output error distribution, known as the Maximum Correntropy Criterion. This has several possible interpretations, one being that we try to maximize the similarity between the system output and the target or desired values.

When applying this concept to SE, one may think of maximizing the similarity between the vector of measurements and the vector of electric values that result from the estimated values for the state variables (usually voltages and angles).

The Correntropy function, unlike LS, is not unimodal. The kernel function originally adopted is the Gaussian function. The cost term for each residuals becomes:

$$cost(r) = 1 - exp(-|r/\alpha|^2) \quad (2.8)$$

where:

- α scaling factor, denoted as Parzen window width for other versions of the Correntropy function*.

**The function described above is written in such way to match with how the SE problem was defined (minimisation) in section 2.1.2. Its most usual presentation is as a maximisation process (MCC – Maximum Correntropy Criterion).This also brings the formulation closer to the LS formulation. The LS process induces a Euclidean metric in the search space; Correntropy is associated with a distinct metric (CIM – Correntropy Induced Metric* [14]*) which has a form similar to the objective function adopted in this thesis.*

α is related to the range of sensibility / the width of the kernel. It defines the point at which the function becomes practically insensitive to a residual.

For smaller residuals, the cost function bears resemblance to LS regarding the growth of the marginal cost. However, if a residual grows indefinitely, its cost becomes restrained and will converge asymptotically to a certain positive limit (that, regarding the above version of the Correntropy cost function, equals the unit). As such, its marginal cost becomes fairly small, falling down to zero. The cost of a very large residual is practically insensible to a small variation of the residual, contrasting greatly with LS.

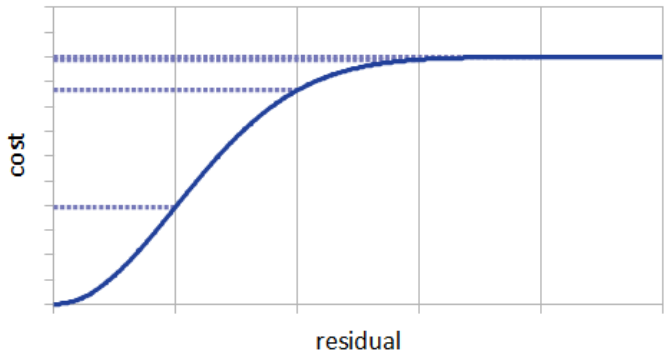


FIgure 2.4 – Correntropy cost function.

As was discussed in section 2.2.1, when using LS, the marginal cost of a residual is highly impacting on the result of the SE exercise. Residuals with larger marginal costs pull the estimate away from residuals with smaller ones, resulting in that a residual is brought to a certain degree so that all marginal costs cancel out. It was also discussed that a sample that "does not fit well" with the remaining data would cause a residual to have a large marginal cost, noticeably displacing the estimate relatively to the case of it not being present.

Correntropy is an attempt at preventing the unrestrained impact of marginal costs on the estimate. The aim is to split the data into different groups of concordant samples (that are mutually coherent), meaning that these samples have small deviations relatively to one

another. The largest group, which would cover the regular samples, should generate the global optimum, whereas less numerous groups, which would cover anomalous samples, should generate local optima. The marginal cost of a large residual at a certain estimate should be null so that it can no longer disrupt the estimate. In other words, it's as if samples with high deviations are "expelled". In order for it to work, the parameters for the cost function must be adequately set. A sufficiently narrow window (small enough α), should be applied.

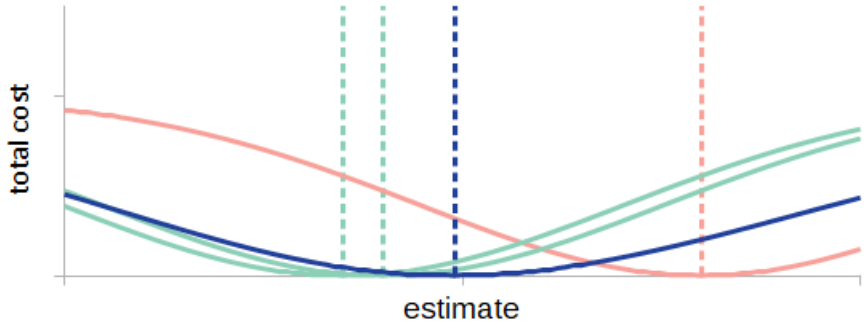


FIgure 2.5 – Correntropy estimate with insufficient sensibility.

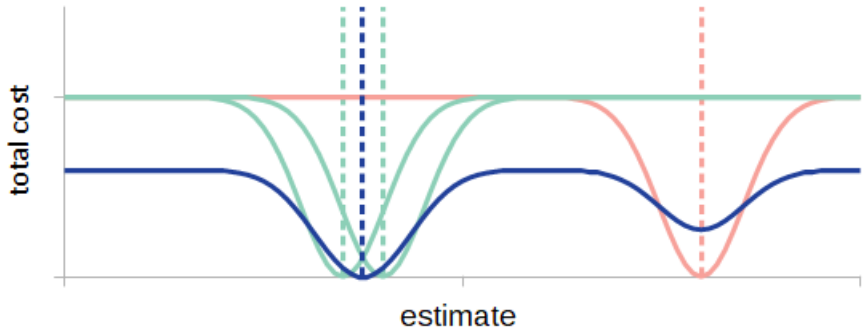


FIgure 2.6 – Correntropy estimate with adequate sensibility.

The cost of the global optimum in the goal function is practically equal only to the sum of the (restrained) costs of the few residuals with respect to the outliers. For the previously enunciated version of the Correntropy function, it equals roughly the number of said measurements that exist in the data.

Although the globally optimal solution corresponds to a much more likely / accurate estimate, there is the issue that the landscape will acquire a rough shape with more local optima as there are more incoherent samples. This is particularly problematic if our intention is to employ classical numerical methods to solve the problem. We will discuss in section 2.3.1 that these methods are very reliant on initialisation. If a numerical method is to be initialised at a typical rule-of-thumb solution, there is a risk that it becomes trapped in a local optimum which happened to be "on the way". A random initialisation is possible but it might require multiple runs in hope of placing the agent in the desired convergence zone.

A possible way to ease the optimisation exercise could be to employ a dynamic landscape strategy. The position of the global optimum depends on the chosen sensibility. When there is a variation of the sensibility, the optimum shifts its position. It is hoped that, if

α was to be gradually increased, the optimum would also "move" gradually along the landscape. It would make it easier for an agent to follow the optimum's trail, given a slow enough sensibility variation and enough time for the agent to relocate properly.

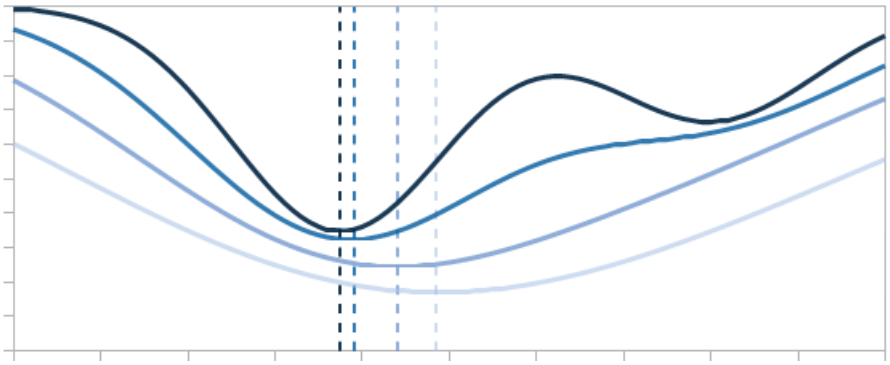


Figure 2.7 – Pursuit strategy (varying α).

Even so, this behaviour is not guaranteed for all cases and depends on the model. It is also possible that a global optimum and a local optimum that are separated by a finite distance suddenly switch roles, becoming local and global respectively.

There is also the necessary caution of not being too aggressive on the sensibility setting. If α is too small, measurements which would be considered coherent might be themselves split into different groups of concordance. In other words, small deviations are scaled up to become large deviations. A greater number of concordance groups translates to more local optima on the landscape.

$$\alpha \to 0 \Rightarrow cost(r) \simeq \begin{cases} 0, & r=0 \\ 1, & r\neq 0 \end{cases} \tag{2.9}$$

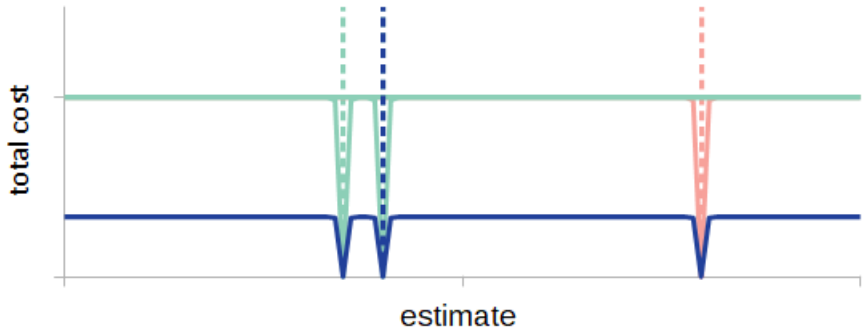


Figure 2.8 – Correntropy estimate with exaggerated sensibility.

2.3 – Methods for optimisation

2.3.1 – Numerical methods

All problems in the field of engineering can be formulated mathematically. However, not all problems can be solved analytically or, at least, require too many symbolic constructions and operations to be performed for that end. Symbolic mathematics is much more expensive than number crunching, in terms of computational power.

Numerical analysis is an iterative process for solving these more complex problems. While it does not ditch mathematical analysis in its entirety, it does ditch symbolic operations and, instead, solves the problem by making use of numerical properties.

Some of these methods are based on the analysis of the function's gradient. The gradient is obtained by differentiation and corresponds to the (approximate) per-unit variation of the function's return value relative to a variation of the value of each variable, depending on the point of analysis. It is also commonly described as a "slope".

$$f'(x) = \frac{f(x + \Delta x) - f(x)}{\Delta x}, \quad \Delta x \to 0 \tag{2.10}$$

where:

- f function of variables;
- f' gradient;

Since a function's gradient, which is a vector, always points towards whichever "path" leads to an increase in the goal function's return value, that is what will happen if we follow such vector with a small enough step. Reversely, following in the opposite direction leads to a decrease.

The Gradient Descent method (or Ascent, depending on the goal) is a first-order optimisation method. Iteratively, the method will move an agent across the landscape, always following the gradient with a step that is (usually) proportional to the gradient. In other words, it follows the variation of the landscape's height. The step factor may either be fixed or change with progress, based on some criterion.

$$x_{next} = x + s*f'(x) \tag{2.11}$$

where:

- x current solution;
- x_{next} next solution;
- s step – positive for maximisation, negative for minimisation.

GD follows the "steepest" path to an optimum. This approach guarantees that an optimum, which is a point of null gradient, is reached with a degree of precision that depends on how small the step factor is. If the landscape has only one optimum, such point is globally optimal and GD will always find it, regardless of initialisation.

The caveat is that, if the landscape has multiple optima, there is a possibility that the obtained solution is locally (rather than globally) optimal, meaning it is not the best solution.

This is why adequate initialisation might be vital for GD's success, when employed to solve complex non-linear problems. Our knowledge of some problems may help us to build an initial solution intelligently with either heuristics or rules of thumb, so as hopefully to place the agent in the convergence zone that contains the global optimum, although this might fail sometimes. Heuristics and rules of thumb usually are employed with the belief that the problem's data meets certain typical conditions and this is not always the case.

2.3.2 – Meta-heuristics

Since the early 50s, researchers had already been studying the concept of "meta-heuristics" that were supposed to be methods that relied on mechanisms of improvement whose main driving factor is randomness, as an answer to the difficulty on approaching non-linear complex optimisation problems. Arguably the first noticeable method of sorts was Simulated Annealing, which consists of causing (small) random perturbations to a candidate solution and keeping either the original solution or the newly obtained one, based on whether or not the solution has improved and, in case it has not improved, according to a time-decreasing-probability. This method employs only one agent. Other methods had been conceived that manage a population of many agents and employ mechanisms for information exchange and coordination between agents. Despite all this, arguably the main culprit for setting the development of robust meta-heuristics for solving complex problems in motion was John Henry Holland, with his book *Adaptation in Natural and Artificial Systems* [24] that showcased the fairly popular Genetic Algorithms.

Meta-heuristics, unlike numerical methods, throw mathematical analysis out the window and, instead, employ mechanisms that search the surroundings / neighbourhood of candidate solutions, with reliance on random number generation to guide the search. The difference between a meta-heuristic and a "brute force" Monte Carlo approach is that, although a meta-heuristic relies on randomness, it searches the landscape not by sampling but rather by improvement.

Meta-heuristics may be applied to solve any problem that can be formulated as an optimisation problem. The goal function is usually addressed as the "fitness" function that evaluates how "fit" an individual / candidate solution is for the problem at hand.

The aim of meta-heuristics is not to be faster than classical numerical methods in finding solutions (since randomness is arguably the slowest strategy for that end) but rather to perform an extensive search of the landscape and, thus, to be more accurate on finding the best solutions.

The Genetic Algorithms family [26] was the first widely-known set of improvement mechanisms to mimic the biological evolution of living beings and their interaction with a natural environment. Conceived by John Henry Holland, it draws inspiration from Darwin's theory on the evolution of species and suggests the use of recombination and mutation as mechanisms for gradually improving the fitness of agents.

The recombination mechanism mimics the way that living beings generate offspring. A child solution / agent could be generated either simply as a perfect copy of its parent solution or from crossing the traits (variables' values) of each parent with a respective mate parent (or multiple) that could be chosen either at random or based on some criterion: the fittest agent of the population; the next agent in order; so on.

$$child[x_i] = par_1[x_i] \text{ or } par_2[x_i] \text{ according to } \rho_{crossover} \quad (2.12)$$

where:

- child — child solution;
- par — parent;
- x_i — variable i, to be inherited;
- $\rho_{crossover}$ — crossover rate.

Traits may also be passed down either directly or as the result of arithmetical operations between themselves. Differential Evolution, which is a variant of GA, credited to Kenneth Price and Rainer Storn. suggests building a "donor" vector from a series of arithmetical operations between the traits of multiple mate parents and then finally crossing it with the original parent.

The mutation mechanism mimics the way that living beings undergo changes in their anatomy in order to better fit in with the environment. The traits of a newly born child may suffer perturbation. Whatever traits are to be mutated / perturbed could be chosen according to a rate. The severity of the mutation could be quantified by a factor. Each one of these parameters could be either fixed or time-varying according to certain intelligent rules.

$$child[x_i] \mathrel{+}= 0 \text{ or } gauss(0, \sigma_{mutation}) \text{ according to } \rho_{mutation} \quad (2.13)$$

where:

- x_i — variable i, to be mutated;
- $\sigma_{mutation}$ — mutation factor;
- gauss — random Gaussian distribution;
- $\rho_{mutation}$ — mutation rate.

After these mechanisms have been applied, children and parents compete in an intra-family tournament. The fittest individual from a family becomes that family's original parent for the next generation.

Particle Swarm Optimisation [27][28] is a more recent retake on the concept of emulating the behaviour of living beings for solving optimisation problems, although it moves away from a biological perspective and into a psychological and sociological one. Conceived by James Kennedy and Russell Eberhart, in an article of the same name (1995), it suggests the implementation of personality and tendency into the agents of a population, in terms of habit and trust (in oneself and in others).

The velocity equation is related to the shift in the agent's position for each instant (iteration) and essentially is the sum of three parcels – habit, memory and cooperation. Habit is proportional to the agent's velocity for the previous instant. Memory and cooperation are proportional to the difference vectors from the agent's current position to specific respective references. In the memory parcel, such reference is the agent's personal reference (its own best finding). In the cooperation parcel, such reference is the swarm's shared reference, based on what swarm topology is being used*. Each of the three terms is affected by a

weight. The memory and cooperation terms also are affected by a random uniform factor, so as to promote a more randomised and, thus, wider search.
**Usually, a "global" topology is used, where the shared reference for every particle is the best solution yet found by the whole swarm. Other topologies are possible, although they are not used as often.*

$$v = w_{habit} * v + w_{memory} * r() * (b_{personal} - p) + w_{cooperation} * r() * (b_{shared} - p) \quad (2.14)$$

where:

- v velocity;
- p position;
- $w_{...}$ (habit; memory; cooperation) weight;
- $b_{...}$ (personal; shared) reference;
- r random uniform distribution.

The position equation formulates the agent's position for the current instant as the addition of the velocity to the position for the previous instant. It is possible to affect the velocity with a factor, after it has been updated, when adding it to the position, although it is unclear whether or not this would have a desirable effect.

$$p = p + v \quad (2.15)$$

Once all positions are updated, they will compete with the current references. Whichever is the fittest becomes the reference for the next instant.

While GA and PSO seem to have radically different architectures, they share some similarities. Apart from their selection mechanisms, both also employ mechanisms of crossover between existing admissible solutions in order to generate new ones. In GA, agents are crossed with one another during recombination. In PSO, agents are crossed with their references though the velocity equation.

It is arguably difficult to guess which method is better suited for tackling a specific kind of problem. Yet, we could argue that a certain mechanism is better suited for a specific kind of exploration, either wider or narrower:

- GA's mutation mechanism directly causes small perturbations to a solution, checking their immediate surroundings. A narrow search may be useful when an agent is positioned somewhere near an optimum, since it gives a better chance of quickly spotting the optimum.
- PSO's movement mechanism allows agents to fly across the landscape with big steps, spreading the search to various zones of the landscape. A wide search may be useful when the agents haven't explored much of the landscape, since it may enable an agent to jump into a different and possibly more promising convergence zone.

On these thoughts, perhaps with a junction of the mechanisms of both methods could be done, we could take advantages from each method in order to improve the search at each stage of the optimisation process.

Evolutionary Particle Swarm Optimisation [30], credited to Vladimiro Miranda and Nuno Fonseca and firstly exposed in an article containing its name (2002), is an attempt at marrying the concepts of GA and PSO together. It borrows the particle swarm's architecture and adds evolutionary mechanisms on top.

In EPSO, unlike PSO, the weights are not restrained under a specific static strategy, are different for each agent and may change in any possible way during the optimisation problem. This is achieved through perturbing them with mutations in a similar way to how child solutions are perturbed in GA. However, the mutated weights do not belong to the original particle, belonging instead to a child particle that is to be spawned afterwards. The intention is to then apply selective pressure and keep whichever set of weights placed the respective particle in a better position.

$$\begin{aligned} w_{child,habit} &= w_{par,habit} + gauss(0,\sigma_{habit}) \\ w_{child,memory} &= ... \\ w_{child,cooperation} &= ... \end{aligned} \tag{2.16}$$

where:

- $w_{child, ...}$ (habit; memory; cooperation) weight of the child particle;
- $w_{par, ...}$ (habit; memory; cooperation) weight of the parent particle;
- $\sigma_{...}$ (habit; memory; cooperation) weight's mutation factor.

The references for each agent are also subject to mutation, although they are only intended to be "fuzzy", meaning they are in effect only for the current instant and do not override the original ones. It is possible either to mutate the shared reference only once (sharing it among the whole swarm), to mutate the shared reference differently for each agent, to mutate the personal references or to apply any combination of these approaches.

The velocity and the position must be updated for both the parent particle and a newly spawned child particle. Both particles have the same starting velocity and starting position. Since the mutation mechanism already comports randomness, the random uniform factors that would otherwise be present in PSO's velocity equation are absent in EPSO. Another reason for such approach is that "unintelligent" random factors could conflict with the adaptive strategy for the weights and have undesirable effects. Additionally, a "stochastic star" topology is adopted, in order to tune the degree of aggregation / communication of the swarm. The cooperation term should be present in the velocity equation only according to a certain rate. The intention is to allow particles to momentarily free themselves from their attraction towards the shared reference.

$$\begin{aligned} bb_{personal}[x_i] &= b_{personal}[x_i] + gauss(0,\sigma_b) \\ bb_{shared}[x_i] &= ... \end{aligned} \tag{2.17}$$

where:

- $bb_{...}$ (personal; shared) fuzzy reference;
- x_i variable i, to be mutated;
- σ_b (personal; shared) reference's mutation factor.

$$
\begin{aligned}
v_{par} &= w_{par,habit} * v + \\
&\quad w_{par,memory} * (bb_{personal} - p) + \\
&\quad c * w_{par,cooperation} * (bb_{shared} - p) , \\
c &= 0 \text{ or } 1 \text{ according to } \rho_{cooperation} \\
v_{child} &= w_{child,habit} * v + \\
&\quad \ldots
\end{aligned}
\tag{2.18}
$$

where:

- v_{par} parent particle's velocity;
- v_{child} child particle's velocity;
- $w_{par, \ldots}$ (habit; memory; cooperation) weight of the parent particle;
- $w_{child, \ldots}$ (habit; memory; cooperation) weight of the child particle;
- $\rho_{cooperation}$ cooperation rate.

$$
\begin{aligned}
p_{par} &= p + v_{par} \\
p_{child} &= p + v_{child}
\end{aligned}
\tag{2.19}
$$

where:

- p_{par} parent particle's position;
- p_{child} child particle's position.

After all characteristics have been updated, the population is subjected to two stages of selection:

1. Children and parents compete in an intra-family tournament, as in GA. Whichever agent is the fittest overrides the previous one with its position, velocity and weights.
2. Agents now compete with their current references, as in PSO. Whichever is the fittest becomes the reference for the next instant.

Although having two selection stages instead of only one infuses EPSO with the possibility of adaptation, it has the downside of doubling the number of evaluations that must be performed for each iteration. However, the extra computational effort may be compensated by an improvement in the exploration and convergence capabilities of the method.

Chapter 3

On errors

3.1 – Noise and anomalies

Summing up section 2.1, we need a SE process because the data that are gathered are imperfect, being affected by errors. Fundamentally, two error profiles may occur, namely:

- noise – small errors which spread across the whole data set. This usually is due to:
  - imperfections in the characteristics of the instrument;
  - asynchronism / lag between measurements;
  - transmission noise.
- anomaly – gross errors which contaminate specific individual measurements. This may be due to:
  - instrument malfunctioning;
  - incorrect topology (from the topology processing step).

The issue of topology errors has different implications than the other issues and is dealt with in a particular way. Since it deserves a study by itself, it will not be addressed in the present work.

Noise rarely ever causes any issues. The first run of the SE procedure finds an estimate of the state which is fairly convincing and coherent with the data and there is no need for further data processing.

Anomalies require some extra effort. An analysis on the residuals of the first obtained estimate will likely reveal a noticeable degree of incoherence between some data, hinted at by the sizes of some residuals which are significantly bigger than that of residuals caused by typical noise. In this situation, we can be certain that the obtained estimate is not at all accurate and must be restored. Bad data processing is then done with this objective. Suspicious measurements, characterised by large residuals, get flagged. Afterwards, a second estimate for the state is obtained, this time without considering the flagged data. This may be repeated until no more anomalies are present.

It is important to note that data which gets flagged as “bad” may actually not be contaminated.

3.2 – Conforming errors

The traditional methodologies for identifying and removing bad data from the SE process are fairly successful in tackling most cases of anomalies, although some. However, there are exceptional cases where the methodologies may fail in spotting the anomaly.

The difficulty depends on specific details regarding its characteristics. Each anomaly case is different and is mainly distinguishable by:

- how many gross errors exist in the data set;
- which system variables participate in the equations for the quantities respective to the affected measurements (and how);
- how conforming the gross errors are among each other, if there are multiple.

Conforming errors may be exceptionally difficult to deal with. They will be a big focus in the upcoming experiments of the thesis. In these cases, these (multiple) errors usually "work against" a single measurement that, in reality, may not be contaminated. This results in the appearance of an optimal estimate with a misleadingly large residual that is actually respective to the uncontaminated measure, wrongly flagging it to be swept off. In power systems, error conformance may happen when errors of similar magnitude occur in:

- two buses which are met by a same branch. If the errors in the buses' power injection measurements are opposite in polarity, they are conforming and cause conflict with the measurement of the branch's power flow.
- one bus and one branch which meet / are adjacent. If the errors in the measurements of the bus' power injection and of the branch's power flow towards the other bus are equal in polarity, they are conforming and cause conflict with the measurement of the other bus' power injection.

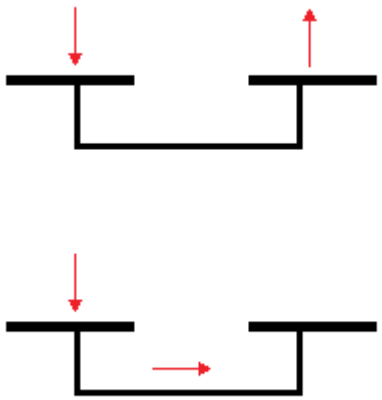


Figure 3.1 – Possible cases of error conformance.

Let us consider the example of two buses, 1 and 2, and their interconnecting branch, which are part of a larger power system. The system is operating at a stationary point. Bus 1 has generation and bus 2 has load. A transitory effect occurs. Now, the measurement at bus 2 accuses an increase in load and the measurement at the branch accuses a similar increase in flow (from bus 1 to bus 2). However, the measurement at bus 1 accuses a decrease in

generation as well. We can confidently declare that there is an anomaly in the measurement process, as there is a great mismatch of injection. If considering this section of the system in isolation (ignoring the neighbouring elements), one of two cases has happened:

1. The measurements at both the branch and bus 2 are contaminated. There actually was a decrease in load at bus 2 and a decrease in flow, conforming with lesser generation at bus 1.
2. The measurement at bus 1 is contaminated. There actually was an increase in generation at bus 1, conforming with greater flow and greater load at bus 2.

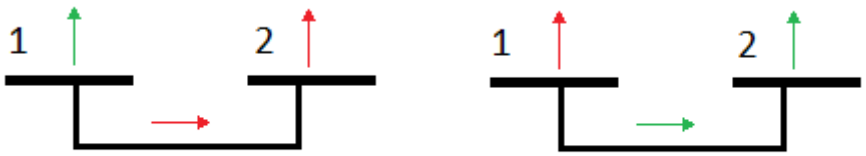


Figure 3.2 – Conflict between measurements.

The tie breaker for this issue can only be the remaining data. The measurements of the power flows and injections of the suspicious elements must also match with the measurements for the neighbouring elements. Mismatches generate residuals at the estimate. If, for that estimate, there are many residuals that are noticeably larger than residuals which are usually caused by noise, it hints that the estimate is not coherent with the respective measurements. Of course, we must find these estimates in order to be able to analyse them in the first place. The criterion and method to be used in solving the SE exercise is decisive for that end.

In order to help in understanding the impact of these scenarios on the exercise, a "toy" case was experimented with. A very small test system was used, consisting only of 3 buses connected triangularly by 3 branches. By adopting the DC model, there are 2 independent variables (2 phases) and 6 dependent variables (3 injections and 3 flows). A scenario of conforming gross errors was simulated and some solutions / estimates were taken from the solutions space.

The following figures display the distribution of residuals, in absolute value, for the dependent variables of each of these estimates. The horizontal line marks a value that is considered as being the maximum possible magnitude of typical noise-induced residuals.

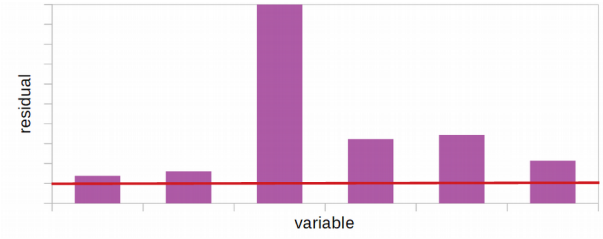


Figure 3.3 – Estimate (1) for the toy case.

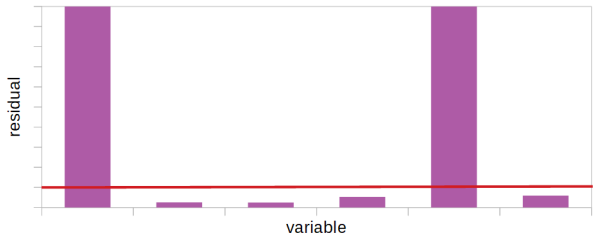


Figure 3.4 - Estimate (2) for the toy case.

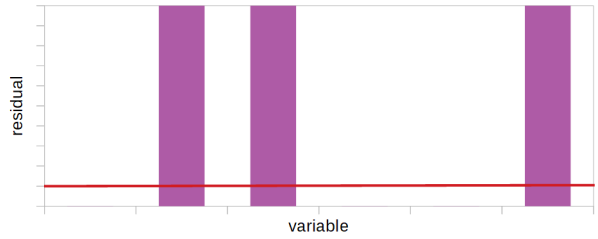


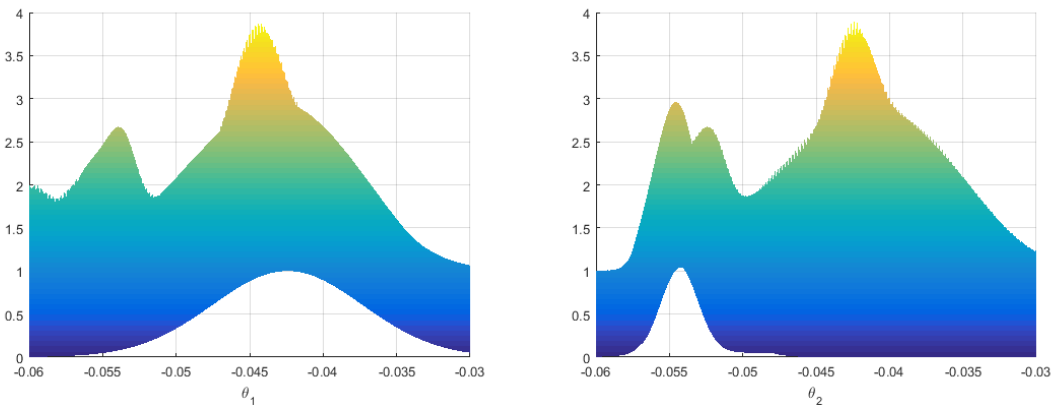
Figure 3.5 - Estimate (3) for the toy case.

Notice that:

1. The first estimate is incoherent in general, as many (all, in this case) its residuals stand above the line (one of which is very large). Therefore, it may be ruled out.
2. The second estimate shows good coherence, as many residuals are seemingly only caused by noise. It is highly incoherent with 2 particular measurements, suggesting that they are outliers.
3. The third estimate also seems fairly coherent, suggesting that 3 particular measurements are outliers. However, we already previously have found an estimate that is coherent with a greater number of measurements.

If we formulate the SE problem using the maximum Correntropy criterion and set its scaling factor equal to the magnitude marked by the horizontal line, these estimates correspond to optima in the problem's solutions space. One of them is the most coherent estimate, corresponding to the global optimum, whereas the other solutions are not as coherent, corresponding to local optima.

Since there are only 2 independent variables in this toy case, it is possible to do a 3-dimensional reconstruction of the solutions space, to allow us to visualise its optima. This experiment is presented in the following figures, each of which is a projection of the maximum Correntropy space from a certain perspective.



Figures 3.6 - Projection of the solutions space on the variable-function planes (X-Z and Y-Z).

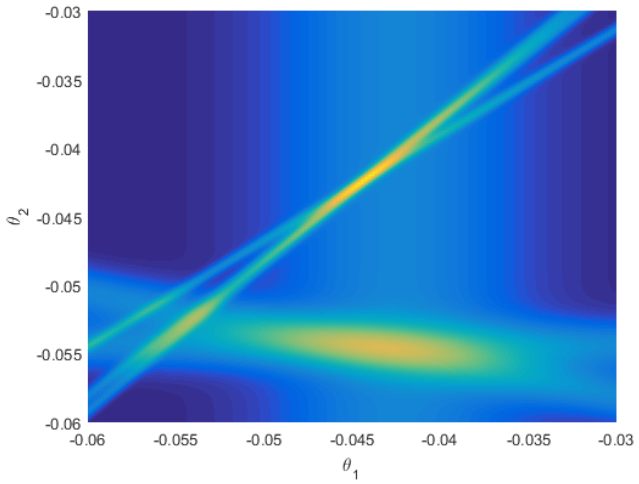


Figure 3.8 - Projection of the solutions space on the variable-variable plane (X-Y).

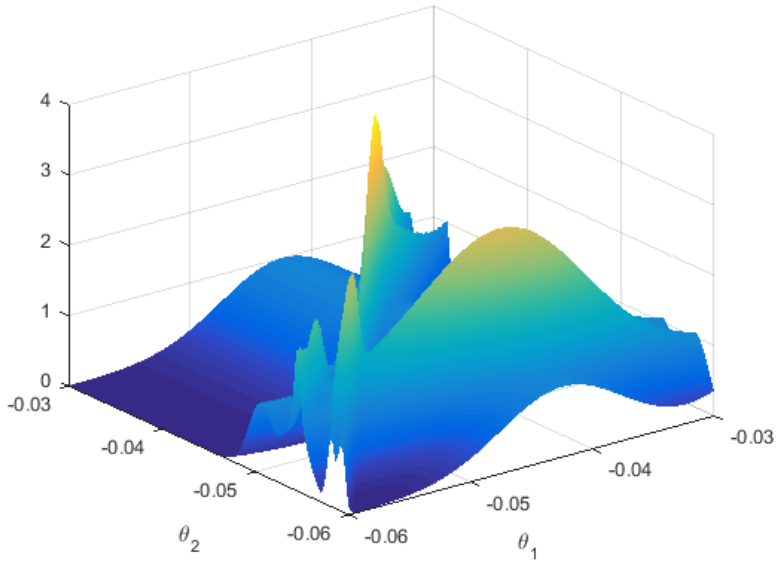


Figure 3.9 - Projection of the solutions space from an overview perspective.

Observing the shape of the landscape, we can assess the difficulty of solving the SE exercise when adopting Correntropy. It is undoubtedly a useful criterion as a measure of localised similarity. However, there might be many optima in the landscape, some of which could have especially misleading basins of attraction.

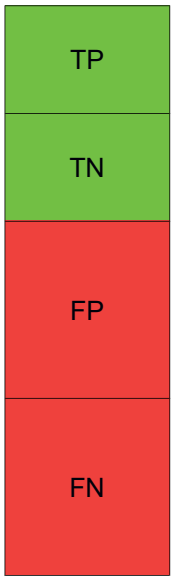
Chapter 4

SE via Least Squares

In the upcoming tests, the test system and simulated scenarios described in section A.1 of the Annexes are to be considered.

4.1 – SE as a diagnosis procedure

The performance of the LS approach to SE will be evaluated based on its sensitivity (detection of anomalies) and specificity (detection of regularities). The decision for accusing a measurement and removing it from future SE processes is based on whether the respective residual of the obtained estimate surpasses a certain threshold. For the upcoming tests, the threshold to be considered is **10%** of the system's base power for both injection and flow measurements.



True positives — Number of correctly identified anomalous measurements.

True negatives — Number of correctly identified regular measurements.

False positives — Number of regular measurements accused as anomalous.

$$FP = n_{regularities} - TN \quad (4.1)$$

False negatives — Number of anomalous measurements considered as regular.

$$FN = n_{anomalies} - TP \quad (4.2)$$

For cases where every anomaly is detected (FN = 0), we could consider each of them as a success. However, in some of those cases where also regularities are falsely accused and, consequently, removed from the process (until the actual anomalies are found, through other

means), the redundancy of the system decreases. It would be preferable if it was possible to keep the number of false positives as low as possible (FP = 0).

For cases where not every anomaly is detected, we will consider each of them as a failure.

4.2 – Results of the LS approach

Since we are addressing the DC model, which is linear, the LS estimate can be found algebraically. If we were addressing the AC model, a numerical method such as Gradient Descent or Gauss-Newton would be needed instead. In that case, it would be initialised at the flat start (all angles at zero).

Table 4.1 - Results of the LS approach.

Case	TP	TN	FP	FN
1, 2, 4, 7, 10, 14, 15, 16, 17, 19, 20, 21, 22, 23, 24, 27, 29, 31	1	32	0	0
5, 11, 12, 13, 18, 25, 26, 28, 30	1	31	1	0
3, 8, 9	1	30	2	0
6	0	31	1	1
35, 38, 44, 47, 49, 53, 55	2	31	0	0
36, 39, 41, 42, 43	2	30	1	0
33, 34, 40, 45, 54	2	29	2	0
32	1	31	0	1
37, 46, 48, 50, 51, 52	1	30	1	1

The following figures show the residuals of the estimates for some test cases. Notice how, sometimes, gross errors can have a significant widespread influence on the residuals profile.

Please refer to section A.2 of the Annexes for the mapping of the residuals (which residual corresponds to which measurement).

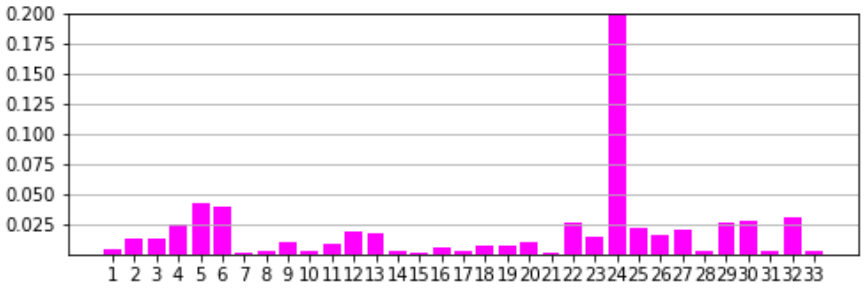


Figure 4.1 - Residuals of the LS estimate for case nº 22, with 1 TP (correct detection).

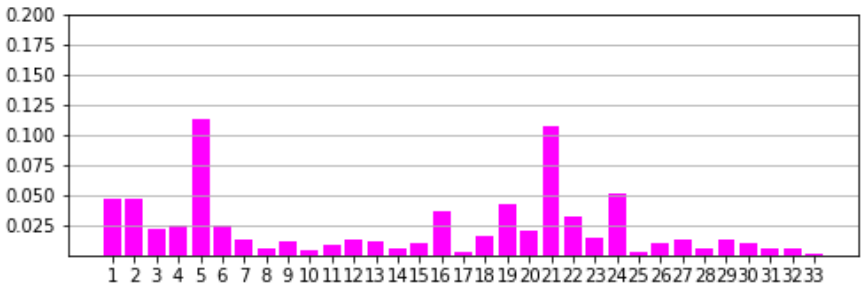


Figure 4.2 - Residuals of the LS estimate for case nº 5, with 1 TP and 1 FP (false accusation).

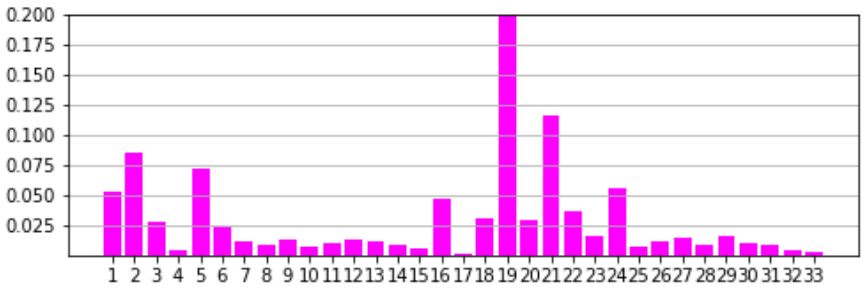


Figure 4.3 – Residuals of the LS estimate for case nº 48, with 1 TP, 1 FP and 1 FN (missed detection).

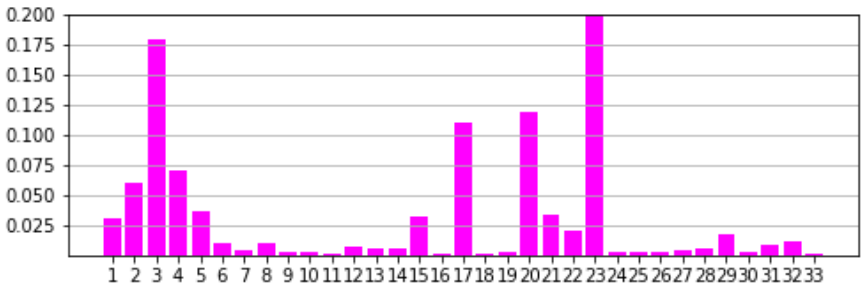


Figure 4.4 - Residuals of the LS estimate for case nº 34, with 2 TP and 2 FP.

Comments

- The traditional strategy works well for a great majority of the cases with a single gross error. In 3 out of the 12 cases that target injection measurements, 2 FP occur. This hints that gross errors are more problematic when they contaminate these kinds of measurements, having a greater influence on the estimate and on the residuals that it contains.
- On case nº 6: If the residual threshold was 5% (instead of 10%), it would actually result in a success with 0 FP. The gross error is not "gross enough" to spawn residuals of greater magnitude, making it more difficult to detect.
- It does not work as well for the cases with two gross errors, where only 7 out of those 24 cases cause 0 FN and 0 FP. In other 7 cases, not all anomalies are detected. 5 of those cases are part of the group of (6) cases which simulate conforming errors.

Chapter 5

SE via maximum Correntropy

5.1 – Attempt with numerical method

In order to try to give convincing and successful results where LS has failed, the Correntropy criterion will now be employed. The same test system and scenarios of chapter 4 will be considered. Since the problem can no longer be solved algebraically for Correntropy, a standard Gradient Descent method will be employed instead. An intelligent initialisation is to be used so as hopefully to place the gradient agent in the best convergence zone (that contains the intended solution / best estimate). The strategy is as follows.

1. Find the LS estimate.
2. Find the Correntropy estimate by initialising GD at the LS estimate.

The standard GD method has only one parameter. However, it was given a second parameter to slightly reduce its step proportionally for every iteration that the solution does not improve (addressed below as "cooling factor"), making the method increasingly more precise. The setting to be used is as follows.

initial step	0,14
cooling factor	0,1

The total number of iterations to be performed by GD is 2000.

The setting for the Correntropy function is as follows.

α	0,1

Table 5.1 - Results for Correntropy approach with numerical method.

Case	TP	TN	FP	FN
1 to 31	1	32	0	0
32 to 49, 53 to 55	2	31	0	0
50, 51, 52	0	30	1	2

The following figures show the residuals of the estimates for some successful cases. It is now possible to more clearly identify the anomalies (for the successful cases), which are given away by the only large residuals. All the other smaller residuals are due only to noise.

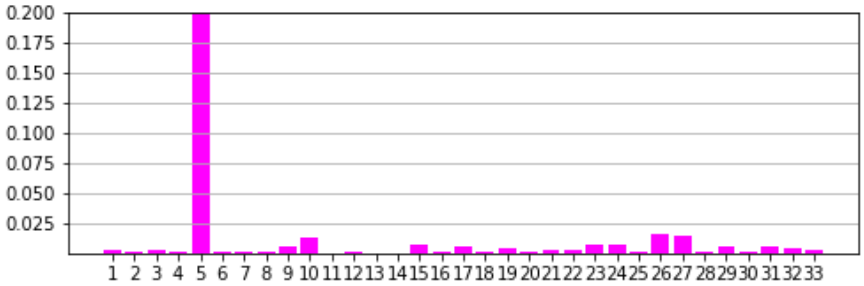


Figure 5.1 - Residuals of the Correntropy estimate for case nº5, with 1 TP.

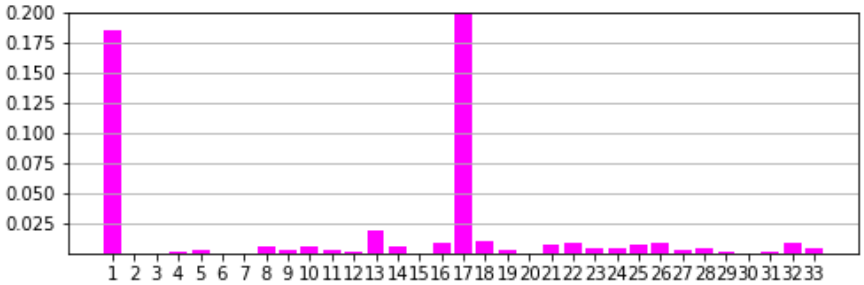


Figure 5.2 - Residuals of the Correntropy estimate for case nº32, with 2 TP.

Comments

- This strategy is convincingly successful for 51 out of all the cases. The estimates that are obtained in these cases generate large residuals which always match the anomalous measurements and no FP occur.
- It fails for cases nº 50, 51 and 52. The total cost of each estimate is lesser than the total cost at the intended solution. This indicates that the intended solution is locally rather than globally optimal and, as such, a smaller value for α is needed in order to further zone out the gross errors.

Case	Cost at estimate	Cost at intended solution
50	1,98	2,00
51	1,87	2,00
52	1,94	2,00

Another attempt was made at solving cases nº 50, 51 and 52, with the same method but this time with the following Correntropy setting.

α	0,06

All results remain the same as shown in table 5.1.

Comments

- This approach still did not solve the problematic cases in study. Nonetheless, the total cost of each estimate is now greater than the total cost at the intended solution, which indicates that the sensibility is adequate.

Case	Cost at estimate	Cost at intended solution
50	3,15	2,00
51	3,07	2,00
52	3,11	2,00

It is thought that the locations of the optima in the landscape change according to the settings of the Correntropy parameters. An hypothesis is formulated that, if parameter α is set at a greater value and made to decrease gradually rather than being fixed and instantly establishing a tight sensibility, it would be possible for the gradient agent to follow the global optimum's relocation.

The next stage of the current strategy consists of gradually changing landscape so as hopefully to guide the method towards the best estimate. After finding the LS estimate, the method will be run a few successive times. For each run, a new setting of parameters is to be used and the method will be initialised at the solution which was obtained for the previous run.

The Correntropy settings to be used for each run are as follows.

Run		
1	α	0,2
2		0,16
3		0,12
4		0,1
5		0,08
6		0,06

All results remain the same as shown in table 5.1.

The following figures show the errors for each of the Correntropy estimates for cases nº 50, 51 and 52 which were unsuccessful.

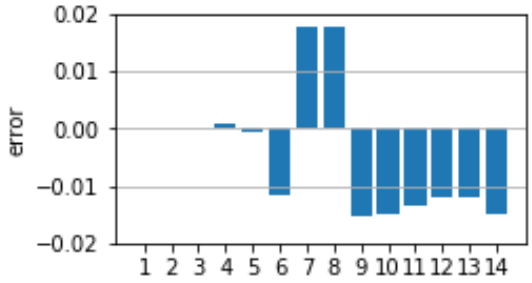


Figure 5.3 - Correntropy estimate for case nº 50 (unsuccessful).

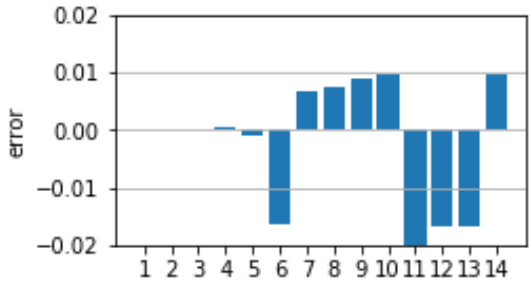


Figure 5.4 - Correntropy estimate for case nº 51 (unsuccessful).

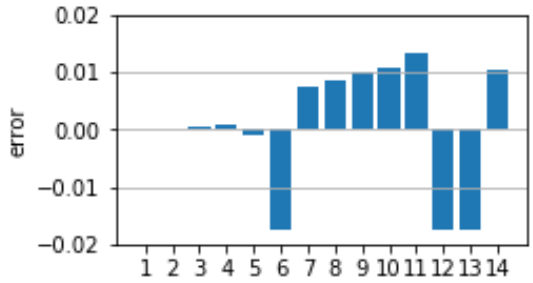


Figure 5.5 - Correntropy estimate for case nº 52 (unsuccessful).

The following figures show the residuals of each of the estimates. The largest residual corresponds to the regular measurement which the anomalies / gross errors work against, being falsely accused. Oppositely, the residuals of the two anomalous measurements which should be accused blend in with the data. Notice that there are some residuals which, despite not being large enough to be flagged, are fairly larger than the noise-induced residuals.

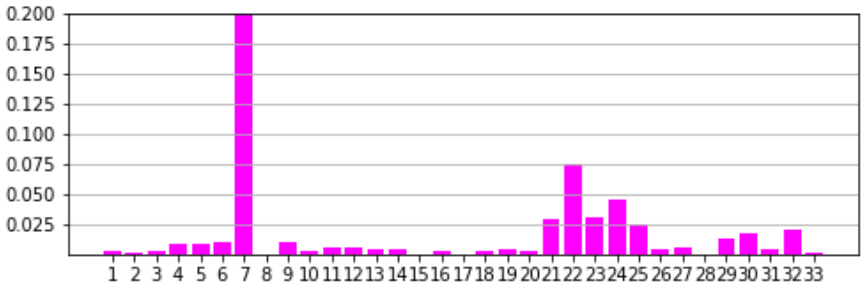


Figure 5.6 - Residuals of the Correntropy estimate for case nº 50 (unsuccessful).

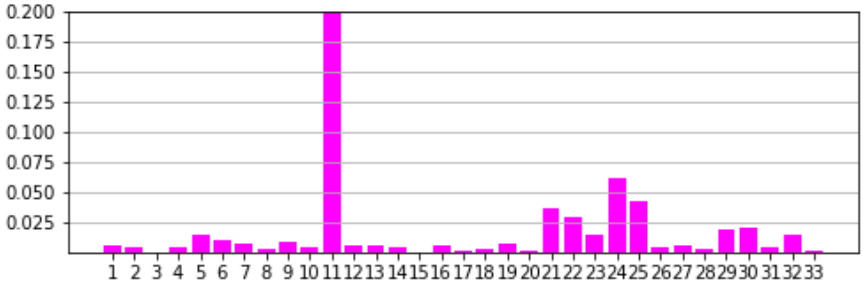


Figure 5.7 - Residuals of the Correntropy estimate for case nº 51 (unsuccessful).

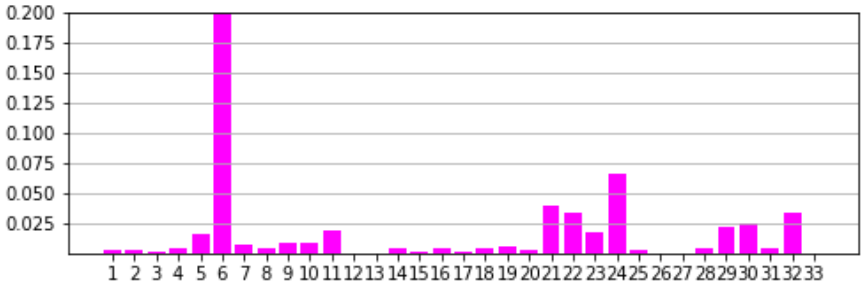


Figure 5.8 - Residuals of the Correntropy estimate for case nº 52 (unsuccessful).

- Cases nº 50, 51 and 52 still remain to be solved. Despite a tighter sensibility which sets the true solution as the global optimum, the agent gets trapped at a local optimum. Interestingly, its position barely shifts between all the successive runs. This suggests that, rather than an optimum which gradually relocates, its position is fixed and only its height (cost) changes so that, for a certain value for parameter α, it stops being the global optimum.
- An additional experiment was made for each of the unsolved cases: the method was instead initialised at the intended solution for all runs. It was found that, in the early runs, the agent always "dropped down" to the same optimum but, suddenly, no longer moved away from its starting point. This happens for all the aforementioned cases for a value of α somewhere between 0,09 and 0,08.
- It is important to remember that the aforementioned cases are part of the roster of gross error conformance cases and, as such, are exceptionally difficult to tackle. For all three cases, there were 0 TP (no anomalies detected) and 1 FP (one regularity accused).

5.2 – Hybrid method to solve Correntropy SE

The failure of the previous strategy in finding the best estimates for the aforementioned problematic cases is due to the inadequate initialisation of the gradient agent, although it does work for all the other cases.

An alternate possible strategy could be to randomise the starting position for the agent, either sampling it from the whole domain of the problem or, more intelligently, from the vicinity of the LS estimate. The mechanism that is used could be a slight perturbation / mutation of the reference point. It could otherwise spawn the agent inside a delimited area, centred on said point. Yet, it is very likely that many agents have to be spawned until one of them gets properly placed in the convergence zone that contains the best estimate, since such strategy actually is not sensitive to the shape of the landscape.

Instead, we could firstly develop a strategy for looking more widely for multiple convergence zones in the landscape and then placing the agent in whichever is the most “interesting”. We have already discussed, on the matter of meta-heuristics in section 2.3.2, that these kinds of methods can be quite effective in wide explorations of the landscape and in quickly finding zones of interest. If a particle swarm manages to find the convergence zone that contains the global optimum, a gradient agent could then take it from there, converging much faster towards the optimum.

The third and final strategy which we will employ to tackle cases nº 50, 51 and 52 keeps the Correntropy criterion but will Evolutionary Particle Swarm Optimisation alongside with GD. It is as follows.

1. Find the LS estimate.
2. Launch EPSO particles / agents, taking the LS estimate in account.

3. After a few iterations, initialise a GD agent at the best solution which was found by the swarm (the shared reference).

The exact rule for spawning each particle is as follows:

1. Spawn the particle at random in the domain. This was chosen to be the interval from $-\pi/2$ to $\pi/2$ for every angle.
2. Pull the particle towards the "anchor" (LS estimate) according to a certain weight. The idea is to have each the particle spawn close to but nor right on top of the anchor. The weight defines how proportionally close to the anchor the particle is placed, relative to its initial spawn. A weight of 0,95 was used.

The Correntropy setting is kept the same from the previous strategy.

α	0,06

EPSO has many parameters to be tuned. Some affect the method's performance more than others. A fairly standard setting is to be used and is as follows.

number of particles	14
maximum initial speed	0,09
habit initial weight	0,8
memory initial weight	0,2
cooperation initial weight	0,2
habit maximum weight	1,0
memory maximum weight	2,0
cooperation maximum weight	2,0
habit learning factor	0,1
memory learning factor	0,15
cooperation learning factor	0,15
global reference mutation rate	random
global reference mutation factor	0,008
cooperation rate	0,8

While the initial speed for each particle is limited to a threshold, it is left unrestrained afterwards.

The mutation rate (probability for the mutation of a variable) for the global reference is reset randomly for every iteration.

The total number of fitness evaluations to be performed by EPSO is 14000 (500 iterations for 14 particles).

From now on, the performance of this methodology and of the upcoming ones will be evaluated by the success rate of the method in finding the intended solution / global optimum, meaning when 0 FP and 0 FN occur. Some observations will also be made regarding the rate at which the method gets trapped at exceptionally deceptive local optima. For cases nº 50, 51 and 52, these optima correspond to the respective Correntropy estimates which were obtained with the previous strategy. For quick reference and as they are the largest local optima, they will be addressed as "nearly-global" optima.

Table 5.2 - Results of 20 runs for Correntropy approach with hybrid EPSO GD method.

Case	Success rate	Rate of convergence to nearly-global optimum
50	45%	40%
51	35%	45%
52	35%	45%

The following figures show the residuals of the estimates. For all cases, the anomalies have now all been correctly identified with no false accusations on regularities.

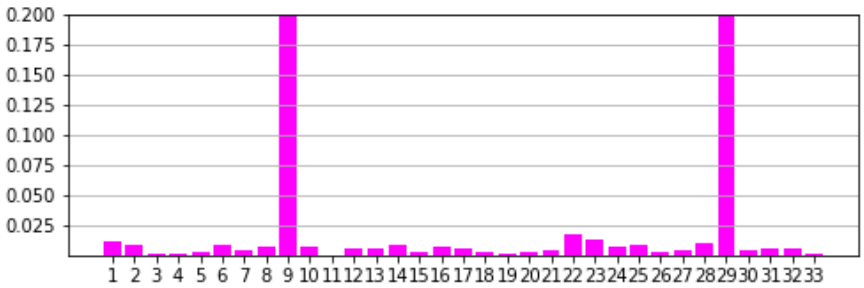


Figure 5.9 - Residuals of the Correntropy estimate, for case nº 50.

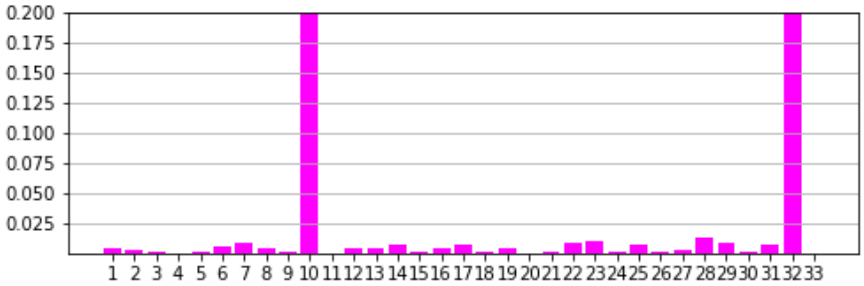


Figure 5.10 - Residuals of the Correntropy estimate, for case nº 51.

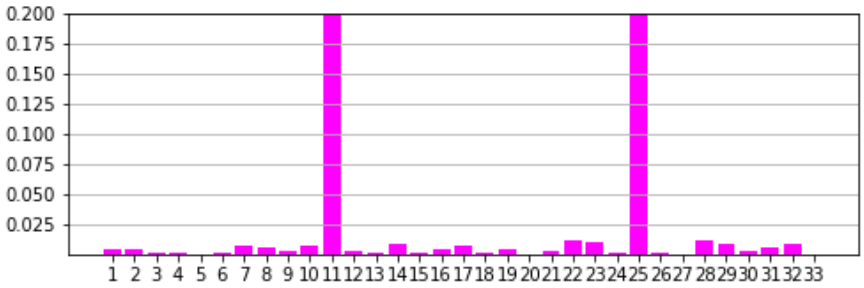


Figure 5.11 - Residuals of the Correntropy estimate, for case nº 52.

The following figures show the progress of successful runs.

Side note: the implementation of the method was done for function maximisation, hence the polarity of the fitness function (negative cost) in the figures.

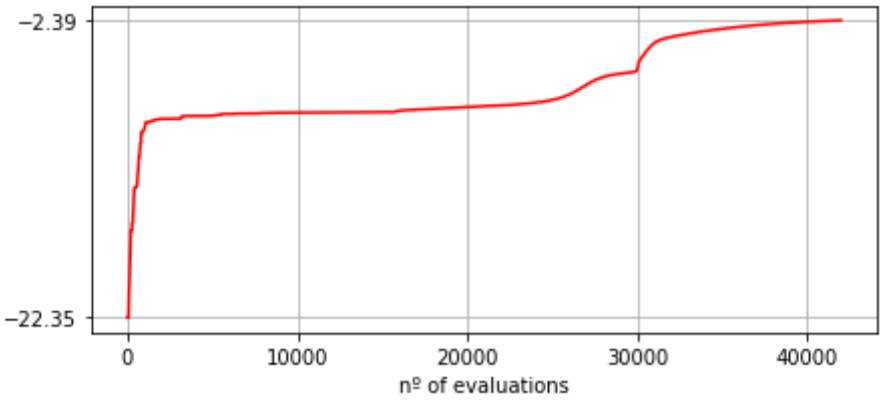


Figure 5.12 - Progress of the hybrid method, for case nº 50.

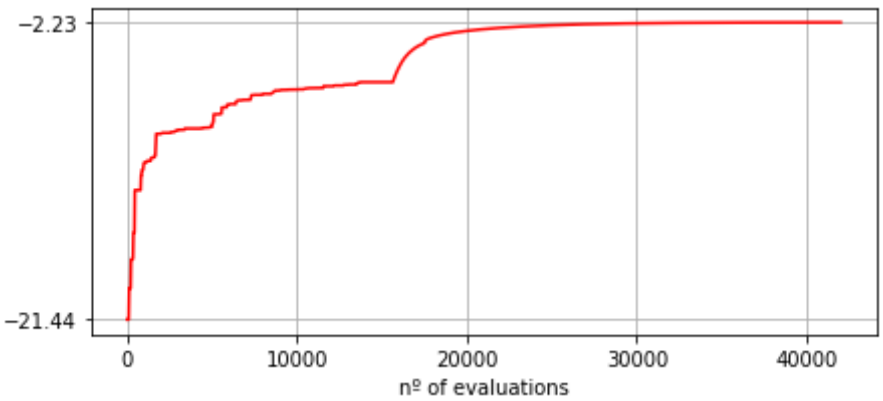


Figure 5.13 - Progress of the hybrid method, for case nº 51.

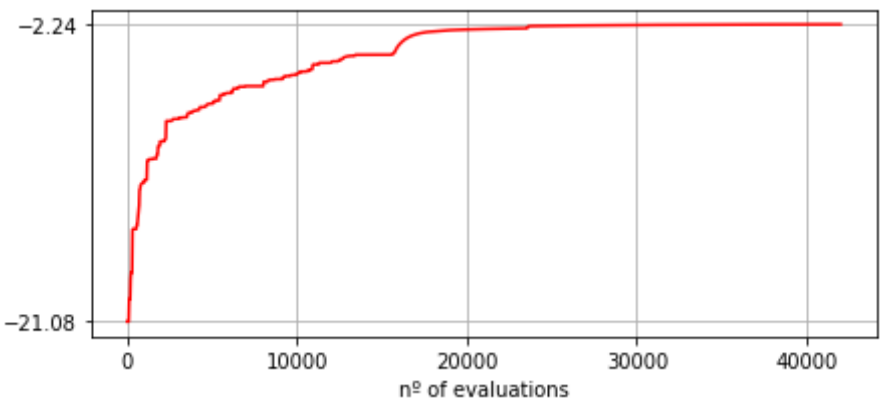


Figure 5.14 - Progress of the hybrid method, for case nº 52.

Comments

- We could say that this strategy is fairly successful for all cases. Even if the method fails to reach the global optimum for some runs, this is not a problem, as long as the method is allowed to do enough runs to arrive at the intended solution at least once. Since this solution is globally optimal (lowest total cost), it will be the estimate of choice in the end of the process.

Even with fairly satisfactory results, it may be possible to improve the method's success rate by adjusting its settings and employing additional mechanisms. The upcoming tests will be directed at this goal.

Only case nº 50 will be addressed for now. When a conclusion has been reached regarding the best combination of settings and mechanisms, then cases nº 51 and 52 will be addressed as well.

5.2.1 – Swarm size and cooperation probability

For the previous tests, the number of particles that were spawned was equal to the number of variables plus one (for good measure), as this is a fairly standard setting. Intuition would tell us that a greater number of particles is always better since it increases the chances that one of them steps into an interesting zone of the landscape. By this logic, we could spawn hundreds or even more. Well… that approach would defeat the purpose of employing a particle swarm method in the first place, since each iteration becomes heavier (and we might as well just do massive random sampling and try to win the lottery). The benefit of having more particles should be assessed considering the same number of allowed fitness evaluations to be performed – how better is a bigger swarm with the same available resources.

The cooperation rate / probability parameter plays an important role in the behaviour of the swarm. For the previous tests, it was set at a standard value of 80%. This translates to a high degree of communication among the swarm which is expected to make all particles

share their efforts on the same zone. On the other hand, it is expected that a low cooperation probability disaggregates the swarm, leaving each particle to its own personal exploitations. This seems to be beneficial since a particle is allowed to leave the flock and come across yet an undiscovered zones. However, since the swarm is disaggregated, an interesting zone may be left poorly explored as other members do not come helping.

Some tests were made for different configurations of the aforementioned parameters (while keeping all the others).

Table 5.3 - Results of 20 runs for Correntropy approach for case nº 50 with different swarm settings.

Configuration			Success rate	Rate of convergence to nearly-global optimum
1	number of particles	70	55%	35%
	cooperation rate	80%		
2	number of particles	14	40%	30%
	cooperation rate	20%		
3	number of particles	70	35%	55%
	cooperation rate	20%		

Comments

- A bigger swarm seems to help slightly with the convergence, even considering a proportionally smaller number of EPSO iterations (only 100).
- On the other hand, a low cooperation rate seems to badly affect the performance of the method when employed either by itself or even with a greater swarm size. Possibly, one or more particles would come across the convergence zone of the global optimum but, because of lack of focused effort, the zone is left poorly explored.

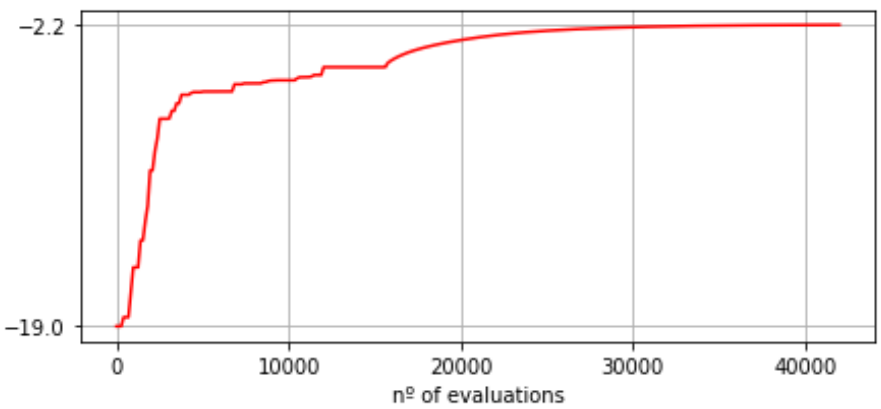


Figure 5.15 - Progress of the hybrid method with 5 times more particles.

5.2.2 – Penalised re-initialisation

In the previous tests for the problematic cases that were done with the hybrid EPSO GD method, the rates of convergence towards the nearly-global optima have been interestingly high. This makes some sense since the conformance of gross errors generates a locally optimal solution with nearly as low a total cost / as high Correntropy as the best estimate, making it exceptionally deceptive.

If we were to place a "probe" at this local optimum and gradually push it towards the global optimum in a straight line, we would notice that the density distribution of the residuals, built by applying the Correntropy kernel to the residuals, changes very little at a larger scale but undergoes interesting changes if we take a closer look at specific regions.

This experiment may be visualised in the following figure, for case nº 50.

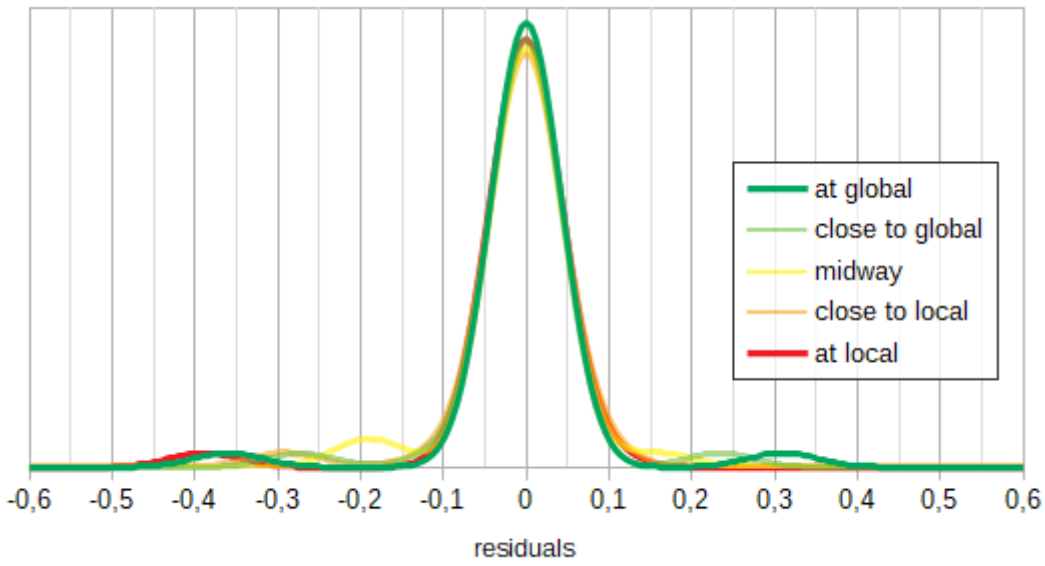


Figure 5.16 - Residuals density distributions at different points between the global and nearly-global optima inclusive for case nº 50.

The two next figures are a zoom on different regions of the previous figure.

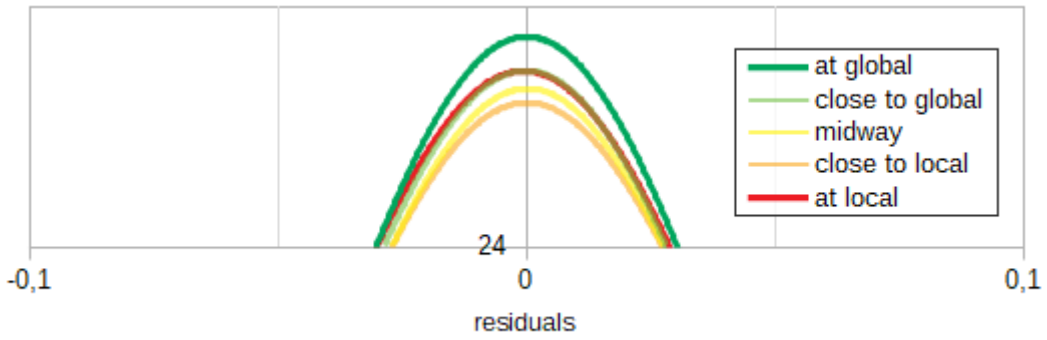


Figure 5.17 - Zoom of figure 5.16 on the highest peaks.

Notice how the global optimum, which is the intended solution for the SE problem, is the point of maximum Correntropy (highest degree of correlation between the estimate and the measurements). However, the Correntropy of the local optimum is very high as well because of gross error conformance. The point at which the convergence zones split for case nº 50 is somewhere between the local optimum and 50% of the way through towards the global optimum.

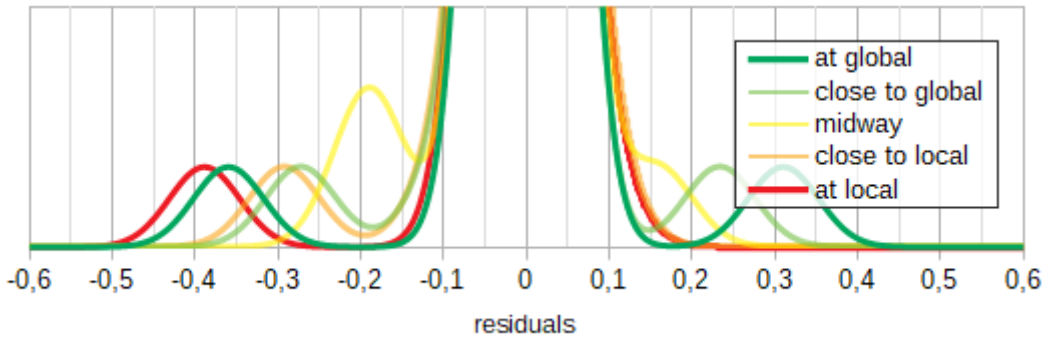


Figure 5.18 - Zoom of figure 5.16 on the lowest peaks.

At the local optimum, such residual is the one which contradicts the conforming gross errors. At the global optimum, the two residuals that stand out correspond to said errors. As the probe moves gradually from one optimum to the other, the regularity is “pulled” into the region of aggregated kernels and the anomalies are “expelled”, each to its side as, in case nº 50, the two errors have opposite polarities.

In an attempt to turn into a success each run where the method converges to the nearly-global optimum, the current strategy will be extended with a re-initialisation as such:

1. After a first run, the hybrid method will be relaunched from the LS estimate.
2. For this next run (and subsequent others), a penalty will be applied to the fitness of particles that get too close to the solution which was found in the first run.

With this approach, we hope that the basin of attraction of the deceptive optimum becomes an “off-limit” zone, preventing the swarm from settling on the same optimum again.

The penalty will be applied if the distance between a particle and the previously found optimum is smaller than a threshold.

The Euclidean distance that separates the two contesting optima for case nº 50 is roughly 0,044 (radians). While this is privileged information (since this is a simulation and we would not have access to it in a real situation), we could take it into account on parameter tuning it only for the sake of proving the mechanism's potential usefulness and then eventually develop a strategy for tuning the parameter in the absence of such information.

Some tests were made with different values for the penalty threshold distance. Also, the cooperation rate will be experimented with again so as to check how well it complements the new mechanism. The swarm size to be set is 70 particles.

Table 5.4 - Results of 20 runs for Correntropy approach for case nº 50 with use of the penalty mechanism.

Configuration			Success rate	Rate of convergence to nearly-global optimum
1	cooperation rate	80%	40%	40%
	penalty distance	0,03		
2	cooperation rate	20%	30%	25%
	penalty distance	0,03		
3	cooperation rate	80%	50%	30%
	penalty distance	0,04		
4	cooperation rate	20%	55%	15%
	penalty distance	0,04		

Comments

- While a penalised re-initialisation could be helpful for pushing the swarm away from the deceptive optimum at which it landed first, it seems not to have as significant of a benefit as was expected at first. In some cases, the swarm is still not repelled away from the deceptive optimum. In other cases, the swarm rather seems to settle more often on other local optima.
- The configuration with longer threshold distance and lower cooperation rate is the better performing. This may be justified in that particles are not "forced" as much to explore prematurely near the nearly-global optimum, which is now "off-limits".
- In a sense, it is somewhat plausible that preventing the swarm from nearing the nearly-global affects the search in a slightly negative way. Maybe, once found (considering no penalty), said optimum would usefully bring the search towards that region, where the global optimum is also located.

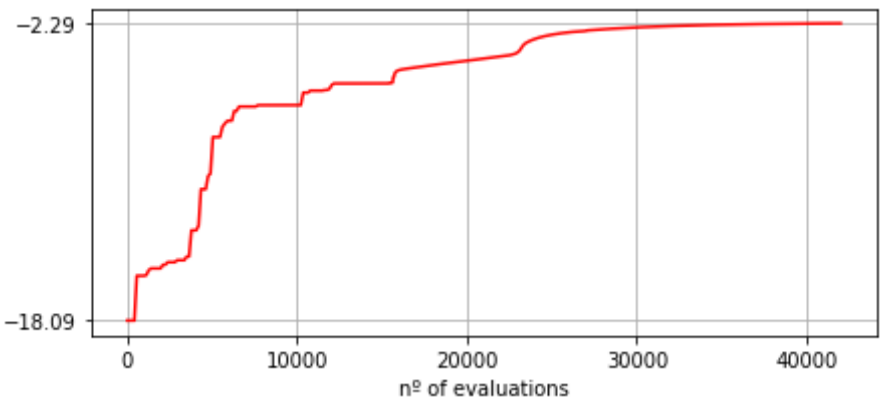


Figure 5.19 - Progress of the hybrid method with penalty.

5.2.3 – Multi-gradient

The expected behaviour of a particle swarm method like EPSO is, through an exchange of information, to bring the particles increasingly closer over time, at a certain rate (cooperation rate), so as to more finely exploit the most interesting zone yet found. As such, the personal references of the particles are more distant from each other at an early instant but, later on, get placed around the best (shared) reference of the whole swarm. A trade-off must be made. Do we:

- give the swarm plenty of time and better communication for more quickly aggregating the swarm around a zone while potentially leaving other zones in the landscape undiscovered?
- prevent the swarm from communicating and aggregating too much around a single zone but risk leaving all the other discovered zones poorly explored?

The problem with the second hypothesis is that a lone particle, by itself, is inefficient at converging to the optimum of a convergence zone because of the rule which dictates its movement – big leaps and oscillation. However, if a reference inside such zone gets turned into a gradient agent, the agent is able to more quickly find the optimum. This actually is the foundation for employing GD after EPSO, only that we have been doing it for the global reference only.

What if we let the swarm spread to many zones and then turn every personal reference into a gradient agent? Whatever the configuration is for the particle swarm, some references might still end up very close if two or more particles happen to come across the same interesting zone. In this situation, spawning a gradient agent for each of these references would be an unnecessary waste of computational effort, since all the agents would meet at the same location (the optimum).

In the upcoming tests, we will assess the benefit of spawning multiple gradient agents along the landscape, after the swarm season, at the cost of a greater consumption of resources. The spawning rule is as follows (after the swarm season).

1. Sort the personal references of the swarm by descending order of their fitness (lower total cost first).

2. Spawn the first gradient agent from the first best reference.
3. For every other reference to follow in order, check its Euclidean distance to every other already existing gradient agent.
   a) If the distance is greater than a threshold, spawn a new gradient agent at the reference;
   b) If not, no agent is spawned.

Some experiments were made with the new aforementioned mechanism. Different configurations for EPSO were used, regarding the cooperation rate. The swarm size was set at 70 particles. Also, the test with configuration nº 4 (presented below) was made using the penalty mechanism.

Table 5.5 - Results of 20 runs for Correntropy approach for case nº 50 with multi-gradient extension.

Configuration			Success rate	Rate of convergence to nearly-global optimum
1	cooperation rate	80%	50%	20%
	agent spawn distance	0,02		
2	cooperation rate	20%	60%	30%
	agent spawn distance	0,02		
3	cooperation rate	20%	75%	20%
	agent spawn distance	0,01		
4	cooperation rate	20%	80%	15%
	agent spawn distance	0,01		
	penalty distance	0,03		

Table 5.6 - Statistics of the spawn of extra agents for Correntropy approach for case nº 50.

Configuration nº	Extra agents spawn rate for ... (number of) agents				Average number of extra agents
	1	2 to 4	5 to 9	10 or more	
1	20%	5%	0%	0%	0,3
2	55%	25%	5%	0%	1,7
3	5%	35%	50%	10%	5,4
4	5%	20%	25%	50%	9,7

Comments

- For the test where a high cooperation rate is set, the number of extra agents that were spawned is generally very small. This matches our earlier discussion of the influence of said rate on the speed of aggregation of the swarm. Because the particles quickly came together, their personal references ended up too close for the spawning of many gradient agents to be justifiable.
- On the other hand, for the tests with a low cooperation rate, more extra agents are spawned and even more so for an ever smaller spawn threshold distance (which is expected). The increase in the success rate, regarding previous tests without the multi-gradient mechanism, hints at what has been discussed earlier in that at least one particle finds the convergence zone but its finding would otherwise get "muffled" by a more interesting estimate found elsewhere.
- However, we should still take into consideration that greater computational effort was needed as the cost for higher success rates.
- Even alongside the multi-gradient, the penalty mechanism seems not to boost the success rate significantly. Actually, it has a higher spawn rate of extra agents, which consumes more resources – roughly double that of not applying penalty. Whereas some of the swarm's references would join close together around the local optimum when there is no penalty (potentially reducing the number of extra spawns), it is possible that they instead scatter around the optimum's "zone of avoidance", becoming more distant from each other, increasing the number of extra spawns. However, since this is the zone of the local optimum, the extra effort is in vain. Thus, the use of this mechanism alongside the multi-gradient may not pay off.

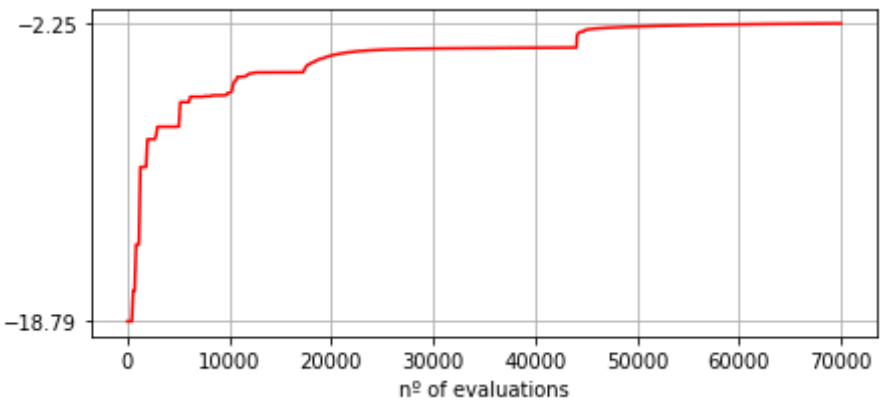


Figure 5.20 - Progress of the hybrid method with multi-gradient mechanism.

Additional tests will be made, now addressing cases nº 51 and 52 employing multi-gradient configuration nº 3 described in table 6.5.

Table 5.7 - Results of 20 runs for Correntropy approach for cases nº 51 and 52 with multi-gradient extension.

Case	Success rate	Rate of convergence to nearly-global optimum
51	65%	35%
52	60%	30%

Table 5.8 - Statistics of the spawn of extra agents for Correntropy approach for cases nº 51 and 52.

Case nº	Extra agents spawn rate for ... (number of) agents				Average number of extra agents
	1	2 to 4	5 to 9	10 or more	
51	20%	5%	0%	0%	6,9
52	55%	25%	5%	0%	7,8

Comments

- Cases nº 51 and 52 seem a little more difficult to tackle than case nº 50, even with the multi-gradient mechanism. Although more agents are spawned for both cases than for case nº 50, the success rate is slightly lower.

5.3 – Generalised Correntropy strategy for SE

So far, we have been using the Gaussian distribution as the Correntropy kernel function. It is fairly satisfactory in giving us the answers we want, provided we find the global optimum of the optimisation problem. This simply was our choice. We could use any other kernel that we feel is adequate for the exercise.

In a generalised version of the Gaussian distribution, instead of squaring the fraction of the variable by the scaling factor, its exponent becomes another function parameter. We can import this parameterisation to Correntropy and the function becomes as follows:

$$cost(r) = 1 - exp(-|r/\alpha|^{\beta}) \tag{5.1}$$

where:

- β shaping factor.

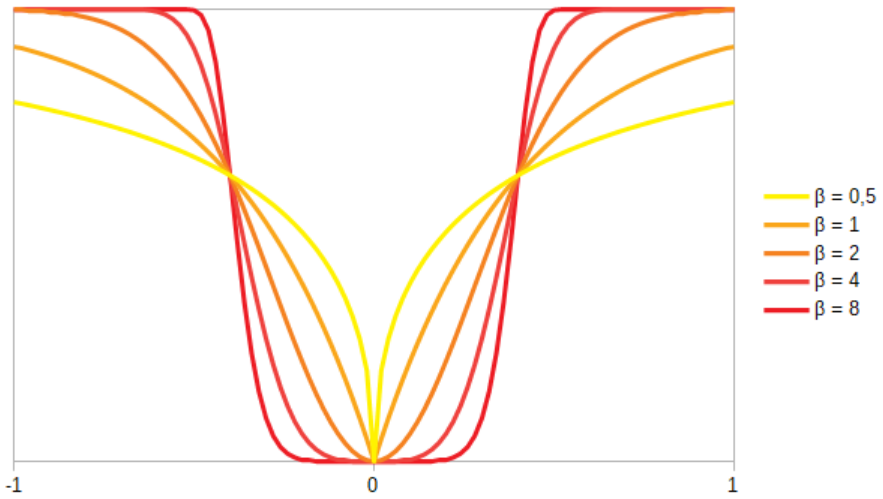


Figure 5.21 - Generalised Correntropy.

β is related to the steepness of the kernel. It defines the contrast in sensibility between residuals which stand at the inner region of the kernel.

In Classical Correntropy, β is set as a square (2). For a larger β, the inner region of the kernel becomes flattened, since, in such region, this parameter is the exponent for powers of numbers which are much smaller than the unit. In theory, an infinitely large β would give the kernel the shape of a random uniform distribution or, in another perspective, a low-pass filter.

$$\beta \to \infty \Rightarrow cost(r) \simeq \begin{cases} 0, & |r| < \alpha \\ 1 - exp(-1), & |r| = \alpha \\ 1, & |r| > \alpha \end{cases} \tag{5.2}$$

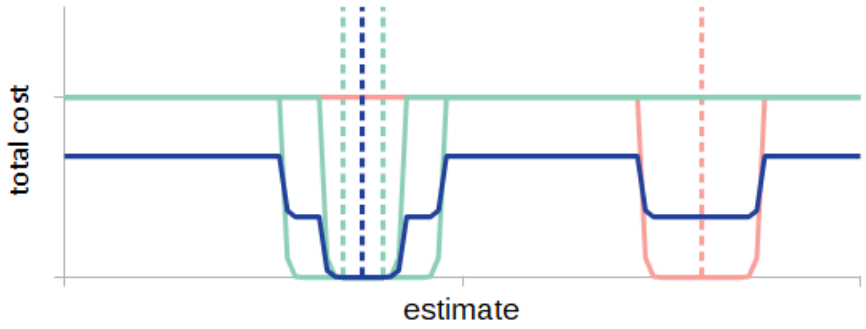


Figure 5.22 – Generalised Correntropy estimate with large ß. The landscape becomes divided into various flat sections at different heights.

In this light, ß could change the behaviour of the method that is employed to tackle the problem, as there could be certain points that, with an increase in ß, could be placed at a much lower height than with a classical Gaussian kernel. This may be particularly useful in a hypothetical situation where an agent comes across both the precise position of a local optimum and a point which is barely inside the global optimum's convergence zone / at its border. Depending on the Correntropy settings, the best candidate solution of the two could change. A larger ß gives convergence zones a wider "event horizon" and may aid in capturing agents which happen to pass nearby, potentially leading to a shift in the focus of the search.

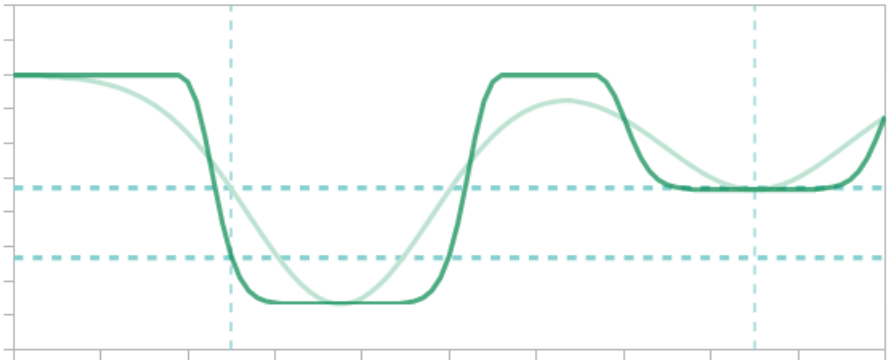


Figure 5.23 – Capture strategy (large ß). Convergence zones are better delimited.

Taking the same toy example that was experimented with in chapter 3 for the purpose of visualisation (3-bus system, conforming gross errors simulation), a reconstruction of the same solutions space was made, still adopting the maximum Correntropy criterion with the same α but now in its generalised version with a large ß. This is displayed in the following figures. Notice how the convergence zones are now more clearly delimited.

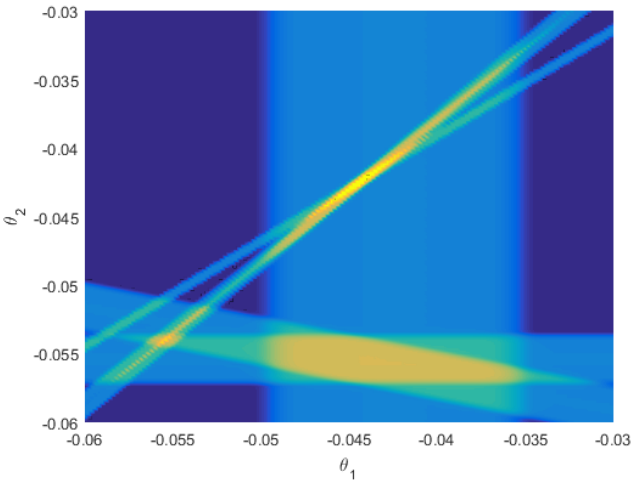


Figure 5.24 - Projection of the solutions space on the variable-variable plane (X-Y).

However, ß must also be tuned with care so as not to convert a local optimum into the global optimum (taking that role away from the true / intended solution). This may happen since, for a larger ß:

- residuals which are smaller than α shrink in cost;
- residuals which are slightly larger than α grow in cost;
- residuals which are much larger than α already have their costs peaked.

If a certain optimum whose residual profile / spectre contains mostly small residuals that sit slightly below α, a great increase in the shaping factor means that the costs of all these residuals get greatly reduced and become insignificant. This may significantly reduce the total cost of an (local) optimum.

The final upcoming tests are aimed at assessing the benefit of using the Generalised Correntropy function with a different value for ß. As has been just discussed, it is possible that flattening the convergence zones on the landscape by increasing β helps to better delimit those zones. Particles which then enter those zones will find points of lower cost than before, which may be a deciding factor for which references are kept at certain stages of the search. Nonetheless, the effectiveness of a larger ß is tightly intertwined with α. It is possible that, for an increase in ß, the deceptive optimum which was otherwise local becomes global if more of its residuals stand below α than for the otherwise global optimum. A few runs of the EPSO GD hybrid method were made for each case, with the following settings for the Correntropy cost function, in order to test this hypothesis.

α	0,06
ß	8

In actuality, this is exactly what happens for cases nº 50, 51 and 52. Here follow the costs for the most optimal estimates which were found.

Case	Cost at estimate	Cost at intended solution
50	1,24	2,00
51	1,18	2,00
52	1,27	2,00

The following figures shows the residuals of the estimate for case nº 50 and the costs of the residuals, both with Classical Correntropy and with Generalised Correntropy. The obtained estimate for this setting of ß is slightly different than the deceptive optimum which had been found in previous tests. Notice that all residuals with the exception of one now stand below α (0,06). With the larger ß, these residuals are not penalised in cost as much and, thus, this estimate gets a lower cost than the deceptive optimum for the Classical Correntropy setting. For the same reason, it gets a lower cost than the intended solution since it only has a single large residual instead of two, whereas the cost of the intended solution is roughly equal to the number of anomalous measurements (= 2).

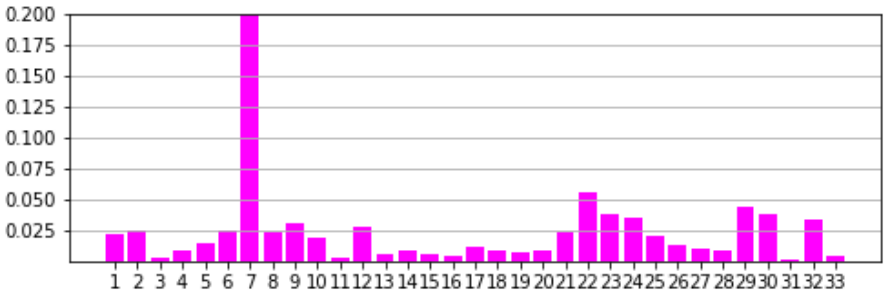


Figure 5.25 - Residuals of the Generalised Correntropy (ß = 8) global optimum (not the intended solution).

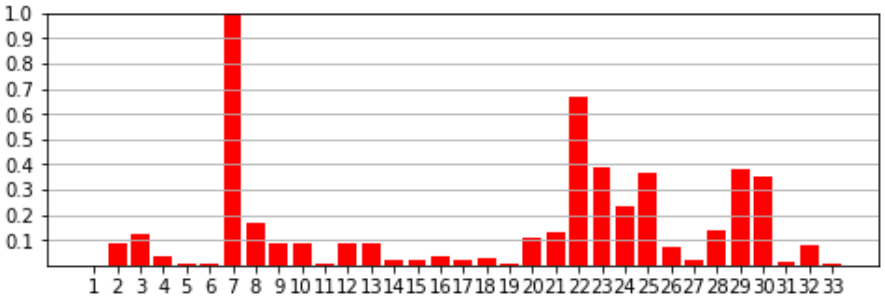


Figure 5.26 - Costs of the residuals shown in figure 7.4, for Classical Correntropy.

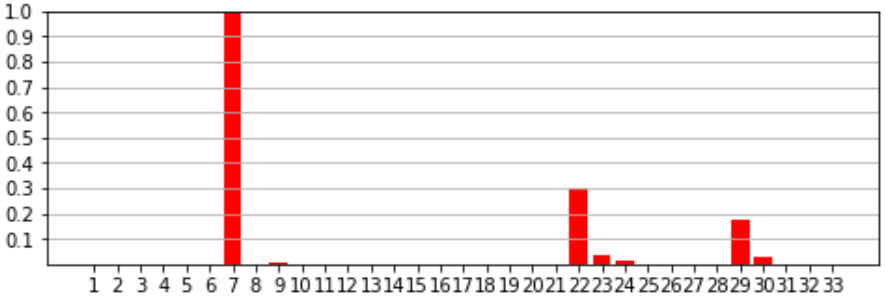


Figure 5.27 - Costs of the residuals shown in figure 7.4, for Generalised Correntropy.

As such, if we launch the EPSO GD hybrid method for a few runs with this static Correntropy setting, the "best" (lowest cost) solution that will be found will be the one which is not intended. Of course, if parameter α is tuned to be even tighter, eventually the deceptive optimum becomes local again because its smaller residuals become larger than α and grow in cost. However, finding an optimum becomes even more difficult as the convergence zones become narrower.

Instead, we will employ a dynamic landscape strategy in order to still take advantage of the better delimitation, where two different settings for Correntropy are used, swapping from one to the other amidst the process. These two settings are the following.

Correntropy setting				
1	α	0,06	β	8
2		0,06		2

The hybrid EPSO GD method with the will still be employed, firstly without additional mechanisms or multi-gradient. The strategy is as follows:

1. Find the LS estimate (as always).
2. Launch EPSO particles from said estimate, with Correntropy setting nº 1.
3. After the swarm season, launch the gradient agent from the swarm's shared reference, keeping the same Correntropy settings.
4. After the gradient season, relaunch the gradient agent from its previously found solution, this time with Correntropy setting nº 2.

Table 5.9 - Results of 20 runs for Generalised Correntropy approach with hybrid EPSO GD method.

Case	Success rate	Rate of convergence to nearly-global optimum
50	45%	30%
51	30%	50%
52	25%	35%

Comments

- The success rates with the new Correntropy strategy barely change relatively to the Classical Correntropy approach. This is justifiable by what has been verified on what happens to the nearly-global optima. Since their residuals only contain a residual larger than α, that respects to the regularity and whose cause is gross error conformity, their costs become much lower and they are kept as the solution of the first Correntropy stage every time they are found.

The problem is that, even if the intended solution was found in the first Correntropy stage, it gets replaced by the deceptive optimum.

Next, we will experiment with the same Correntropy strategy but now using multi-gradient extension, launching it firstly from the personal references of the swarm season for Correntropy setting nº1 and then relaunching it from the solutions of the first gradient season now with Correntropy setting nº2. It is expected that the method firstly finds the deceptive optimum (since it is the global optimum for Correntropy setting nº 1). However, it is hoped that at least one particle falls into the intended solution's convergence zone and keeps its reference there. This is will then guarantee that, when the second multi-gradient season is launched with the Classical Correntropy setting (β = 2), the intended solution, which will now be the global optimum instead, will be found. The strategy is as follows:

Table 5.10 - Results of 20 runs for Generalised Correntropy approach with multi-gradient extension.

Case	Success rate	Rate of convergence to nearly-global optimum (1st Correntropy stage)	Rate of convergence to nearly-global optimum (2nd Correntropy stage)
50	100%	100%	0%
51	95%	100%	5%
52	85%	100%	15%

Table 5.11 - Statistics of the spawn of extra agents for Generalised Correntropy approach.

Case	Extra agents spawn rate for ... (number of) agents				Average number of extra agents
	1	2 to 4	5 to 9	10 or more	
1	0%	10%	20%	70%	13,6
2	0%	0%	30%	70%	12,3
3	0%	20%	30%	50%	13,7

The following figures show the progress of a successful run for case nº 50 during each of the Correntropy stages.

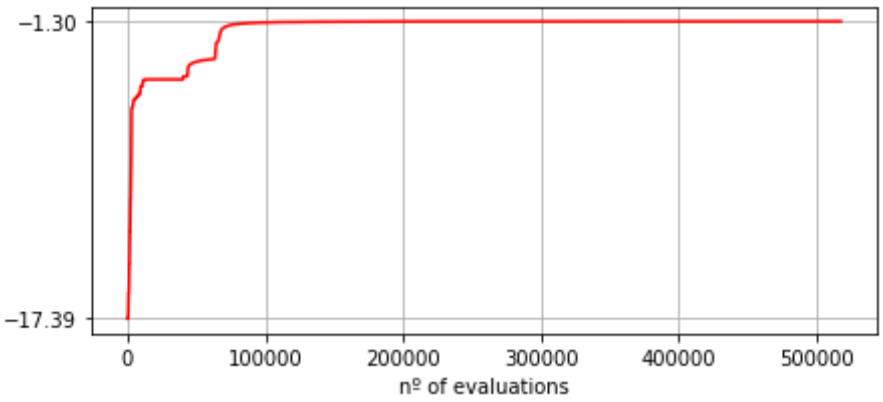


Figure 5.28 - Progress of the hybrid method for case nº 50 (swarm season and multi-gradient season) with Generalised Correntropy.

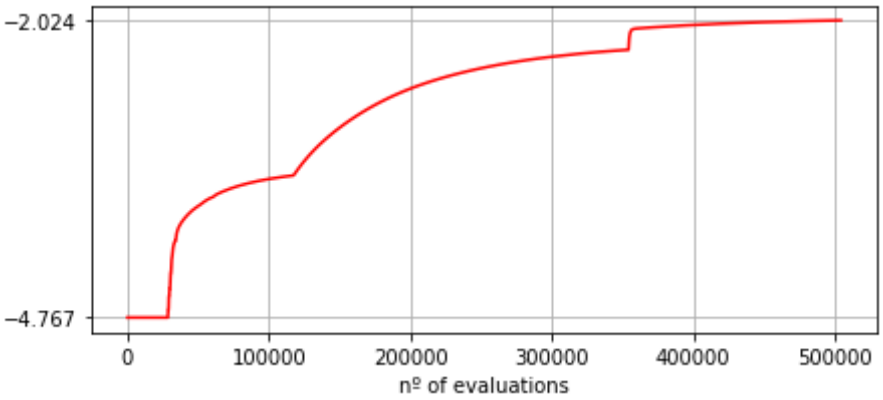


Figure 5.29 - Progress of the relaunched multi-gradient season for case nº 50 with Classical Correntropy.

The following figure shows the aggregated progress of a successful run for all cases. The sudden "dip" of progress that is seen roughly halfway through is the instant at which the Generalised Correntropy stage / first gradient season ends and the Classical Correntropy stage / second gradient season starts. This dip indicates that the cost of the deceptive optimum grows as we change the landscape. Somewhere nearby, and agent that had previously settled on the zone of the intended solution now finds its way to the exact location of the optimum, which is now global.

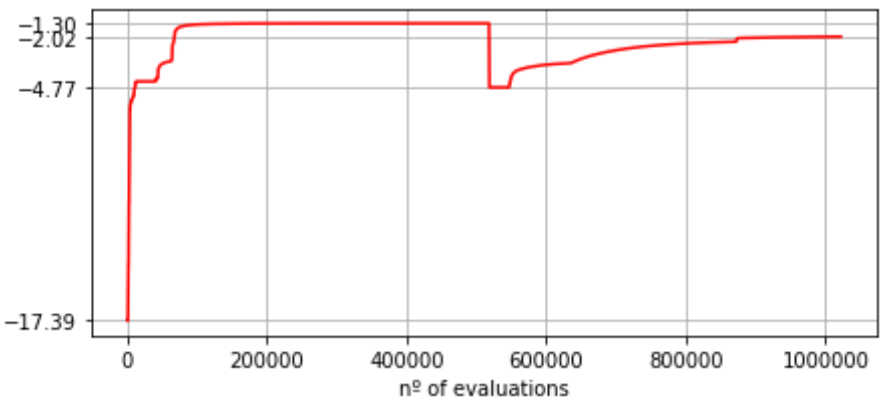


Figure 5.30 - Full progress for all aggregated seasons, for case nº 50.

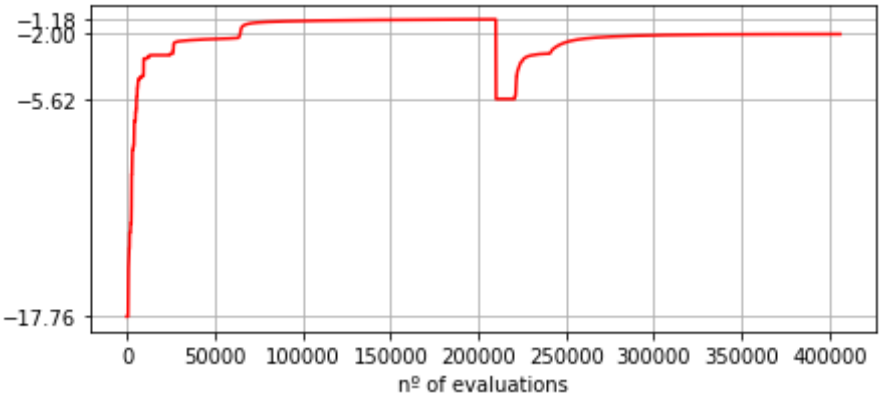


Figure 5.31 - Full progress for all aggregated seasons, for case nº 51.

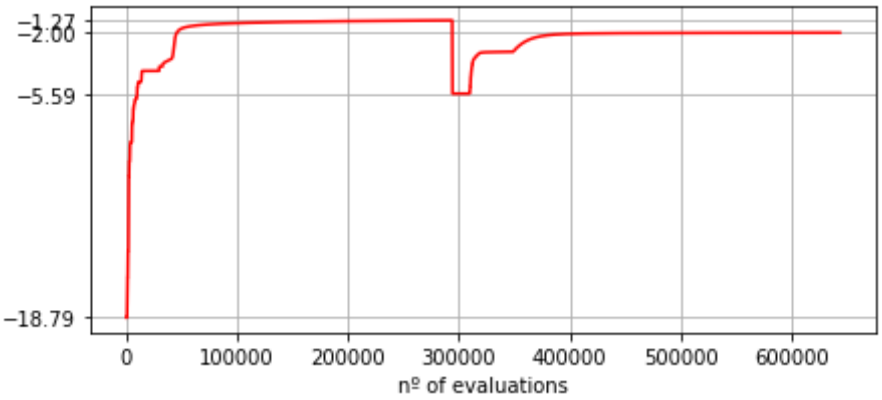


Figure 5.32 - Full progress for all aggregated seasons, for case nº 52.

Comments

- The success rates have spiked up noticeably with this new approach but so has the computational effort for each run. Even if the methodology was refined and hypothetically achieved a 100% success rate for all cases, it is still uncertain whether there would be advantage in using this strategy over running the hybrid method without the multi-gradient extension a few times.
- Arguably, the number of iterations for the swarm season and for the gradient seasons could be better adjusted in order to save some effort, by applying a stop / convergence criterion for each season. Also with that same aim, a limit could be set on the number of extra gradient agents that may spawn, tuned hand in hand with the penalty threshold distance.
- For the first multi-gradient season, the deceptive optima were always the estimates at which the method settled. This is confirmation that a Correntropy setting with a larger value for ß than for Classical Correntropy may not be reliable as a static setting.

Chapter 6

Conclusions

6.1 – On the results

The thesis focused on providing insight on the difficulty of handling anomalous situations that occur in power systems SE, specifically the particular problem of conforming gross errors that may demand an alternate approach than what is currently adopted. Some alternative methodologies where tested to tackle this issue, with varying degrees of success but all commonly confirming that the difficulty exists. In more detail:

- SE via LS is still a reliable approach in the majority of scenarios. It deals well with single gross errors, sometimes resulting in 1 or 2 FP (false positives) but always detecting the anomaly. For cases with two gross errors, the occurrence of FP is more common. It fails to detect all anomalies for some cases with gross error conformity.
- Using a criterion based on Correntropy helps to cast out the outliers and find the most coherent estimate. A numerical method with intelligent initialisation is sufficient for most cases, although yet again conforming gross errors cause problems since they generate a deceptive local optimum in the solutions space at which the method gets trapped.
- The employment of a population based meta-heuristic for exploration of the solutions space, alongside the numerical method for faster convergence, proves successful in finding the desired solution / global optimum given that enough runs are performed. Nonetheless, the rate of convergence towards the aforementioned deceptive optimum is nearly as high as the success rate, even for multiple different settings for the method's parameters and with the use of additional mechanisms. An extension of the hybrid method by launching the numerical method from multiple interesting solutions found by the meta-heuristic improves the success rate but might require much greater computational effort.
- The generalisation of the Correntropy function has usefulness, as we can better delimit convergence zones for the agents to find, but it may affect the optima undesirably or require a more aggressive Correntropy setting that hinders the

optimisation process. The adoption of a dynamic landscape with different stages corresponding to different Correntropy settings boosts the success rate significantly.

Although all tests in this thesis were done adopting the DC power model, the work demonstrates how hard it is to overcome the problems caused by conforming gross errors. Conforming gross errors create a very deceptive landscape that makes convergence difficult , even for meta-heuristics that usually show quite a good success rate in multi-modal problems.

The Maximum Correntropy Criterion (including in its modern form of Generalised Correntropy) allows the desired (correct) state variables estimation to stand out as the global optimum of the optimization problem defined as the SE formulation – however, this global optimum, in the cases tested, seems to have a very narrow basin of attraction, in contrast with a deceptive local optimum formed by the conforming gross errors. This explains why the optimization algorithms that were experimented could not easily find the correct solution and were lured into an erroneous solution.

It must be said that conforming gross errors is most certainly a rare event, so in a control centre this situation is will not occur often. However, the confidence of operators can only improve if the methods and tools available, that work on data collected and put together at the SCADA, are proven to be robust. This thesis has demonstrated how hard this task might be, and why it is necessary to continue to develop work to understand the impact of conforming gross errors and algorithmic solutions to overcome them.

6.2 – Future work

The tests that were carried out are by no means exhaustive. A relatively small test power system was considered and its DC model was adopted. Also, the same point of operation (load profile) was considered throughout. More experimentation could be done with the hybrid method in terms of different parameter settings and additional useful mechanisms and also in terms of resource usage / iterations, so as to save resources (time) while keeping the same effectiveness.

Even so, the conclusions taken from the outcomes of the tests can be generalised and used as reference for future experiments. Concretely, in the future:

- Test power systems could be approached using the AC model. Different topologies, redundancy factors and the consideration of reactive power in the measurements may originate other interesting potential situations of anomaly of varying difficulty.
- Situations with a greater number of conforming gross errors (three or more) could be simulated. It would be interesting to see up to how many can be successfully detected for a same instance.
- In order to make the hybrid EPSO GD method, or other methods, more efficient and competitive at tackling Correntropy SE, a parallelised implementation could be done in a GPU (Graphical Processing Unit) or any other device with that purpose.
- Further research could be done on the usefulness of Generalised Correntropy. Other parameter settings that are more carefully tuned and eventually used as the stages of a dynamic landscape strategy as the one presented may result in higher success rates. Also on the topic of Correntropy, new strategies could be developed in which different kernels could be applied to certain measurements with varying degrees of coherence or suspicion, based on some rule.

References

[1] A. Monticelli, *State Estimation in Electric Power Systems: A Generalised Approach*, 1st edition, Springer US, 1999.

[2] A. Monticelli, A. Garcia, "Reliable Bad Data Processing for Real-Time State Estimation", in *IEEE Power Engineering Review*, 1983.

[3] A. Abur and A. G. Expósito, *Power System State Estimation: Theory and Implementation*, 1st edition, CRC Press, 2004.

[4] E. M. Lourenço, L. B. Souza, A. S. Costa, "A Unified Approach for Bad Data and Topology Error Identification in Generalized State Estimation", in *16th Power Systems Computation Conference*, 2008.

[5] M. Kezunovic, S. Meliopoulos, V. Venkatasubramanian, V. Vittal, *Application of Time-Synchronized Measurements in Power System Transmission Networks*, 1st edition, Springer International Publishing, 2014.

[6] B. N. Rao, R. Inguva, "Power System State Estimation Using Weighted Least Squares (WLS) and Regularized Weighted Least Squares (RWLS) Method", in *Int. Journal of Engineering Research and Applications*, 2016.

[7] V. Freitas, A. S. Costa, "Integrated State & Topology Estimation Based on A Priori Topology Information", in *IEEE Eindhoven PowerTech*, 2015.

[8] E. Parzen, *Statistical Methods on Time Series by Hilbert Space Methods*, 1st edition, Applied Mathematics and Statistics Laboratory, Stanford University, 1959.

[9] E. Parzen, "On the estimation of a probability density function and the mode", in *The Annals of Mathematical Statistics*, 1st edition, Institute of Mathematical Analysis, 1962.

[10] A. W. Bowman, A. Azzalini, *Applied Smoothing Techniques for Data Analysis*, 1st edition, Oxford University Press, 1997.

[11] N. Cardoso, R. Abreu, "Kernel Density Estimate-Based Approach to Component Goodness Modeling", in *27th AAAI Conference on Artificial Intelligence*, 2013.

[12] A. Gunduz, A. Hedge, J. C. Príncipe, "Correntropy as a Novel Measure for Nonlinearity Tests", in *International Joint Conference on Neural Networks*, 2006.

[13] W. Liu, P. P. Pokharel, J. C. Príncipe, "Correntropy: A Localized Similarity Measure", in *International Joint Conference on Neural Networks*, 2006.

[14]W. Liu, P. P. Pokharel, J. C. Príncipe, "Correntropy: Properties and Applications in Non-Gaussian Signal Processing", in *IEEE Transactions on Signal Processing*, 2007.

[15]B. Cheng, L. Xing, H. Zhao, N, Zheng, J. C. Príncipe, "Generalized Correntropy for Robust Adaptive Filtering", in *IEEE Transactions on Signal Processing*, 2015.

[16]C. A. Loza, J. C. Príncipe, "Generalized Correntropy Matching Pursuit: A novel, robust algorithm for sparse decomposition", in *Internation Joint Conference on Neural Networks*, 2016.

[17]V. Freitas, A. S. Costa, V. Miranda, "Robust State Estimation based on Orthogonal Algorithms and Maximum Correntropy Criterion", in *IEEE Manchester PowerTech*, 2017.

[18]J. K. Opara, V. Miranda, "Information Theoretic State Estimation in Power Systems", Ph.D. dissertation, Faculdade de Engenharia da Universidade do Porto, 2014.

[19]A. Santos, V. Miranda, "Application of Entropy Concepts to Power System State Estimation", M.S. thesis, Faculdade de Engenharia da Universidade do Porto, 2009.

[20]R. J. Bessa, V. Miranda, J. Gama, "Entropy and Correntropy Against Minimum Square Error in Offline and Online Three-Day Ahead Wind Power Forecasting", in *IEEE Transactions on Power Systems*, 2009.

[21]J. C. Príncipe, D. Xu, "Introduction to Information Theoretic Learning", in *Internation Joint Conference on Neural Networks*, 1999.

[22]J. C. Príncipe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, 1st edition, Springer-Verlag New York, 2010.

[23]D. Erdgomus, J. C. Principe, "An error-entropy minimization algorithm for supervised training of non-linear adaptive systems", in *IEEE Transactions on Signal Processing*, 2002.

[24]J. H. Holland, *Adaptation in Natural and Artifical Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, University of Michigan Press, 1975.

[25]H. G. Beyer, "Toward a Theory of Evolution Strategies: Self-Adaptation", in *Evolutionary Computation*, 1995.

[26]U. Bodenhofer, *Genetic Algorithms: Theory and Applications*, 2nd edition, Institut fur Algebra, Stochastik und wissensbasierte mathematische Systeme, 2001/2002.

[27]J. Kennedy, R. Eberhart, "Particle Swarm Optimization", in *IEEE International Conference on Neural Networks*, 1995.

[28]S. Das, A. Abraham, A. Konar, "Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives", in *Advances of Computational Intelligence in Industrial Systems*, 1st edition, 2008.

[29]M. Clerc, "The Swarm and the Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization", in *Proceedings of the 1999 Congress on Evolutionary Computation*, 1999.

[30]N. Fonseca, V. Miranda, "EPSO – Evolutionary Particle Swarm Optimization, a New Method with Applications in Power Systems", in *Transmission and Distribution Conference and Exhibition 2002: Asia Pacific*, 2002.

[31]N. Fonseca, V. Miranda, "New Evolutionary Particle Swarm Algorithm (EPSO) Applied to Voltage / VAR Control", in *14th Power Systems Computation Conference*, 2001.

[32]M. Lovberg, "Hybrid Particle Swarm Optimiser with Breeding and Sub-populations", in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2001.

[33]J. J. Jamian, M. W. Mustafa, H. Mokhlis, M. N. Abdullah, "Comparative Study on Distributed Generator Sizing Using Three Types of Particle Swarm Optimization", in *3rd International Conference Intelligent Systems Modelling and Simulation*, 2012.

[34]J. Pereira, J. Alves, M. Matos, "Optimization of Electrical Distribution Network Operation based on EPSO", in *18th International Conference on Intelligent System Application to Power Systems*, 2015.

Annexes

A.1 – Test system and simulated scenarios

The test system to be used is an example from IEEE's database of power systems for this purpose and is composed by 14 buses and 19 branches. The problem of State Estimation is formulated considering the system's DC model.

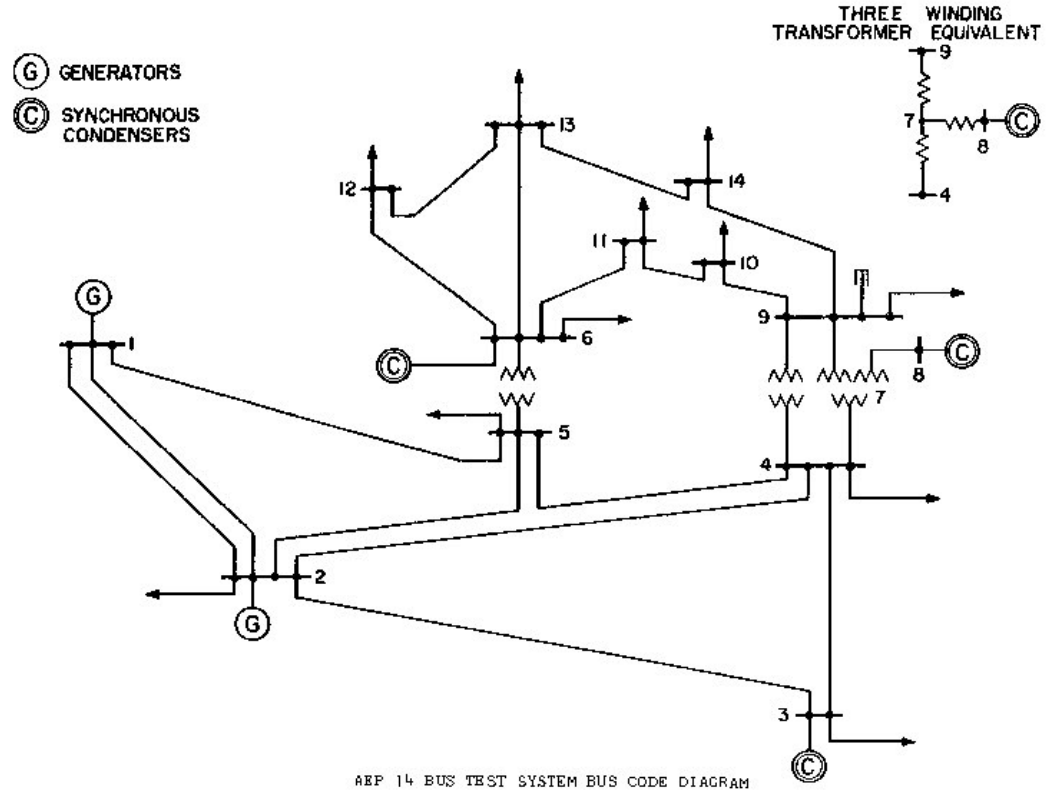


Figure A.1 – IEEE's 14-bus test system.

A fixed point of operation is to be considered for all tests. The quantities for such point of operation regarding the DC model are presented in the following table.

Table A.1 - Point of operation (DC) for the test system.

Bus	angle (rad)	Power injection (p.u.)
1	0,0000	2,190
2	-0,0875	0,183
3	-0,2263	-0,942
4	-0,1850	-0,478
5	-0,1585	-0,076
6	-0,2632	-0,112
7	-0,2464	0,000
8	-0,2464	0,000
9	-0,2787	-0,295
10	-0,2792	-0,090
11	-0,2733	-0,035
12	-0,2711	-0,061
13	-0,2718	-0,135
14	-0,2976	-0,149

Branch	Power flow (p.u.)
1 – 2	1,479
1 – 5	0,711
2 – 3	0,701
2 – 4	0,553
2 – 5	0,408
3 – 4	-0,241
4 – 5	-0,628
4 – 7	0,294
4 – 9	0,168
5 – 6	0,415
6 – 11	0,107
6 – 12	0,065
6 – 13	0,131
7 – 8	0,000
7 – 9	0,294
9 – 10	0,018
9 – 14	0,149
10 – 11	-0,072
12 – 13	0,004

In order for the system to have a moderate redundancy, regarding the available data for the SE exercise, the measurements to be considered are:

- power injections for all buses.
- power flows for all branches.

In total, there are **33** measurements in the measurement set. Since there are **13** system variables (voltage angles, one for each bus, excluding bus 1 which is the reference bus), this set gives the system a redundancy degree of (roughly) **2,5**.

With regards to noise and gross error simulation, the power injections for buses 7 and 8 are to be excluded from the possible targets for contamination since these buses have neither active generation nor active load. The injections, which equal the sum of all power flows into said buses, should always cancel out. As such, they may be considered as perfect (pseudo) measurements.

The anomaly profiles to be simulated are:

1. a single gross error in a flow measure, for every branch.
2. a single gross error in an injection measure, for every bus.
3. a gross error in a flow measurement and another in an injection measure, for some combinations of a bus and a branch which are non-adjacent.

4. a gross error in a flow measurement and another in an injection measure, for some combinations of a bus and a branch which are adjacent, subdivided into:

   a) conflicting errors.

   b) conforming errors.

The magnitude for each gross error is sampled from between **20%** and **40%** of the system's base power (per unit). In cases of non-adjacent elements, the polarity of each error is chosen at random. In cases of adjacent elements, the polarity was purposefully chosen to forge conforming errors. Also, small noise was spread across the data, being sampled from a normal distribution with standard deviation equal to 1% of the base power.

The following tables list the cases to be simulated. The "target(s)" field refers to the measurement(s) which is (are) to be contaminated and the "error(s)" field refers to the error(s) of said measurement(s), where:

- P *i* active power injection at bus *i*;
- P *i* - *j* active power flow from bus *i* to bus *j*.

Here follow the cases with single gross errors.

Table A.2 – Test cases with single gross errors.

Nº	Target	Error
1	P 1	-0,20
2	P 2	-0,40
3	P 3	+0,40
4	P 4	-0,34
5	P 5	+0,35
6	P 6	-0,20
7	P 9	-0,36
8	P 10	+0,39
9	P 11	+0,38
10	P 12	-0,23
11	P 13	-0,30
12	P 14	-0,38

Nº	Target	Error
13	P 1 - 2	+0,33
14	P 1 - 5	-0,22
15	P 2 - 3	-0,24
16	P 2 - 4	+0,37
17	P 2 - 5	-0,31
18	P 3 - 4	+0,38
19	P 4 - 5	-0,25
20	P 4 - 7	+0,24
21	P 4 - 9	-0,23
22	P 5 - 6	+0,30
23	P 6 - 11	-0,33
24	P 6 - 12	-0,21
25	P 6 - 13	+0,32
26	P 7 - 8	-0,34
27	P 7 - 9	+0,31
28	P 9 - 10	+0,33
29	P 9 - 14	-0,22
30	P 10 - 11	+0,38
31	P 12 - 13	+0,23

Here follow the cases with two gross errors, which were each built from merging two cases of single errors together. The cases with non-adjacent elements are on the left table

and the cases with adjacent elements are on the right table. Cases with conforming errors are highlighted since they are a focus of the thesis.

Table A.3 - Test cases with two gross errors.

Nº	Targets	Errors
32	P 1	-0,20
	P 2 - 3	-0,24
33	P 2	-0,40
	P 4 - 5	-0,25
34	P 3	+0,40
	P 4 - 9	-0,23
35	P 4	-0,34
	P 1 - 5	-0,22
36	P 5	+0,35
	P 12 - 13	+0,23
37	P 6	-0,20
	P 10 - 11	+0,38
38	P 9	-0,36
	P 2 - 5	-0,31
39	P 10	+0,39
	P 5 - 6	+0,30
40	P 11	+0,38
	P 2 - 4	+0,37
41	P 12	-0,23
	P 3 - 4	+0,38
42	P 13	-0,30
	P 7 - 9	+0,31
43	P 14	-0,38
	P 6 - 12	-0,21

Nº	Targets	Errors
44	P 1	-0,20
	P 1 - 2	+0,33
45	P 2	-0,40
	P 2 - 4	+0,37
46	P 3	+0,40
	P 3 - 4	+0,38
47	P 4	-0,34
	P 4 - 7	+0,24
48	P 5	+0,35
	P 2 - 5	-0,31
49	P 6	-0,20
	P 6 - 13	+0,32
50	P 9	-0,36
	P 7 - 9	+0,31
51	P 10	+0,39
	P 10 - 11	+0,38
52	P 11	+0,38
	P 6 - 11	-0,33
53	P 12	-0,23
	P 6 - 12	-0,21
54	P 13	-0,30
	P 12 - 13	+0,23
55	P 14	-0,38
	P 9 - 14	-0,22

A.2 – Figure mapping

A residuals figure displays the distribution of the residuals for a certain estimate. Each residual is mapped to its respective measurement as follows, where:

- P *i* active power injection at bus *i* (per unit of system base power);
- P *i* - *j* active power flow from bus *i* to bus *j* (per unit of system base power).

Table A.4 - Mapping of residuals.

1	P 1	2	P 2	3	P 3	4	P 4
5	P 5	6	P 6	7	P 7	8	P 8
9	P 9	10	P 10	11	P 11	12	P 12
13	P 13	14	P 14	15	P 1 – 2	16	P 1 – 5
17	P 2 – 3	18	P 2 – 4	19	P 2 – 5	20	P 3 – 4
21	P 4 – 5	22	P 4 – 7	23	P 4 – 9	24	P 5 – 6
25	P 6 – 11	26	P 6 – 12	27	P 6 – 13	28	P 7 – 8
29	P 7 – 9	30	P 9 – 10	31	P 9 – 14	32	P 10 – 11
33	P 12 – 13						

An errors figure displays the errors of the variables' values for a certain estimate. Each error is mapped to its respective variable as follows, where:

- θ *i* voltage angle at bus *i* (radians).

Table A.5 - Mapping of errors.

1	θ 1	2	θ 2	3	θ 3	4	θ 4
5	θ 5	6	θ 6	7	θ 7	8	θ 8
9	θ 9	10	θ 10	11	θ 11	12	θ 12
13	θ 13	14	θ 14				