Håkon Kile

Evaluation and Grouping of Power Market Scenarios in Security of Electricity Supply Analysis

Thesis for the degree of Philosophiae Doctor

Trondheim, March 2014

Norwegian University of Science and Technology
Faculty of Engineering Science and Technology
Department of Electric Power Engineering

NTNU – Trondheim
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology, Mathematics and Electrical Engineering
Department of Electric Power Engineering

ISBN 978-82-326-0094-6 (printed ver.)
ISBN 978-82-326-0095-3 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2014:85

Printed by NTNU-trykk

Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) for partial fulfilment of the requirements for the degree of philosophiae doctor (PhD).

This work has been financed by the research project "Integration of methods and tools for security of electricity supply analysis", administrated by SINTEF Energy Research, Trondheim, Norway.

I would like to thank my supervisors Professor Kjetil Uhlen and Adjunct Professor Gerd H. Kjølle for their guidance and support.

Trondheim, February 2014
Håkon Kile

Summary

In power system reliability assessment, at the transmission system level, a basic requirement of the analysis is models describing load and generation profiles at the individual buses, as these variables are needed as input to the power flow problem. As a consequence of deregulation and the increasing share of renewable energy sources in the generation system, the modelling of load and generation patterns has, over the last decades, become more difficult.

For electric power systems, it is often the case that a certain mix of load and generation (spread out at the individual buses) cause the system to be vulnerable, as it might lead to, e.g., too high load on certain transmission lines or too little generation reserve in certain areas. An important aspect of power system reliability assessment is to reveal and analyse such vulnerabilities. Thus, it is necessary to check different mixes of generation and load in conjunct with (a large number of) component outage combinations to reveal as many system vulnerabilities as possible.

Typically, only peak load situations are analysed in power system reliability assessments. However, the transmission system can be vulnerable due to bottlenecks in the system, as a consequence of large local loads and/or non-dispatchable generation, even though the total demand might be well below the peak demand. To improve the generation and load modelling, the reliability assessment framework discussed in the thesis use a power market model to generate (future) load and generation scenarios. The generated scenarios are interpreted as a sample of future power market behaviour, and are used as basis for a reliability assessment. Depending on the type of power market model, this analysis can take into account different market structures, renewable energy sources, hydro inflow scenarios, etc.

The power market models tends to produce a large number of load and generation scenarios, and to include all these scenarios in a reliability assessment results in excessive computation time. In this thesis, the scenario selection

method is proposed and discussed. The scenario selection method finds groups of similar load and generation scenarios, and then, for each group, chooses one scenario to represent the group characteristics. The set of chosen scenarios is denoted the representative set, and only this set of scenarios is used as basis for a reliability assessment. Essentially, the scenario selection method keeps the sample variation of the full sample of load and generation scenarios more or less intact, but at the same time severely reduces the computational requirements of the reliability assessment. Based on the results of the case studies in this thesis, these main conclusions (with respect to application of the scenario selection method) are drawn:

- Feature selection, i.e., selection of variables used for quantifying similarity among load and generation scenarios, is a case dependent process, and ideally the feature selection should be customised to suit the analysis to get optimal results. However, the power injections (P and Q at the buses in the power system under study) give good results when used as features for scenario selection, and is the best general recommendation.

- The problem of finding groups of similar load and generation scenarios is most likely a segmentation problem. Thus, there is no objective method for determining the number of natural groups k in the set of load and generation scenarios, and k must be selected based on experience. The results of the case studies indicate that $k \approx 0.1 \cdot n$ is a sensible choice (where n is the total number of load and generation scenarios).

- In a segmentation problem, the goal of the clustering algorithm is to group together load and generation scenarios which are very similar. An agglomerative clustering algorithm, with complete linkage, tends to produce small and compact groups, and is therefore thought of as an appropriate algorithm for this problem. The results of the case studies support this assumption, as agglomerative clustering with complete linkage is, overall, the algorithm with the best results when applied for scenario selection. The k-means algorithm gives similar results as the agglomerative clustering algorithm, while more advanced clustering algorithms perform at the same level, or worse, than the agglomerative clustering algorithm when applied for scenario selection.

- The scenario selection method can reduce the computation time of the reliability assessment by about 90% (by setting $k \approx 0.1 \cdot n$), and simultaneously keep the error in the reliability indices within a 5-10% margin.

This is thought of as an appropriate trade-off between reduced computation time and maintaining the accuracy of the analysis.

An alternative method for reducing the computation time of the reliability assessment, is to only analyse those scenarios with the highest overall system load, e.g., exclude 90% of the load and generation scenarios from the analysis by taking away those with a "low" system load. The scenario selection method is a more systematic and consistent method for reducing the scenario set compared to such (rule-based) methods, and it is shown that the scenario selection method gives much better results than such (rule-based) methods.

Abbreviations

ACCL	Agglomerative Clustering with Complete Linkage
ACPF	AC Power Flow
ATC	Available Transfer Capacity
CFM	Cascading Failure Model
CENS	Cost of Energy Not Supplied
DCPF	DC (linearised) Power Flow
EMPS	The Multi-Area Power Market Simulator
EMPS-NC	The Multi-Area Power Market Simulator with Network Constraints
EENS	Expected Energy Not Supplied
HILP	High Impact Low Probability
IGMM	The Infinite Gaussian Mixture Model
KM	K-Means
MCMC	Markov Chain Monte-Carlo
MRM	Minimal Rescheduling Model
OPAL	Norwegian abbreviation for Optimisation of Reliability of Supply
SAC	System Available Capacity
SAMREL	Integration of Methods and Tools for Security of Electricity Supply Analysis
SOM	Self-Organising Map
SoS	Security of Electricity Supply
TSO	Transmission System Operator
TV	Target Values

Nomenclature

Reliability Indices (OPAL)

d	delivery point index
$EENS^a_{i,d}$	annualised expected energy not supplied at delivery point d for operating scenario i
$EENS^a_{y,d}$	annual expected energy not supplied at delivery point d for year y
h_i	duration (in hours) of operating scenario i
h_{year}	duration (in hours) of an "EMPS year"
i	operating scenario index
j	index for contingencies and minimal cuts
λ_j	(equivalent) yearly failure frequency for minimal cut j
$n_{mc,i,d}$	number of minimal cuts for delivery point d for operating scenario i
$P_{i,d}$	demand at delivery point d for operating scenario i as given by EMPS-NC
$P^{\text{inter}}_{i,j,d}$	interrupted power at delivery point d after the occurrence of contingency j for operating scenario i
r_j	(equivalent) mean time to repair for minimal cut j
$SAC_{i,j,d}$	available capacity to supply the load (P) at delivery point d after the occurrence of contingency j for operating scenario i
$U^a_{i,d}$	annualised expected interruption duration at delivery point d for operating scenario i
$U^a_{y,d}$	annual expected interruption duration at delivery point d for year y

Scenario Selection

$\widehat{EENS}^a_{i,d}$	estimated annualised expected energy not supplied at delivery point d for operating scenario i
$\widetilde{EENS}^a_{y,d}$	estimated annual expected energy not supplied at delivery point d for year y
k	number of groups in the representative set
p	dimension of the feature space
$\mathbb{R}^p$	feature space, of dimension p, for operating scenarios
s_i	feature vector representing operating scenario i
$\mathbf{x}_i$	feature vector representing operating scenario i
$x_{i,j}$	feature j representing operating scenario i
X_f	feature set number f
$\widehat{U}^a_{i,d}$	estimated annualised expected interruption duration at delivery point d for operating scenario i
$\widetilde{U}^a_{y,d}$	estimated annual expected interruption duration at delivery point d for year y
W_k	within cluster dispersion

Contents

Chapter 1

Introduction

The objective of electric power systems is to provide a reliable, sustainable, and economical supply of electricity to end customers. In order to try to maintain a continuous (non-interrupted) supply of electricity at all times, including those periods with generator and/or network component outages due to maintenance or forced outages, extra generation and network capacity is built into the power system. Theoretically, a continuous supply of electricity can be guaranteed at all times if enough redundancy is built into the system. However, this requires huge investments which cannot be economically justified. In addition, excessive generation and network capacity leads to an extremely complex power system, which makes the system very difficult to control and operate. This also increases the sources of errors as more components are built into the power system. Thus, from a practical point of view, the amount of redundancy needs to be economically justified.

In reliability evaluation of power systems, the ultimate goal is to balance the worth of a reliable supply of electricity against the costs associated with reaching (and maintaining) this level of reliability in the electricity supply. The "right" reliability level in the electricity supply is dependent on the society's needs and expectations towards a reliable supply, the customers willingness to pay for a certain reliability level (in terms of the price of electricity and costs incurred due to interruptions), and operation and investment costs related to the power system itself.

1.1 Power System Reliability Assessment

To simultaneously analyse the reliability level of the electricity supply and the corresponding costs is a very complicated procedure, especially for large systems, and is not feasible in practice. To simplify the problem, a common approach is to define some reliability criteria, and use a reliability assessment to check if the power system satisfies these criteria or not. These reliability criteria can be either deterministic, like the loss of largest unit criterion, or probabilistic like the loss of load expectation. Either way, these criteria are defined to represent (approximate) the appropriate reliability level of the electricity supply, considering, e.g., customer expectations and economic consequences. In long-term investment (or operation) analysis, the objective can be to, e.g., minimise the costs subject to meeting the given reliability criteria.

Power system reliability assessment aims at quantifying the system ability to perform its intended task - maintaining a continuous supply of electricity to the end customers, by calculating a set of reliability indices. A range of different reliability indices are used to describe different aspects of the system reliability [1, 2].

In its most basic from, a reliability assessment is done to check if there is enough generation capacity in the system to meet the total demand (at time T). For such an analysis, a model describing the availability and maximum power output of each individual power generator, and a model for the total demand in the system, (at time T) are required. To analyse the effect of transmission constraints, a model of the transmission system has to be incorporated in the reliability assessment. Further extensions of the reliability assessment include incorporation of models describing the distribution system, distributed generation, and/or the protection system [2]. The (power system) model, used in the reliability assessment, quickly turns into a very complex one as more detailed descriptions are included.

In power system reliability assessment, the description of the physical components in the power system is often not the main problem. For instance, in a situation where the reliability of a transmission network is investigated, a model for solving the power flow problem of the transmission network itself is usually available. The main challenges are rather to come up with sensible models for the load and generation profiles at the individual buses, and models describing the availability of the power system components. As a consequence of deregulation and the increasing share of intermittent energy sources in the generation system, the modelling of load and generation patterns has, over the last decades, become more difficult but simultaneously become a more important part of the

reliability assessment. The calculations required to do a reliability assessment increases as the model complexity increases.

At the most basic level, a common approach to limit the computational requirements of the reliability assessment is to only analyse the peak load situation and/or situations where a large portion of the generation system is out due to scheduled maintenance. However, the transmission system can be vulnerable due to bottlenecks in the system, as a consequence of large local loads and/or non-dispatchable generation, even though the total demand might be well below the peak demand. This is the motivation behind this PhD project, where the objective has been to find a set of load and generation patterns to be used as input to the reliability assessment, where the set should represent the variation in load generation composition at the individual buses and at the same time be small enough as to severely restrict the computation time of the reliability assessment. Thus, this thesis focuses on the load and generation patterns themselves, and not the problem of finding the cut-off point for the contingency level to be included in the (analytical) reliability assessment, e.g., first, second, of higher order outages [3].

1.2 Integration of Methods and Tools for Security of Electricity Supply Analysis

This work is a contribution to the research project "Integration of methods and tools for security of electricity supply analysis" (SAMREL). The SAMREL project is a knowledge-building project for industry, carried out by SINTEF Energy Research and The Norwegian University of Science and Technology (NTNU).

The main objective of the SAMREL project is to establish a comprehensive methodology for security of electricity supply (SoS) analysis, by the integration of power market models with detailed network and reliability models for contingency and reliability analysis [4, 5]. SoS is defined as the ability of the power system to supply final customers with electricity. SoS is composed by energy availability, power capacity, and reliability.

The SAMREL SoS analysis framework consists of three main modules - power market analysis, contingency analysis, and reliability analysis, as shown in Figure 1.1. Analytical contingency enumeration techniques are used for evaluating reliability indices. The SAMREL SoS analysis aims at (adequately) representing the risks and uncertainties in the electricity supply related to the stochas-

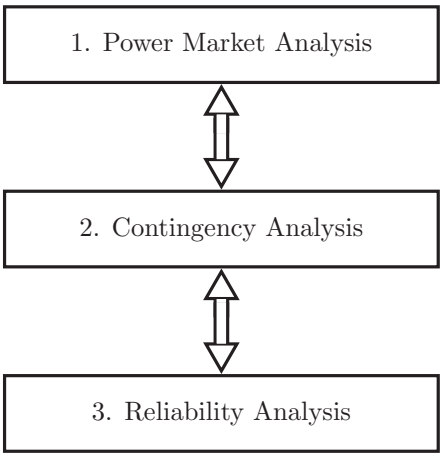


Figure 1.1: Illustration of the three main modules in the SAMREL SoS analysis framework.

tic nature of renewable generation (such as wind and hydro power), stochastic component failures, and uncertainties in the price of electricity and the power market itself. The two major goals of the SAMREL project are to develop the data exchange (interface) between the power market module and contingency and reliability analysis, and to improve the analysis methodologies used in the contingency and reliability analysis modules.

The SAMREL SoS analysis framework, in its current version, is most suitable for long-term analysis. However, with some adjustments, the framework can be used for operation planning and for on-line operation as discussed in [5]. Only long-term analysis is considered in the remaining parts of this thesis.

1.2.1 Power Market Analysis

The power market analysis module, in the upper part of Figure 1.1, focuses on (long-term) market analysis, where the stochastic nature of future load and generation schedules is included in the analysis. Uncertainties related to the power market and the price of electricity are also included in the analysis in this module. A recent study of the Nordic power system [6] revealed risks associated with capacity and energy shortage, as the power market itself does not always respond fast enough to prevent such problem. Thus, there is a

need for connecting power market models with network and reliability models to represent and analyse risks and vulnerabilities related to all aspects of the electricity supply.

The Multi-area Power-market Simulator (EMPS) [7] is used for the power market analysis. EMPS is designed for simulation of hydro-thermal power systems, where the market analysis is done by finding the optimal socio-economic dispatch considering, e.g., different hydro inflow scenarios and unit commitment cost. EMPS takes into account the stochastic nature of hydro inflow, wind speed, and temperature by using historic weather data. The historic weather data are regarded as expected future scenarios, where typically 50-75 years of historic data are used for a planning horizon of 3-5 years. The load model (in EMPS) is set to approximate the expected load in this 3-5 years period.

The annual variation in hydro inflow, wind speed, and temperature (in the analysis period) is taken care of by using the sample of 50-75 years as representative for the planning horizon. For each (historic) year, EMPS optimises the operation of the hydro-thermal power system per week by splitting the week into load periods or do the dispatch per hour. Reference [7] and [8] provide details regarding the within-week optimisation procedures in EMPS. For each hour or load period, EMPS gives a generation and load schedule. For instance, if a week is split into five load periods, EMPS generates $5 \times 52 = 260$ load and generation schedules which represent the historic "EMPS year", for which the operation is currently optimised. (An "EMPS year" is 52 weeks, not 365 days.) These load and generation schedules represent the daily, weekly, and seasonal variation in load and generation, and are regarded as representative for the load and generation profile over the whole year. This is an improvement over most conventional methods of power system reliability assessment, where typically only the peak load situation is analysed. These load and generation schedules are denoted operating scenarios (or operating states).

1.2.2 Contingency and Reliability Analysis

The operating scenarios are the main input from the power market analysis to the contingency and reliability analysis, as indicated by the upper arrow in Figure 1.1. In fact, the operating scenarios are the focal point of this thesis, as they are both the basis for the contingency and reliability analysis, and one of the main reasons behind the high overall computation time of the SAMREL SoS analysis. The contingency and reliability analysis is briefly discussed below, and further elaborated on in chapter 2. The reasons for the high computation time is also briefly discussed in that chapter. Different alternatives for reducing

the computation time, are discussed in detail in chapter 3 and 4.

For each operating scenario, the consequences of a set of contingencies are determined. A contingency is here defined as an outage of one or more components due to failures, where the failures can be caused by, e.g., technical or human factors. A typical analysis depth is to analyse all first and second order contingencies (for each operating scenario).

The contingency analysis determines which delivery points experience an interruption during a contingency (if any), and determines the severity of the interruption in the form of, e.g., load shedding or voltage deviation at the delivery points.

Reliability indices are evaluated per delivery point per operating scenario, by using analytical techniques, based on minimal cuts and approximate techniques [9]. The evaluated indices include the frequency and duration of interruptions, expected energy not supplied (EENS), and the corresponding cost of energy not supplied (CENS). The per operating scenario indices are combined to give annual indices for each of the historic years (included in the EMPS analysis). System indices are derived from the delivery point indices.

1.2.3 Computation Time

By considering 5 operating scenarios per week for 75 years, and 1000 contingencies to be tested for each operating scenario, a total of about 19.5 million power flow problems have to be solved in the contingency analysis, compared to 1000 power flow problems if only looking at one operating scenario, e.g., the peak load situation. The number of required power flow solutions grows rapidly as the number of parameters increases (e.g., system size and analysis depth), and quickly turns into an infeasible problem with today's computer technology.

This high computational cost of the contingency analysis in the SAMREL SoS analysis framework is the main motivation behind this PhD work, as the goal of the PhD project has been to reduce the number of operating scenarios which has to undergo a full contingency and reliability analysis must, while at the same time maintaining the (stochastic) variation in the annual reliability indices per historic (hydro inflow) year.

1.3 Main Contributions

The main contribution of this PhD project is the definition and implementation of the scenario selection method, for reducing the number of operating scenarios

used as input to the contingency and reliability analysis, in the SAMREL SoS analysis framework.

The scenario selection method finds groups of similar operating scenarios (generated by the power market analysis module in Figure 1.1), and then, for each group, chooses one scenario to represent the group characteristics. The set of chosen scenarios is denoted the representative set, and only this set of scenarios is transferred to the contingency analysis module. Reliability indices are found for each scenario in the representative set. It is assumed that all the scenarios within a group have the same value of the reliability indices as the group representative, and this information is used to estimate the annual reliability indices per historic year.

Essentially, the scenario selection method keeps the sample variation of the full sample of operating scenarios more or less intact, but at the same time severely reduces the computational requirements of the contingency analysis module in the SAMREL SoS analysis.

In the development of the scenario selection method, several research questions had to be answered, where the three main questions are given below.

- What makes operating scenarios similar?
- What is the best algorithm for finding groups of similar operating scenarios?
- How many groups of operating scenarios should the representative set consist of?

As the SAMREL SoS analysis currently is under development as a part of the SAMREL research project, an SoS analysis for testing and benchmarking of the scenario selection method had to be defined and implemented. This specific SoS analysis version is defined with respect to an analysis of the Nordic power system. This SoS analysis required:

- The establishment of a linearised power flow model for the Nordic power system, where Sweden is replaced with a 30 bus network equivalent. This power flow model is used in both the power market analysis and the contingency and reliability analysis.
- Definition and implementation of the minimal rescheduling model (MRM), which is used in the contingency analysis of operating scenarios. With MRM, the consequences of forced outages (contingencies), which lead to violations of the operating criteria, are minimised with respect to the

cost of the corrective actions needed to bring the system back within its operating limits.

Within the defined SoS analysis, the scenario selection method can reduce the time it takes to complete the contingency and reliability analysis by 90%, if a 10% error in the (estimated) reliability indices is deemed acceptable. Thus, the main contributions of this thesis can be seen as twofold - theoretical and practical. The theoretical contributions lie in the definition of the scenario selection method - a data reduction framework for contingency and reliability assessment, while the case studies are used to determine the optimal values of the parameters used in the scenario selection process.

An alternative method for reducing the computation time of the contingency and reliability analysis, is to only analyse those operating scenarios with the highest overall system load, e.g., exclude 90% of the operating scenarios from the analysis by taking away those with a "low" system load. The scenario selection method is a more systematic and consistent method for reducing the scenario set compared to such (rule-based) methods, and it is shown that the scenario selection method gives much better results than such (rule-based) methods. The scenario selection is marginally slower than such (rule-based) methods with respect to reducing the scenario set, but the time difference vanishes when compared to the time it takes to complete the contingency analysis.

As the scenario selection method is based on unsupervised learning, the scenario selection method of data reduction has (theoretically) been compared with the application of supervised learning as a method for data reduction (within the SAMREL SoS framework).

1.4 List of Publications

During the course of this PhD project, several scientific papers have been published. The following publications are directly related to the content of this thesis:

- **Publication A** *Supervised and Unsupervised Learning in Composite Reliability Evaluation*, H. Kile and K. Uhlen, IEEE PES General Meeting 2012, San Diego, California, USA, July 2012.
- **Publication B** *Data reduction via clustering and averaging for contingency and reliability analysis*, H. Kile and K. Uhlen, International Journal of Electrical Power & Energy Systems, Volume 43, Issue 1, December 2012, p. 1435-1442.

- **Publication C** *Averaging Operating States With Infinite Mixtures in Reliability Analysis of Transmission Networks*, H. Kile and K. Uhlen, 12th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS) 2012, Istanbul, Turkey, June 2012.

- **Publication D** *Scenario Selection by Unsupervised Learning in Reliability Analysis of Transmission Networks*, H. Kile, K. Uhlen and G. Kjølle, PowerTech 2013, Grenoble, France, June 2013.

- **Publication E** *Scenario Selection in Composite Reliability Assessment of Deregulated Power Systems*, H. Kile, K. Uhlen and G. Kjølle. Submitted to International Journal of Electrical Power & Energy Systems.

- **Publication F** *The Important Role of Feature Selection when Clustering Load and Generation Scenarios*, H. Kile, K. Uhlen and G. Kjølle. Accepted for presentation at The 5th IEEE PES Asia-Pacific Power and Energy Engineering Conference, Hong-Kong, December 2013.

Some of the chapters in this thesis are either directly related to the content, or modified versions, of these publications. This is stated at the beginning of each such chapter.

The research resulted in these additional publications, which falls outside the main scope of this thesis, but are related to research done as a part of this PhD project:

- **Publication G** *Incorporating power market scenarios in the adequacy assessment of the Norwegian power system*, H. Kile and R. Karki, The 2013 Canadian Conference on Electrical and Computer Engineering (CCECE), Regina, Saskatchewan, Canada, May 2013.

- **Publication H** *Extension of Area Risk Concepts to Incorporate Wind Power in Unit Commitment Risk Evaluation*, S. Thapa, R. Karki, R. Billinton, and H. Kile, The 2013 Canadian Conference on Electrical and Computer Engineering (CCECE), Regina, Saskatchewan, Canada, May 2013.

1.5 Thesis Outline

This chapter has introduced the context of, and the motivation behind, the PhD project, as well as the main scientific contributions from this research project.

Chapter 2 briefly discusses the general concept of power system reliability assessment. The general SAMREL SoS analysis framework is discussed in more details in section 2.2, while section 2.3 defines and discusses an SoS analysis set-up, which is suitable for an analysis of the Nordic power system.

Chapter 3 describes the general concept of how supervised and unsupervised learning algorithms are used to reduce the computation time in power system reliability assessment, and draw parallels between the research done within this PhD project and other applications of learning algorithms in both power system reliability assessment and other fields of research.

Chapter 4 introduces and elaborates the scenario selection method. The main focus of this chapter is on the theory behind the method, and how to implement the method into the SAMREL SoS analysis framework. The scenario selection method is the main focus of this PhD study, and thus is this chapter considered as the core of this thesis.

Chapter 5 contains the main findings of the case studies done as a part of this research project. Different version of the scenario selection methods are compared, where the differences relate to both algorithms and parameters. A comparison of the scenario selection method and rule-based methods for data reduction is also included.

Chapter 6 discusses the results and practical implications of the findings in chapter 5. Some concluding remarks, and suggestions for further work, are given in chapter 7

Appendix A contains a detailed mathematical derivation of the multivariate infinite Gaussian mixture model. This model is introduced later in chapter 4. An outline of the Markov chain Monte-Carlo sampling scheme, which is used for drawing inference about the parameters of this model, is also included in the appendix.

Failure rate and average outage time per component type are given in appendix B.

Chapter 2

Power System Reliability Assessment

This chapter contains a discussion of power system reliability assessment. The general idea behind SAMREL SoS analysis was introduced in the previous chapter, while in this chapter, the methodology is discussed in more details. The last part of this chapter describes the specific SoS analysis set-up which has been used for testing and benchmarking of the scenario selection method. This SoS analysis is defined with respect to an analysis of the Nordic power system.

2.1 A General Introduction

As mentioned in the previous chapter, both probabilistic and deterministic techniques can be used for power system reliability assessment. A third possibility is well-being analysis, a hybrid analysis framework, applying both probabilistic and deterministic criteria [10–13]. In the remainder of this thesis, only probabilistic analysis is considered.

2.1.1 Adequacy vs Security

It is common to split the concept of power system reliability into two categories.

- **Adequacy analysis:** This analysis focus on the static (stationary) state of the power system, where the objective is to determine whether or not

the power system can deliver the electric power and energy required by the customers, taking into account forced outages and maintenance schedules [2]. For instance, is there enough transmission capacity in the system to transfer the required power from the generation facilities to the delivery points?

- **Security analysis:** The focus is on the transition between states, i.e., the system response when the system operating point is changed [2], e.g., as a consequence of a contingency. A common question is whether or not the new operating point is stable or unstable, i.e., can the system use the available resources to get to a new stable operating point?

Both adequacy and security analysis are intrinsic parts of power system reliability assessment, and both are needed to fully analyse the system reliability, and should not be seen as separate entities. The split is merely done out of a practical need for structure and clarification of what aspect of the system reliability which is currently analysed (and quantified).

Security analysis typically deals with a time frame from milliseconds up to a few hours, while adequacy analysis often deals with the long term perspective, e.g., from hours to many years ahead.

Probabilistic reliability assessment techniques are most commonly applied in the context of adequacy analysis, but it is an on-going research task to develop probabilistic analysis for security analysis as well [2, 14, 15]. In this thesis, only adequacy analysis is considered.

2.1.2 Hierarchical Levels

As electric power systems are very large and complex system, it is common to split the system into subsystem which are analysed separately. Most commonly, the power system is split into three hierarchical levels (subsystems), as shown in Figure 2.1:

- **HL-I:** This is usually denoted generation reliability assessment, and the objective is to check whether or not there is sufficient generation capacity in the system to supply the total demand.

- **HL-II:** This is often denoted composite reliability assessment, and includes an analysis of both the generation and the transmission system. The objective of the analysis is to quantify the reliability of electricity supply to individual delivery (bulk supply) points in the system, as well as for the whole system.

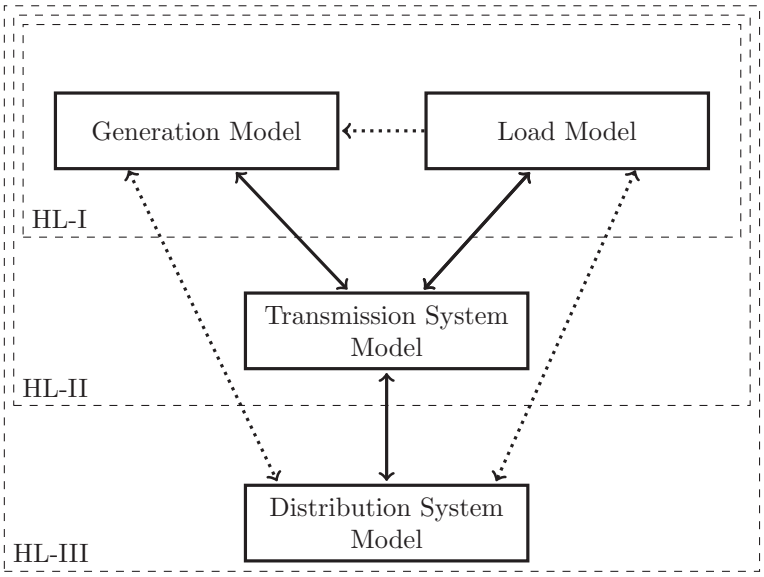


Figure 2.1: Illustration of the three hierarchical levels in power system reliability assessment.

- **HL-III:** This includes an analysis of the generation, transmission, and distribution system, where the objective is to analyse the reliability of supply to the individual customers connected to the distribution (and transmission) network. However, as the distribution system mainly consists of radial feeders connected to the bulk supply points, distribution system reliability is often done separately (and independently) of the two other systems. The results of an HL-II analysis can be used as input to the distribution system analysis.

With the introduction of distributed generation, deregulated power markets, and smart grids, the validity of splitting the system into these three levels has become questionable [3]. However, attempting to analyse a large power system as a whole leads in general to excessive computation time, and more importantly, it can be very difficult to come up with a meaningful interpretation of the results [2]. Thus, these three hierarchical levels remain as the most popular method for splitting up the power system in power system reliability assessment.

2.1.3 Reliability Indices

Two categories of reliability indices are of interest - absolute and relative measures. Absolute measures represent the actual (true) value of the reliability indices for a given system. Absolute measures are typically only evaluated based on past performance data. For planning purposes, future values of the reliability indices are of interest. There is in general much uncertainty related to the value of future indices, and typically only the expected value of these indices is evaluated. Predicted indices are thus considered as relative measures, and are typically used for comparison purposes. For instance, the expected gain in reliability level of investing in a new transmission line can be found by predicting future reliability indices for the system with and without the new line. Only predicted indices are considered in this work.

The reliability assessment in this work is concerned with HL-II analysis. A range of indices are defined to describe different parts of the system reliability at this hierarchical level. Typically, both system and delivery point indices are evaluated. See, e.g., [3] for an overview of the different types of indices that can be evaluated in HL-II analysis.

2.1.4 Monte Carlo vs Contingency Enumeration

In probabilistic reliability assessment, reliability indices are calculated by one of two conceptually different techniques - analytical contingency enumeration [2] or Monte Carlo simulation [16]. In the Monte-Carlo approach, the generation and load levels, and component outages, are determined by random sampling. In the analytical contingency enumeration approach, the generation and load levels, and the outage combinations to be analysed, are determined based on some criteria chosen prior to the analysis, and the reliability indices are evaluated based on analytical methods (models).

The concept of Monte Carlo simulations in reliability assessment is discussed in detail in, e.g., [2, 16, 17], while more detailed discussions of analytical contingency enumeration are found in, e.g., [2, 18, 19]. Monte Carlo simulation and contingency enumeration are compared, in terms of a reliability assessment of the Brazilian power system, in [20].

2.2 Security of Electricity Supply Analysis

The integrated methodology for security of electricity supply analysis, as defined in the SAMREL project [5], is shown in Figure 2.2. The SAMREL SoS analysis

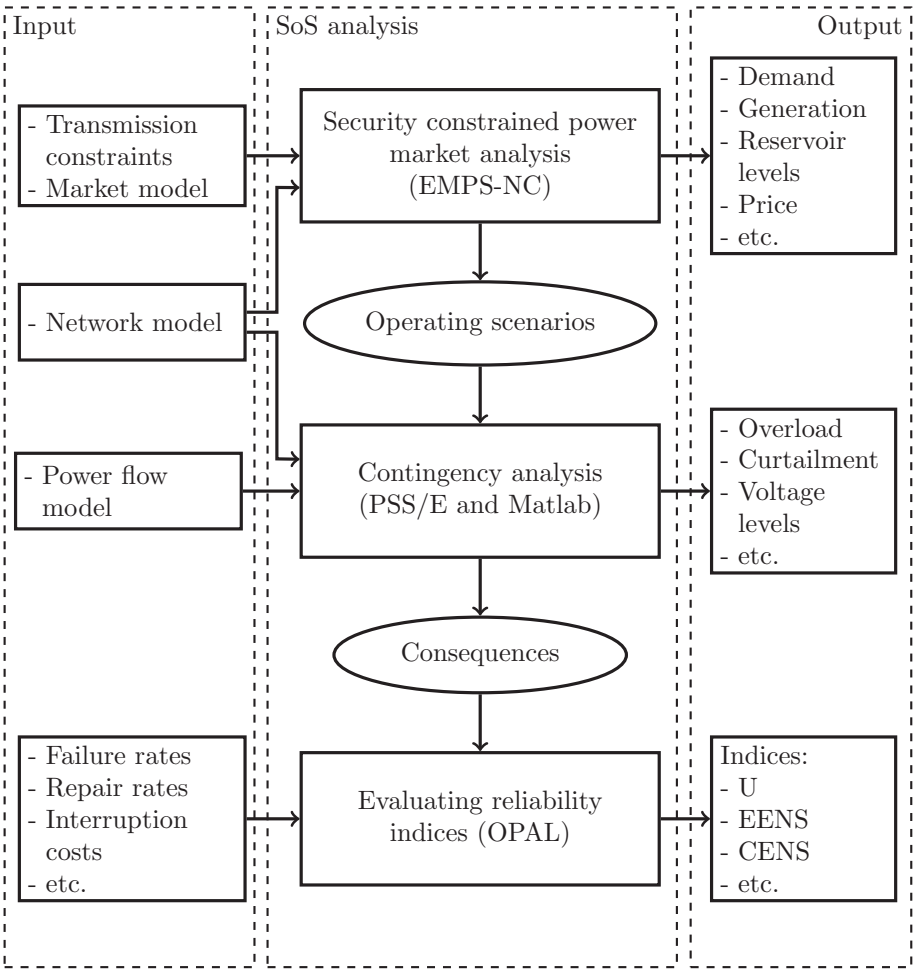


Figure 2.2: Illustration of the general SAMREL SoS analysis framework. Each module requires some input parameters. The results from each module can be analysed separately, or given as input to the next module.

first focus on the reliability of electricity supply at individual delivery points in the power system, and based on these results the system reliability level is determined.

The role of the power market analysis module, in the SAMREL SoS analysis framework, is to generate a set of operating scenarios, for which reliability indices can be evaluated as discussed in the previous chapter. The operating scenarios should represent the generation and load profile for each delivery point in the system (and thus also for the system itself) over the whole year, and represent the stochastic nature of, e.g., hydro inflow and temperature. As mentioned, these operating scenarios are the focal point of this work, and their role as input to the contingency (and reliability) analysis is illustrated in Figure 2.2.

For an operating scenario and a given contingency, the contingency analysis determines whether or not the system can deliver the power required by the customers (at individual supply points), i.e., is there enough generation and transmission capacity in the system to maintain a continuous supply of electricity after the occurrence of this contingency? If not, the amount of necessary load shedding is found. In the reliability analysis, the load shedding is weighted with the probability of failure of components constituting the contingency.

The SAMREL SoS analysis thus takes into account the stochastic nature of both renewable generation (provided by the power market model) and component failures (as failure and repair rates are used for evaluating reliability indices). System reliability indices are derived from the individual delivery point indices.

In the SAMREL SoS analysis framework, reliability indices are evaluated based on the OPAL (Norwegian abbreviation for optimisation of reliability of supply) methodology [9]. OPAL uses analytical models for evaluating reliability indices, based on approximate techniques and minimal-cuts.

In the OPAL methodology, an operating scenario is defined as "... a system state valid for a period of time, characterized by load and generation composition including the electrical topological state (breaker positions etc.) and import/export to neighbouring areas". Reliability indices are evaluated per operating scenarios, where protection system faults can be included in the analysis. To get annual reliability indices, the reliability indices for the set of operating scenarios (which represents a given year) are combined.

Only evaluation of reliability indices used in the case studies in chapter 5 are presented here. A more complete description of the OPAL methodology is found in [9], and more general discussions of the SAMREL SoS analysis framework are found in [4, 5, 9, 21].

2.2.1 Security Constrained Power Market Analysis

EMPS-NC (Network Constraints) [8, 22], an extension of EMPS, is used for the power market analysis. In EMPS-NC, transmission constraints are included in the weekly dispatch optimisation, as a linearised power flow is used to check each scenario to ensure that the dispatch does not result in too high (active) power flow through predefined corridors. These corridors, and the corresponding constraints, should ideally reflect thermal and stability limits in the system. In practice, these corridors and constraints are often defined according to some deterministic criteria, e.g., the $N-1$ criterion. As shown in Figure 2.2, this module takes a network model and transmission constraints as input. This type of power market analysis is referred to as security constrained power market analysis [4, 8].

EMPS-NC models loads per delivery point, and production per generator, in the analysed power system. Thus, the generated operating scenarios are suitable as basis of both generation and transmission system reliability assessment.

2.2.2 Contingency Analysis

The operating scenarios provided by the security constrained power market analysis only hold information about active power flow. If AC power flow is to be used for the contingency analysis, the first step of this module is to establish a solution of the AC power flow problem for the operating scenarios coming from EMPS-NC. If there exists a solution of the AC power flow problem for the given network, one possibility is to, in small steps, adjust the active power injections of the existing solution towards the active power injections of an operating scenario coming from EMPS-NC. After each adjustment, the AC power flow problem is solved. See, e.g., [4, 5] for some more details.

The objective of the contingency analysis is to determine which of the delivery points that experience an interruption for a contingency, where the severity of the interruption is quantified by, e.g., the load shedding at the delivery point. If a contingency event leads to a breach of the system operating criteria (e.g., overload on a component in the transmission system), corrective actions are used to bring the system back within its operating limits. The amount of remedial (corrective) actions required for this operation determines the consequences seen at the delivery points.

The consequences (if any) of such contingencies are found for all delivery points for all operating scenarios. The analysis covers all contingencies in a predefined analysis depth up to second, third, or higher order. In addition, con-

tingencies which are known to cause trouble can be included in the contingency list.

The main result of the contingency analysis is the system available capacity (SAC). SAC is defined per delivery point, and is the available power to supply the load at the delivery point after the occurrence of a contingency [4]. Thus, the SAC depends on the amount of remedial (corrective) actions needed to resolve the system problems (if any) for the given contingency. For operating scenario i (from EMPS-NC), contingency j, and delivery point d, the SAC is denoted $SAC_{i,j,d}$

Among the three modules in Figure 2.2, the contingency analysis is by far the one that takes the most time to complete. The set of operating scenarios that needs to be processed at this step is very large, and the power flow analysis of each operating scenario is in itself quite complex. The time it takes to process one operating scenario depends on, among others, the power flow algorithm (AC power flow or DC power flow), the size of the power system, and the number of contingencies to be analysed (which increases rapidly with increasing analysis depth).

2.2.3 Reliability Analysis

The results (the SACs) of the contingency analysis are taken as input to the reliability analysis, as shown in Figure 2.2. The interrupted power at delivery point d, for operating scenario i and contingency j, is defined in [9, 21], as:

$$P^{\text{inter}}_{i,j,d} = P_{i,d} - SAC_{i,j,d} \quad [\text{MW}], \tag{2.1}$$

where $P_{i,d}$ is the demand at delivery point d (as given by EMPS-NC), and $SAC_{i,j,d}$ is as defined above.

The reliability analysis then determines the minimal cut sets per delivery point, where each cut in the set causes an (partial) interruption at the given delivery point ($P^{\text{inter}}_{i,j,d} > 0$). The interrupted power per cut, together with failure and repair rates, is used to find the expected consequence per cut per delivery point. The expected consequences, for each cut, are accumulated into reliability indices for the delivery point. Each step of this process is described below.

Delivery Point Indices Per Operating Scenario

Reliability indices are calculated for all delivery points. The expected annualised interruption duration, at delivery point d, is:

$$U^a_{i,d} = \sum_{j=1}^{n_{\mathrm{mc},i,d}} r_j \cdot \lambda_j \quad [\mathrm{h}],$$

where λ_j and r_j are the equivalent yearly failure frequency and mean time to repair for minimal cut j, and $n_{\mathrm{mc},i,d}$ is the number of minimal cuts for delivery point d for operating scenario i. The annualised expected energy not supplied (EENS) for delivery point d is:

$$EENS^a_{i,d} = \sum_{j=1}^{n_{\mathrm{mc},i,d}} P^{\mathrm{inter}}_{i,j,d} \cdot r_j \cdot \lambda_j \quad [\mathrm{MWh}], \tag{2.2}$$

where λ_j, r_j, and $n_{\mathrm{mc},i,d}$ are as described above.

System Indices Per Operating Scenario

System indices are also found per operating scenario i. The system annualised EENS is:

$$EENS^a_i = \sum_d EENS^a_{i,d} \quad [\mathrm{MWh}], \tag{2.3}$$

where the sum is over all delivery points, and $EENS^a_{i,d}$ is given by (2.2).

The average interruption duration per delivery point in the system (U^a_i) is defined as the mean of the expected annualised interruption durations of all the delivery points.

Annual Delivery Point Indices

EMPS-NC analyse different (hydrological inflow) years, and annual indices are found for each of those years. It is assumed that the historic data are representative for future scenarios. For delivery point d, the annual indices for year y are:

$$U^a_{y,d} = \sum_{i=1}^{n_{os}} U^a_{i,d} \cdot \frac{h_i}{h_{year}} \quad [\mathrm{h}], \tag{2.4}$$

$$EENS^a_{y,d} = \sum_{i=1}^{n_{os}} EENS^a_{i,d} \cdot \frac{h_i}{h_{year}} \quad [\mathrm{MWh}], \tag{2.5}$$

where n_{os} is the number of operating scenarios in year y, h_i the duration (in hours) of operating scenario i, and $h_{year} = 8736$, the number of hours in an "EMPS-year".

This method neglects that outages of higher order than those of the minimal cuts might have an additional impact on the EENS, i.e., only outages of order equal to that of the minimal cuts are accounted for in the analysis. This is in general not precise, but is a reasonable approximation as long as the higher order outage combinations have low probabilities.

2.3 Security of Electricity Supply Analysis of the Nordic Power System

As the SAMREL SoS analysis is under development, a part of this PhD work was to develop a version of this SoS analysis which could be used for testing and benchmarking of the scenario selection method. The scenario selection method itself is dealt with later in chapter 4. This version focuses on an SoS analysis of the Nordic power system. However, the analysis set-up should be applicable for SoS analysis of other deregulated power systems as well, but small adjustment might be necessary due to, e.g., different market structure.

In the remainder of this thesis, when SoS analysis is used, it refers to the analysis described in the remaining parts of this chapter. The reliability indices in this SoS analysis are evaluated as described in chapter 2.2.3.

2.3.1 The Nordic Power Market

Deregulation of the Nordic power system took place in the 1990's and early 2000's [23], and the leading power market is Nord Pool[1]. The Nordic transmission system is operated by four TSOs - Energinet.dk (Denmark), Fingrid (Finland), Statnett (Norway), and Svenska Kraftnät (Sweden).

In addition to being responsible for the real time operation of the transmission system, the Nordic TSOs define available transfer capacities (ATCs) between market zones [23]. Market zones are defined such that transmission corridors with a high anticipated load connect different zones. In situations where the market clearing for the whole system leads to too high power flow through one or more of these corridors, the market zones are used to split the system into price areas, to reduce the power flow through these corridors. The

[1] www.nordpoolspot.com

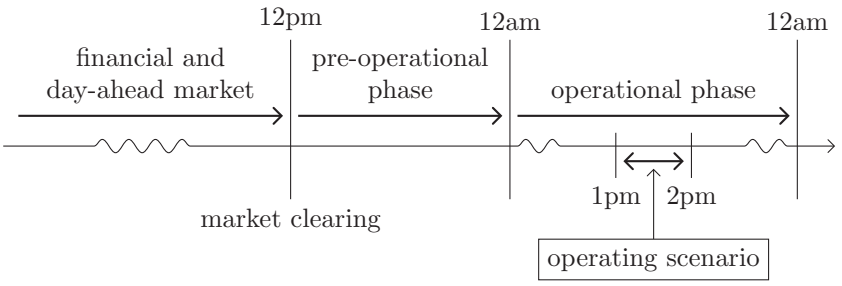


Figure 2.3: The organisation of the Nordic power market. The trading in the financial and day-ahead markets ends with a market clearing at noon (12pm) for each hour of the following day starting at midnight (12am to 12am) [23]. After the market clearing, the TSOs are responsible for trading balancing services to ensure secure system operation. Balancing services are traded in the pre-operational and operational phase. An operating scenario is the system state for a given hour, after the market clearing, but before balancing services are in effect.

TSOs can define ATCs on a daily basis, while market zones are defined for longer time periods.

The organisation of the Nordic power market is illustrated in Figure 2.3. In the financial market, long term contracts are traded, where the main purpose is hedging against price fluctuations. In the day-ahead market, physical power is traded, and at noon the market clearing is done for each hour of the following day according to the supply and demand curves. The price for each hour is determined by the intersection of these two curves.

After the market clearing in the day-ahead market, TSOs trade power in the balancing market to, e.g., resolve congestion problems within market zones or provide spinning reserve. Balancing power is traded in the operational phase in Figure 2.3, where both demand response and reserve generation can be bought. For more details about the Nordic power market, see, e.g., [23, 24].

A special characteristic of the Nordic power system is the high share of large reservoir hydro power plants. Hydro power covers about 95% of the installed capacity in Norway, and about 45% of the installed capacity in Sweden. The rest of the installed capacity consists of nuclear power plants in Sweden and Finland, and thermal generation in Denmark and Finland. In Denmark, 34% of the installed capacity consists of wind power. There are some wind farms in the other countries, with more to be built in the future.

2.3.2 The Operating Scenarios

Depending on the time resolution of the dispatch optimisation (hourly or by load periods), EMPS-NC predicts the system state for that hour or period after the market clearing. This system state has so far been referred to as an operating scenario. These scenarios can equally well be referred to as power market scenarios, as they represent future power market behaviour.

In this work, the EMPS-NC analysis is done considering a constant network topology, thus is the electrical topological state of all the power market scenarios the same. The transmission constraints in the EMPC-NC analysis are in this case related to the corridors connecting market zones and the corresponding ATCs. In this work, the market zones and ATCs are kept constant for the whole analysis period. In the SoS analysis of the Nordic power system, the power flow problem is solved for the whole synchronous area of Eastern Denmark, Finland, Norway, and Sweden. The import/export to Central and Eastern Europe, through HVDC connections, are modelled as negative/positive loads in the power flow problem.

With respect to the SoS analysis framework discussed in this chapter, power market scenarios and operating scenarios are used synonymously. If a different network topology is used for different periods of the year due to, e.g., maintenance schedules, the contingency analysis needs to take this into account. This is easily included in the contingency analysis in the SoS analysis in Figure 2.2 by updating the network model. However, if a non-constant network topology is used, some slight modifications in the application of the scenario selection method are necessary. This is discussed later in 4.8.2.

In the Nordic power system, the ATCs are defined by the TSOs using the $N-1$ criterion, and thus is the probabilistic reliability level unknown. The objective of the SoS analysis is to determine the long-term reliability level (adequacy analysis) of the power system. As the operating scenarios are interpreted as a sample of possible future day-ahead market scenarios, the reliability assessment is concerned with the long-term adequacy assessment of the operational phase in Figure 2.3. In this SoS analysis framework, the reliability indices are first evaluated per operating scenario, and then combined into annual reliability indices as described earlier. Thus, for each operating scenario, this SoS analysis is defined with the objective of checking whether or not the system has enough generation and/or transmission capacity (left after market clearing) to take care of potential system problems due to forced outages of generators, transmission system failures, or within market zone congestion.

2.3.3 Contingency Analysis Models

For each operating scenario, a set of contingencies is analysed with respect to violations of the operating criteria. Three different models are defined and implemented in the SoS analysis framework to find the consequence (if any) of such contingencies.

The Minimal Rescheduling Model

The minimal rescheduling model (MRM), when used in the contingency analysis in the SoS analysis, aims at minimising the consequences, seen by end users at delivery points, of violations of the operating criteria. This includes an analysis of the operating scenario itself, to check for overload on transmission system components.

If a forced outage of a component results in violation of the operating criteria, optimal power flow is used to minimise the cost of the generation rescheduling needed to bring the system back within its operating limits. For linearised power flow, this optimisation problem, which is defined as the minimal rescheduling model, is:

$$\begin{aligned} &\min_{\mathbf{P},\boldsymbol{\theta}_{\text{shift}}} \quad \mathbf{c}_P \cdot \Delta\mathbf{P}, \quad \text{where } \Delta P_k = |P_k - P_k^{\text{sched}}| \qquad (2.6)\\ &\text{subject to } \mathbf{Y}^{\text{DC}}\boldsymbol{\theta} = \mathbf{P}\\ &\qquad\qquad \mathbf{h}_{\min} \le \mathbf{h}(\boldsymbol{\theta}, \mathbf{P}) \le \mathbf{h}_{\max}\\ &\qquad\qquad \boldsymbol{\theta}_{\text{shift,min}} \le \boldsymbol{\theta}_{\text{shift}} \le \boldsymbol{\theta}_{\text{shift,max}}\\ &\qquad\qquad \mathbf{P}_{\min} \le \mathbf{P} \le \mathbf{P}_{\max} \end{aligned}$$

Here, $\mathbf{c}_P$ is a vector of marginal costs of rescheduling per generator included in the optimisation problem, P_k the power output of generator k after rescheduling, P_k^{sched} the scheduled power output of generator k (after market clearing), and $\mathbf{Y}^{\text{DC}}$ is the admittance matrix. The branch flow constraints are defined in $\mathbf{h}(\boldsymbol{\theta}, \mathbf{P})$, and $\mathbf{P}$, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_{\text{shift}}$ are the generator outputs, bus voltage angles, and phase shifting transformer settings.

There is a cost related to both up and downward regulation of the generators, since from the TSOs perspective, this balancing power has to be bought in the balancing market. In (2.6), the marginal cost of upward and downward regulation is the same, which in general is not true. The actual cost of rescheduling is determined by the bids in the balancing market, submitted daily by the generators, and thus it is difficult to estimate this cost in long-term analysis.

However, from an adequacy analysis point of view, the actual cost is not too important, as the analysis mainly is concerned with avoiding load shedding, i.e., avoid interruptions in the electricity supply. Therefore, the cost can, e.g., be set equal to (or higher than) the area price.

If rescheduling is not sufficient to solve the system problems, or if sufficient amounts of demand response cannot be bought in the balancing market, load shedding is included in the objective function given by (2.7).

$$\min_{\mathbf{P},\boldsymbol{\theta}_{\text{shift}}} \mathbf{c}_P \cdot \Delta\mathbf{P} + \mathbf{c}_L \cdot \mathbf{P}^{\text{lshed}} \tag{2.7}$$

Here, $\mathbf{c}_L$ is a vector of marginal cost of load shedding per delivery point with an interruptible load, $\mathbf{P}^{\text{lshed}}$ is the vector of the amount of load shedding done at each of those delivery points. $\Delta\mathbf{P}$, $\mathbf{c}_P$, and the constraints are as above.

The marginal cost of load shedding per delivery point $\mathbf{c}_L$ depends on, e.g., customer type, interruption duration, and time of interruption, see [25–27].

There are two main reasons for using the minimal rescheduling model. First, it minimises the long-term cost of balancing services, and second, it (approximately) minimises the difference between the actual generation scheduling and the (hydro power) scheduling given by EMPS-NC. If a forced outage causes a large deviation from the schedule given by the power market model, this can affect the power market, and thus an update of the power market analysis itself might be necessary. However, the effect of forced outages on the power market can only be analysed if a sequential simulation approach is used, as the time and duration of the forced outages matter, and is not covered by the analysis described here.

If AC power flow is used, it is in principle straight forward to implement this model using AC OPF, as an extension of (2.6) or (2.7). Reactive compensation can also be included in the optimisation problem in this case. However, computationally this is much more demanding.

The Cascading Failure Model

The cascading failure model (CFM) is a simple consequence analysis, which trips the most overloaded component (if any), solves the power flow, and trips the next most overloaded component. This continues until the problem area is isolated, or the whole power system is shut down. This model is only useful in analysis of transmission system outages.

The consequences of most contingencies are quite severe in the CFM, so the model serves as an approximation of worst case scenario.

The Rule Based Model

In real time operation of power systems, a rule based approach is commonly used to resolve system problems. However, this requires specific implementation and knowledge about the power system under study. Thus, this type of consequence analysis is mainly relevant for operational studies. The SoS analysis discussed here is concerned with long-term analysis, thus is this consequence model not used in this work.

2.3.4 Comments

The SoS analysis set-up for the Nordic power system is shown in Figure 2.4, which is a specific version of the general methodology shown in Figure 2.2. Only interruption duration and expected energy not supplied have been discussed in this chapter. The presented analysis framework could be used to estimate the long-term cost of balancing power (as this can be derived from the cost functions in (2.6) and (2.7)), and the cost of energy not supplied [9]. However, some effort should be put into finding sensible (realistic) values to put in the cost vector in the optimisation problem. The scenario selection method can be used in this type of analysis as well.

In the SoS analysis framework, the chronological structure of the operating scenarios is disregarded, as each scenario is analysed independently. However, when combining the results into annual indices, the chronological structure of the operating scenarios is used, and this type of analysis thus distributes the consequences of the forced outages over the whole analysis period. Reference [9] contains more details on how to capture the time dependencies in the analytical OPAL model. The alternative is to use sequential simulations for the reliability assessment, to better capture the time dependence of forced outages.

The load model in EMPS-NC consists of a firm demand and a price sensitive demand. Ideally, load uncertainty should be included in the reliability assessment, but this is not considered in this work.

As a large portion of the generation in the Nordic power system is based on hydro power, there could be situations where there is energy shortage due to very low hydro inflow, or there could be capacity shortage due to, e.g., very low temperature [6]. These problems should ideally be solved by the market itself (by increased prices), but the market might not respond fast enough to sufficiently prevent these problems [6]. In its current version, the SoS analysis is not ready to deal with these problems. One way to include these problems in the SoS analysis, is to update the constraints in (2.6) and (2.7) to take into

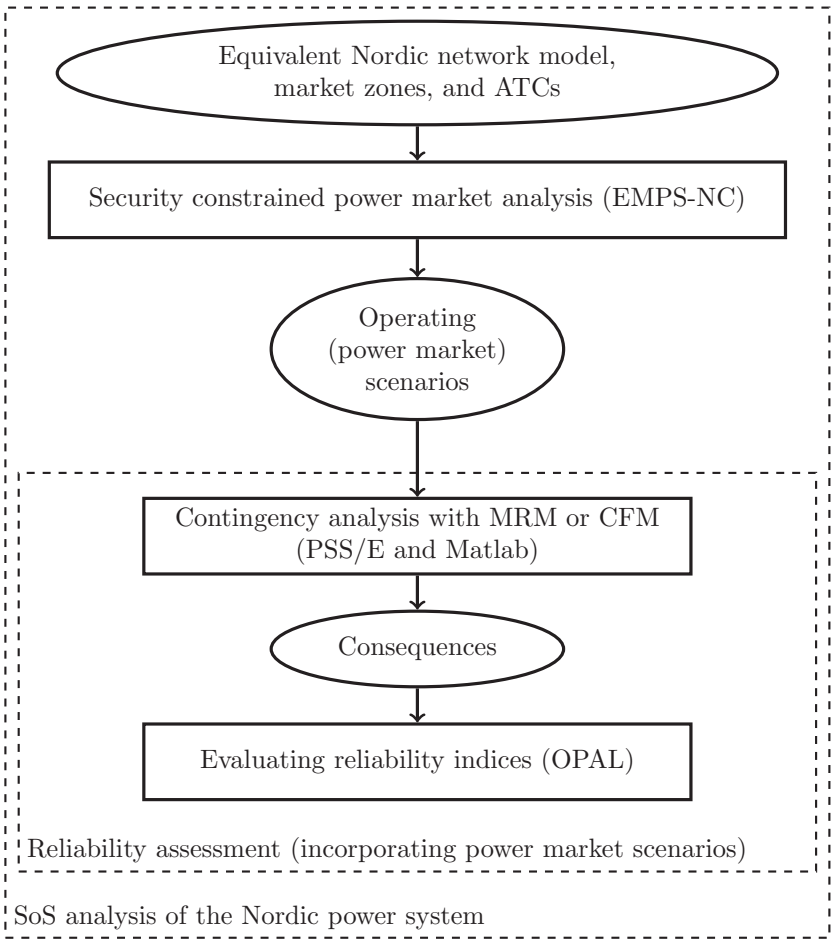


Figure 2.4: Illustration of the SoS analysis used for testing and benchmarking of the scenario selection method. This is a specific version of the general methodology shown in Figure 2.2. Market zones and ATCs are included in the power market analysis, and the operating (power market) scenarios are used as a basis for a reliability assessment.

account, e.g., the reduction in maximum power output of individual generators due to very low reservoir levels. An analysis based on sequential simulations, or the approach taken in [6], can also be used for this purpose.

Chapter 3

Supervised and Unsupervised Learning in Composite Power System Reliability Assessment

In power system reliability assessment, there is a high computational cost associated with calculating the probabilistic reliability indices. The computational costs are affected by factors such as choice of power flow model, the size of the power system, and accuracy requirements. This chapter discusses the background and some of the factors that influence the computational cost, and discusses how statistical learning models can be incorporated in the reliability assessment to reduce the overall computational cost. This chapter is a modified version of publication A.

3.1 Assessment Techniques

A high computational cost is incurred if either analytical or simulation techniques are used to evaluate the probabilistic reliability indices [3, 18], but are due to somewhat different factors.

3.1.1 The Simulation Approach

In the simulation approach, the reliability indices are calculated using non-sequential [2], pseudo-sequential [28, 29], or sequential simulation [16, 17] techniques. Among these, sequential simulation is the method with the highest computational cost, while non-sequential simulation is the computationally cheapest among the three. Here, statistical learning models are discussed in the context of non-sequential simulations - the most commonly used technique among the three. It is possible to incorporate statistical learning models in reliability assessment by sequential and pseudo-sequential simulations as well, but small adjustments might be necessary. This is not further discussed in this thesis.

Non-sequential simulation techniques are based on the state-space algorithm. If a well-being analysis is a part of the reliability assessment, the state-space algorithm follows these mains steps:

1. Generate a system state by random sampling.
2. Determine the status of the system state: success or failure. If a failure state, determine, e.g., the amount of load shedding required to bring the system back within its operating criteria.
3. In well-being analysis: Determine the $N-1$ status of the system state: secure, insecure, or at-risk.
4. Update the reliability indices. If the stopping criterion is not fulfilled, return to step 1.

The main contributors to the computational cost are the number of system states that has to be analysed to fulfil the convergence criterion at step 4, the analysis of the individual system state at step 2, e.g., by AC or DC power flow, and the system size (as increased size drives up the cost of both the analysis at step 2 and 3). For instance, at step 3, to classify a system state as $N-1$ secure or insecure requires up to N power flow evaluations.

In the non-sequential simulation approach, variance reduction [30, 31] and pseudo-sampling [28, 29] can be used to reduce the computational costs by altering the sampling scheme itself; whereas state-space pruning [32] can find the "regions of interest" in the state-space. These methods deal with the evaluation technique directly, whereas learning algorithms can reduce the computational cost by dealing with the systems states themselves.

The most used learning algorithm in the context of reliability assessment is a classifier - a supervised learning algorithm. If the status of each system

state is of interest (success or failure) a classifier can be used to estimate the status of a system state without running a full power flow analysis. Different classifiers have been used in the context of reliability assessment, e.g., neural networks [33,34], Naive Bayes [35], and support vector machines [36,37]. These methods reduce the computational cost related to the analysis of each individual system state at step 2 and/or 3 in the state-space algorithm.

Learning techniques can also be used to limit the number of system states that has to be analysed. This is done by grouping similar system states together, i.e., unsupervised learning, and only analysing one system state per group in step 2 and 3. For power system reliability assessment, this has been done in, e.g., [38,39].

3.1.2 The Analytical Approach

The analytical approach can be based on the state-space algorithm as well. If the more common state-space probability approach is used for the reliability assessment, the factors influencing the overall computational cost have a nearly identical form as the ones discussed below, but usually an operating scenario is denoted a base case. The computational cost of the analytical approach is here introduced and discussed in the context of the SoS analysis described in the previous chapter, where reliability indices are evaluated based on approximate techniques and minimal cuts.

The evaluation of reliability indices follows these main steps:

1. Select an operating scenario.
2. Evaluate reliability indices for the operating scenario, based on an analysis of all (predefined) contingencies.
3. If needed, determine the status of the operating scenario: $N-1$ secure or insecure.
4. Stop if all operating scenarios are analysed, if not, return to step 1.

The computational cost is influenced by the number of operating scenarios (generated by EMPS-NC) that has to be analysed, and the complexity and extent of the contingency analysis done per operating scenario in step 2. The complexity depends, e.g., on the choice of power flow model and the number of contingencies.

A common way of limiting the computational cost, when analytical techniques are used for reliability assessment, is to only analyse the most probable

forced outages (contingencies) [18]. However, to find the most probable forced outage, and define the cut-off criteria, can be very difficult, especially for large systems.

Unsupervised learning can be used to limit the number of operating scenarios that has to be analysed in the consequence analysis, while supervised learning can be used to aid the contingency analysis by limiting the number of power flow solutions required per operating scenario.

The objective of the rest of this chapter is to present the two conceptually different learning techniques - supervised and unsupervised learning; how these can be used in composite reliability assessment, what type of result that can be expected, and when it is appropriate to use these techniques. Learning techniques with origin in statistical and machine learning, artificial intelligence, and bioinformatics are all referred to as statistical learning techniques in this thesis.

The focus is on how to use and incorporate statistical learning models in a reliability assessment based on analytical models. However, it is fairly straight forward to extend the ideas and methods to be used in reliability assessment based on simulation techniques as well. This chapter deals with the problem on a quite general basis, while in chapter 4, an unsupervised learning algorithm specifically suited for the (SAMREL) SoS analysis framework is proposed, elaborated, and discussed in detail.

3.2 Statistical Learning Models

The computational cost of the reliability assessment is dominated by the complexity of the analysis of each individual operating scenario, and the number of operating scenarios that has to be analysed. Statistical learning can be used to either speed up the analysis of each individual operating scenario, or to reduce the number of scenarios that has to be analysed. Two conceptually different learning techniques, supervised and unsupervised learning, can be used for each of these purposes, respectively.

Before discussing the two learning techniques, and how they can be integrated in a reliability assessment, it is necessary to define what statistical learning is. All learning techniques aim at learning the functional relation $f(\cdot)$ between the input $\mathbf{x}$ (here: the operating scenarios) and the response $\mathbf{y}$

$$\mathbf{y} = f(\mathbf{x}),$$

where $\mathbf{x}$ and $\mathbf{y}$ can be vectors. What the response $\mathbf{y}$ is, depends on the objective

of the analysis. It can be a label classifying an operating scenario as $N-1$ secure or insecure, or it can be a group number, where the group consists of similar operating scenarios as $\mathbf{x}$ itself.

Classifiers (supervised learning algorithms) are used to predict the response (the outcome of the reliability assessment) per operating scenario, and thus reduce the complexity of the analysis of each individual operating scenarios. Unsupervised learning algorithms are used to find groups of similar operating scenarios, and thus used as a method for reducing the amount of data taken as input to the reliability assessment.

The main purpose of this section is to describe the different approaches to learning, and how and when these can be applied in a reliability assessment. There are written many books on statistical learning, with [40–42] as excellent examples, and more details can be found there.

3.2.1 The Input-Response Relation

Before deciding which learning technique, and more importantly which specific algorithm, to use, one should consider the structure of the problem. That is to say, what is a likely relation between the input and the response? This relation is of course unknown, but often one has an idea of what it might "look like", e.g., based on knowledge about the physical system that is analysed or based on experience.

Different problem structures are illustrated in Figure 3.1. For a reliability assessment, one can, e.g., think of the blue area as $N-1$ secure, and the orange area as $N-1$ insecure. The goal of a learning algorithm is to describe each area mathematically. Once this mathematical description is established, operating scenarios can be classified according to which area they belong to, and a full $N-1$ analysis of all the individual operating scenarios is no longer necessary.

In Figure 3.1a, it is easy to separate the two groups. The case in Figure 3.1b is similar, but here there is some overlap between the groups. In the grey area, it is hard (impossible) to separate the operating scenarios without doing a full $N-1$ analysis.

In Figure 3.1c and Figure 3.1d, it is much harder to describe the areas mathematically, as one is fully contained in the other. The boundary between the areas in Figure 3.1d is very hard to describe mathematically. Advanced and complex models are needed to describe boundaries like these.

In most practical situations, the shape of the areas is unknown and must be learned from observed (and analysed) operating scenarios. As the shape of the areas in Figure 3.1 gets more complicated, more advanced models are required to

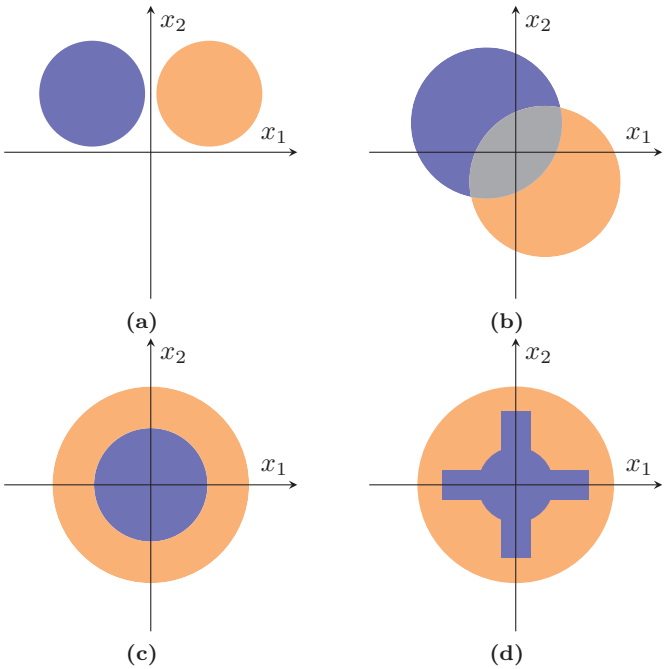


Figure 3.1: Examples of different learning problems and their complexity.

separate the areas. More advanced models require in general more data to train the model, and thus more operating scenarios must be fully analysed before a learning algorithm successfully can be used to classify the operating scenario.

Another very important aspect to consider is variable (feature) selection, i.e., which variables to use as input to the learning algorithm. In Figure 3.1a, only the x_1 axis is useful for separating the groups; thus this axis is "information rich", while x_2 is a redundant feature as it contains no information useful for separating the two groups. In Figure 3.1b-3.1d, both axes are needed to separate the groups.

Identifying the "information rich" features is important for two reasons: being able to separate the groups, and keep the dimension of the input vector low. When the dimension of the input vector increases, it causes problems for the learning algorithms, known as the "curse of dimensionality", see [40]. This "curse" causes classical distance measures to break down, and the feature space

gets sparsely populated. Thus, it gets harder to apply learning algorithms. In practice, it can be very difficult to identify these "information rich" features.

To use statistical learning models in a reliability assessment, a set of features has to be assigned to represent the operating scenarios, and these features should hold information regarding the load, generation, and network configuration. The choice of input features should also reflect what type of analysis is done on each operating scenario; if, say, optimal power flow is used to analyse each operating scenario, the features should contain information regarding how much freedom there is in the control variables in the optimisation problem.

3.2.2 Supervised Learning

Generally, a supervised learning algorithm integrated in a reliability assessment is initiated and applied according to these main steps:

1. Analyse n_t operating scenarios, and determine the response $\mathbf{y}$ for each. Use this as training data.
2. Train a supervised learning model

$$\hat{\mathbf{y}}_\mathbf{i} = f(\mathbf{x}_i),$$

   where $\hat{\mathbf{y}}_\mathbf{i}$ is the predicted response of an input $\mathbf{x}_i$.
3. Predict the response of the remaining operating scenarios $\mathbf{x}_j$, where $j = n_t + 1, n_t + 2, ..., n$

The computational savings are then related to how much faster the statistical learning model can predict the response per operating scenario, compared to a complete analysis of the operating scenario.

In reliability assessment, the response $\mathbf{y}$ is usually a classification label, e.g., success vs. failure state [35, 36] or healthy vs. marginal scenario [33, 34]. If the learning model is unbiased, the classification results are unbiased [40]. However, if, e.g., the load curtailment is calculated for all states classified as failure states, the overall load curtailment is underestimated. This is because all success states, which are misclassified as failure states, are fully evaluated, and thus it is known that no load curtailment is necessary for these states. But the failure states which are misclassified as success states are never fully analysed, resulting in (possibly much) lower overall load curtailment.

Different supervised learning algorithms are suitable for different types of learning problems. Returning to Figure 3.1, classical methods like linear/quadratic

discriminant analysis, logistic regression, naive Bayes, and decision trees, e.g., see [40], can be successfully applied to learning problems of the type shown in Figure 3.1a and Figure 3.1b. For more complex problems, like the ones in Figure 3.1c and Figure 3.1d, more advanced methods are necessary, like the infinite Gaussian mixture model [43–45], neural networks [41], support vector machines [36, 40] and adaptive nearest neighbours [40]. It is important to note that in order to properly train the more advanced methods, more training data is required. Nonparametric methods like k-nearest neighbours (kNN), e.g., see [40], can be very useful for problems where there is no clear parametric structure. To choose the correct learning model for a specific problem is very important. Model selection and evaluation in the context of supervised learning algorithms, see [40], is beyond the scope of this chapter and is not discussed further.

3.2.3 Unsupervised Learning

Generally, an unsupervised learning algorithm is integrated in a reliability assessment, following these main steps:

1. Generate all the operating scenarios.
2. Form k groups of similar operating scenarios.
3. For each group, choose one scenario to represent the group.
4. Evaluate the reliability indices of each chosen scenario. Assume all scenarios within each group to result in the same reliability indices.
5. Estimate the overall/system reliability indices.

The computational savings are then related to how many scenarios that needs to be analysed to get sufficiently accurate reliability indices.

This technique is used in a reliability assessment in [38]. Clustering algorithms are used to form groups at step 2 above, and most such algorithms are constructed to find well separated groups in the data, as in Figure 3.1a and partly in Figure 3.1b. If no well separated groups exist, the problem of finding similar operating scenarios is often denoted as a segmentation problem, as is the case in Figure 3.1c and Figure 3.1d. For segmentation problems, much of the same ideas are used, but there is no point in searching for the natural number of groups, see [46] for further explanation.

The main criticism against unsupervised learning is that there is no objective criteria for verifying the results; only heuristic arguments can be used as opposed

to supervised learning where some model verification is possible as, e.g., cross-validation and bootstrapping [40], can be used for this purpose.

The assumption that all operating scenarios within a group result in the same reliability indices is quite crude. If there is much variation within a group, this can lead to poor results.

Two of the most popular clustering methods are agglomerative clustering and the k-means algorithm, e.g., see [40]. Examples of more advanced unsupervised learning methods are the infinite mixture models, [47], spectral clustering [48], the self-organising map [49], and support vector clustering [50].

3.2.4 A Comparison of The Two Learning Methods

The two learning methods require different type of input data. The supervised approach requires training data where the response must be known. Thus, if n_t operating scenarios are used as training data, at least n_t scenarios has to be fully analysed.

The unsupervised approach is most efficient when it can take all operating scenarios as input data. Thus, all operating scenarios should be generated before the reliability assessment. Only the chosen scenario from each cluster has to be fully analysed, thus in total k operating scenarios are fully analysed. This approach can reduce the computational requirements more than the supervised approach if a small value of k is chosen. However, choosing a (too) small k can lead to large errors in the reliability indices.

A supervised learning algorithm can be implemented on-line, as it can be trained simultaneously as operating scenarios are analysed. When the analysis of n_t operating scenarios is completed, the algorithm can replace the full analysis of each individual operating scenario.

Supervised learning requires some training data as input. If the goal is to train a classifier, there should be a "sufficient" number of operating scenarios from each class in the training data. For instance, if there is a very low proportion of failure states, many states have to be analysed to get a "sufficient" number of failure states in the training set. If there are many failure states, the computational savings related to not having to analyse the secure states might be small. The unsupervised learning approach does not have this problem.

3.2.5 High Dimensional Problems

For "real" size power networks, each operating scenario is described by a high dimensional vector. Thus, the "curse of dimensionality" is a problem. Methods

like support vector machines in combination with kernel principal components [40] can be used for such problems. In general, it gets harder to use learning techniques as the dimension of the problem increases.

Even though it is difficult to apply learning techniques on high dimensional problems, and some care should be taken, these methods are still very useful, as the contingency analysis of each operating scenario has a very high computational cost for large networks.

Chapter 4

The Scenario Selection Method

This chapter propose, elaborate, and discuss the scenario selection method. The scenario selection method is used to reduce the overall computation time of the (SAMREL) SoS analysis, which was the main goal of this PhD project. Thus, this chapter is the core of this thesis.

First, the general idea of the method is outlined. It is followed by detailed discussions of the different aspects and problems related to each step of the scenario selection process. This chapter is directly related to publication B.

4.1 General Idea

The proposed scenario selection method is designed to reduce the number of operating scenarios (generated by EMPC-NC) that has to be analysed in the contingency analysis module in the SAMREL SoS analysis framework, illustrated in Figure 2.2. The method is built around the idea of how to use unsupervised learning models in a reliability assessment, which was discussed in the previous chapter.

The scenario selection method follows these main steps:

1. Find k groups of similar operating scenarios, among the n operating scenarios generated by the security constrained power market analysis.

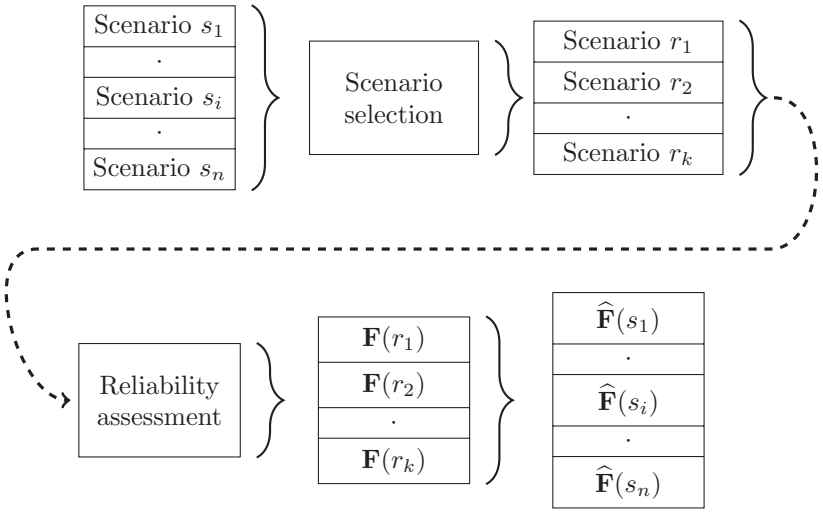


Figure 4.1: Illustration of the scenario selection process. The n scenarios, generated by EMPS-NC, are represented by k scenarios, which form the representative set. For each scenario in the representative set, at set of reliability indices $\mathbf{F}(r_i)$ are calculated, as described in chapter 2.2. As it is assumed that all scenarios within a cluster/group have the same value of the reliability indices, this is used to estimate the value of reliability indices $\hat{\mathbf{F}}(s_i)$ for each of the n original scenarios.

2. For each group: Represent the group characteristics by the group centroid (the operating scenario closest to the group centre). The chosen scenarios form what is denoted the representative set.

3. Evaluate the reliability indices for the k chosen operating scenarios (picked to represent each group).

4. Assume that all operating scenarios within a group have the same value of the reliability indices as the group representative (found in the previous step). Use this to estimate the reliability indices for all the n initial operating scenarios.

This process is illustrated in Figure 4.1. As only the group centroids are fully analysed, the computational cost of the contingency analysis is substantially reduced, while most of the detailed analysis results provided by the power market model are kept intact.

If the probabilities of the different operating scenarios are of interest, as in, e.g., risk analysis, the scenario selection method can keep the empirical probability distribution of the operating scenarios approximately intact. For instance, assuming all operating scenario initially have equal probability $(1/n)$, and group k in the representative set has n_k members, the operating scenario representing group k is assigned a probability of n_k/n.

The scenario selection method has much in common with the scenario reduction method discussed in [51], and the system state reduction technique presented in [38]. A different approach of data (scenario) reduction is adaptive sampling [52], but that approach is mainly useful for very large data sets. Although motivated by reducing the number of operating scenarios, the method and problems discussed in this chapter are quite general, and similar data (scenario) reduction approaches are found within other research fields.

It should be noted that the scenario selection method does not reduce the time it takes to analyse one operating scenario, as the per operating scenario analysis in the contingency analysis module (in Figure 2.2 and Figure 2.4) is the same as before. The computational savings are only related to the reduction of the number of scenarios that needs to be analysed. If the time it takes to analyse each operating scenario is too high, supervised learning algorithms can be used as discussed in the previous chapter.

4.2 The Representative Set: A Qualitative Discussion

The representative set should be an adequate base for the contingency analysis and the evaluation of reliability indices, i.e., it should capture the main/important characteristics of the full set of operating scenarios. Furthermore, the representative set should be substantially smaller than the full set to limit the computational cost of the reliability assessment. These two requirements are contradictory, as a smaller set in general leads to less accuracy in the overall analysis as less data is used as decision support. Thus, an appropriate trade-off between accuracy and data reduction is needed. There are some important issues to consider before the appropriate "trade-off criteria" can be determined, e.g., how much loss of accuracy can be accepted at the gain of less computational effort required to complete the overall analysis? The following discussions deal with some of these issues.

Clustering algorithms essentially search for natural groups within the dataset,

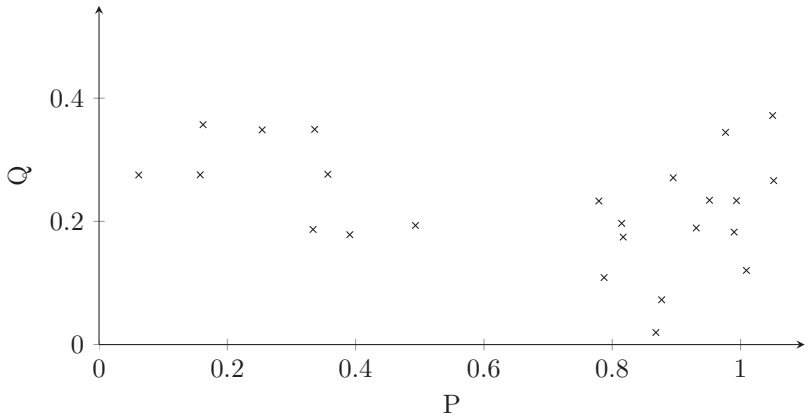


Figure 4.2: Example of natural grouping of data. The P and Q represent hypothetical power production for a power plant. The two groups can be easily separated by, e.g., placing a vertical line at $P = 0.6$. Note that only the P axis is useful in terms of separating the two groups.

meaning that observations in one group should be more similar to each other compared to those in other groups. For instance, it is evident that the hypothetical power generator data plotted in Figure 4.2 can be divided into two groups by drawing a vertical line at $P = 0.6$. The existence of such natural groups within the data, and the number of such groups, is generally unknown. If a natural grouping exists, clustering algorithms can be successful in finding these groups. Different clustering methods are presented and discussed in detail later in this chapter. If there is no such natural grouping within the data, the whole dataset should be interpreted as one (uniform) group of similar operating scenarios (at least when looked at a clustering perspective). If this is the case, the problem of finding a representative set is quite different, as it turns into a segmentation problem [40].

Seen from another perspective, if there is a fixed (and predetermined) number of operating scenarios one can afford to include in the representative set (in terms of the computational complexity of the overall analysis), the scenario selection problem turns into a segmentation problem. This can for instance be the case if one is willing to analyse, say, only 50 operating scenarios.

There is also a problem related to how to sufficiently represent the operating scenarios within one cluster. If the cluster is compact, it should be sufficient to

represent all the operating scenarios within the cluster by the centroid. On the other hand, if the cluster is quite large (where large means high volume), with a high within cluster variation, more scenarios might be needed to sufficiently represent the cluster characteristics. Thus, the amount of averaging done, within each cluster, depends on how much accuracy one is prepared to lose at the gain of more data reduction.

The type and characteristics of the analysed power system influence the number of groups needed in a representative set. In a power system dominated by thermal energy production, the future generation patterns are quite predictable, and possibly a few groups are enough to represent the main characteristics of the full set of operating scenarios. While in systems where generation is dominated by intermittent generation, e.g., wind and hydro power, the future generation patterns vary a lot, and more groups are (probably) needed in a representative set to fully capture the important characteristics of the full set.

To actually have a representative picture of the future load on the power system, the extreme states/outliers are important, as they might represent special cases where the system is under considerable strain. A forced outage, for this load and generation scenario, might have severe consequences. This is usually referred to as "high impact low probability" (HILP) events in the context of reliability analysis. Most clustering algorithms cannot handle outliers. If an outlier is included in a cluster, the characteristics of the cluster, e.g., the mean, is severely affected, and this distorts the clustering procedure. Thus, it is important to take care of outliers in a suitable manner, before searching for groups of similar operating scenarios.

Much of the success of the scenario selection method depends on how much it can reduce the computational cost of the reliability assessment. At the same time, the framework should find the representative set in an efficient matter. That is to say, the representative set should substantially reduce the number of operating scenarios that has to be analysed, and at the same time, the scenario selection method should quickly find the representative set. If the scenario selection method uses much time to find the representative set, little is gained, with respect to overall computation time, by doing the data reduction. Of course, the time one can afford to spend on finding the representative set depends on how computationally intensive the reliability assessment is, e.g., if first, second, or higher order contingencies are analysed.

An additional information gain, when applying the scenario selection method, is that it can find patterns and structures among the operating scenarios. This might be useful for other analyses, where, e.g., seasonal trends are of interest.

4.3 Feature Selection

For clustering algorithms to be able to find groups of similar operating scenarios, a set of features (variables) must be assigned to represent each scenario, such that similarity measures can be used to quantify similarity. Mathematically, each operating scenario s_i is represented by a data vector $\mathbf{x}_i$

$$s_i = \mathbf{x}_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,p}], \tag{4.1}$$

where $x_{i,j}$ is a feature (data point), and p is the dimension of the feature vector/set. For instance, the vector $\mathbf{x}_i$ can be the loads at all delivery points in the analysed power system.

Each feature $x_{i,j}$ should provide relevant information regarding similarity/dissimilarity of the operating scenarios. At the same time, the dimension p should be kept as low as possible to avoid the "curse of dimensionality", see, e.g., [40].

In the context of scenario selection within the SAMREL SoS analysis framework, the question is what makes operating scenarios similar. Is it the output of the EMPS-NC model? How each operating scenario is analysed in the contingency analysis? Or a combination of the two? As feature selection is in fact a case dependent process, these questions are best answered and discussed based on results of case studies. Thus, the feature selection problem, within the scenario selection method framework, is discussed further in chapter 5.3.

On a more general note, it is known that some features are useful for separation among objects, while others are not. The hypothetical example in Figure 4.2 illustrates this. The data in the figure are simulated data for a power generator, where P and Q are in pu. It is clear that there are two natural groups here - high and low production. But only the P value is useful for separating the two production categories.

As the aim of the scenario selection method is to find groups of operating scenarios with similar (value of) reliability indices, it is important that the features provide information about the power system problems encountered in the reliability assessment. For instance, total system load is an important feature in reliability assessment of generation deficient systems.

4.4 Finding Groups of Similar Operating Scenarios

After a set of features is chosen to represent each operating scenario, and linearly dependent and highly correlated features are removed, the process of finding similar operating scenarios can start. To find similar scenarios, a similarity measure and a clustering algorithm must be chosen. Some clustering algorithms can use different similarity measures, while others are built around a specific similarity measure.

4.4.1 Similarity Measures

All clustering algorithms need a measure of similarity/dissimilarity to quantify similarity between operating scenarios. Only quantitative features are considered in this work.

Distance Measures

The most common similarity measure is the distance metric:

$$d(x_{i,j}, x_{i',j}) = ||x_{i,j} - x_{i',j}||^l, \tag{4.2}$$

where $x_{i,j}$ and $x_{i',j}$ are the values of feature j of operating scenario i and i' respectively.

The total dissimilarity between two operating scenarios is the weighted sum of all the pairwise dissimilarities between the features of the two operating scenarios, with indices i and i',

$$D(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} w_j \cdot d(x_{i,j}, x_{i',j}), \quad \sum_{j=1}^{p} w_j = 1. \tag{4.3}$$

Here, w_j is the weight of feature j, and $d(x_{i,j}, x_{i',j})$ is as defined in (4.2). As l increases, more weight is put on large distances between the features. A more thorough discussion of distance measures is found in [40].

Likelihood

Another method for quantifying similarity is the likelihood of a data set $\mathbf{X}$. The likelihood is defined in (4.4).

$$L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} g_{\boldsymbol{\theta}}(\mathbf{x}_i) \tag{4.4}$$

Here, $g_{\boldsymbol{\theta}}(\cdot)$ is a parametric model family, with parameters given by the vector $\boldsymbol{\theta}$. As previously, $\mathbf{x}_i$ is a p dimensional vector. In most practical applications, the log-likelihood is used, as defined in (4.5).

$$l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{i=1}^{n} l(\boldsymbol{\theta}; \mathbf{x}_i) = \sum_{i=1}^{n} \log g_{\boldsymbol{\theta}}(\mathbf{x}_i) \tag{4.5}$$

The log-likelihood is often referred to as cross-entropy or deviance.

To find similar operating scenarios, first a parametric model family $g_{\boldsymbol{\theta}}(\cdot)$ is chosen, and then the parameter vector $\boldsymbol{\theta}$ which maximises the likelihood of the sample of operating scenarios is found. Similarity of operating scenarios is then determined by how well the model fits the data, and the model itself helps determine the number of groups of similar operating scenarios. Different models and parameter vectors are compared with respect to the likelihood.

4.4.2 Clustering Algorithms

There is a great variety of clustering algorithms which can be used to find groups of similar operating scenarios. Two criteria for distinguishing the main types of clustering algorithms are useful in this context; hierarchical methods vs. partitioning methods and hard vs. fuzzy clustering. Partitioning methods aim at finding one partition for the data, while hierarchical methods aim at fitting the data into a hierarchical structure. Hard clustering assign each object to one cluster, while fuzzy clustering can let one object belong to more than one cluster, where the former is more suitable for the scenario selection method.

Below are k-means, agglomerative clustering, self-organizing map clustering, and the infinite Gaussian mixture model briefly discussed and summarised. For more thorough discussions, and introduction to the vast amount of different clustering techniques, see, e.g., [40, 50, 53].

Agglomerative Clustering

Hierarchical clustering is divided into agglomerative and divisive clustering. Divisive clustering starts with all objects in one cluster, and at each level it chooses an optimal split of the clusters. This approach is way too computationally intensive to be useful for scenario selection, except if one search for a very small number of clusters. Only agglomerative clustering is discussed here.

Agglomerative clustering algorithms require a dissimilarity matrix as input. The dissimilarity matrix is a symmetrical matrix where all the pairwise dissimilarities between objects (the operating scenarios) are calculated. Agglomerative clustering starts with all objects in separate clusters, and starts merging clusters together based upon the dissimilarity matrix and the linkage criterion. The most common linkage types are:

- Single linkage: The two clusters with the minimum minimal distance between them are merged.
- Complete linkage: The two clusters with minimum maximal distance between them are merged.
- Average linkage: The clusters with the minimum distance between their cluster centres are merged.

The algorithm yields a dendrogram, and cutting the dendrogram at a certain height forms clusters. The dendrogram in Figure 4.3 represent the agglomerative clustering of the data from Figure 4.2. To get the two natural clusters, cut the dendrogram at the height represented by the horizontal line “Cut 1”. “Cut 2” yields four clusters.

Single linkage often results in chaining, i.e., one large cluster is formed. Agglomerative clustering with complete linkage usually produces small and tight clusters, and is in practice the most used linkage criteria. The average linkage requires more computations. See [40] for a more details.

K-Means

This method requires the number of clusters k as input. It then assigns objects (operating scenarios) to k partitions, such that the “between cluster” variation is maximised with respect to the “within cluster” variation. This algorithm uses the Euclidean distance (setting $l = 2$ in (4.2)) as dissimilarity measure.

The different steps of the k-means clustering process are:

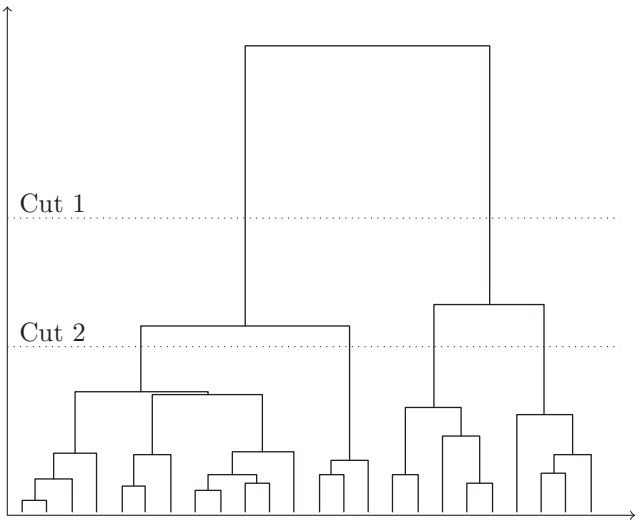


Figure 4.3: The dendrogram build by an agglomerative clustering of the data in Figure 4.2. To cut the dendrogram at the height of the horizontal line cut 1 results in a partition of the data into the two natural groups seen in Figure 4.2.

- Step 1: Choose k and partition the data into k initial clusters, e.g., by random sampling.
- Step 2: Calculate the cluster centres.
- Step 3: Reassign objects to the cluster with the nearest cluster centre.
- Step 4: Repeat step 2 and 3 until no change occur between iterations.

It is also possible to add a fifth step, which takes action when empty clusters appear and/or takes care of outliers.

Essentially, this algorithm partitions the data into hyperspherical clusters, and it is known to be very successful if well separated and compact clusters exits within the data [53]. The algorithm is also quite efficient as it works directly with the data.

In Figure 4.4, the k-means algorithm is used to find clusters among the data in Figure 4.2. The solid dark ellipses indicate the result when setting $k = 2$, and the red dashed ellipses the result of setting $k = 3$.

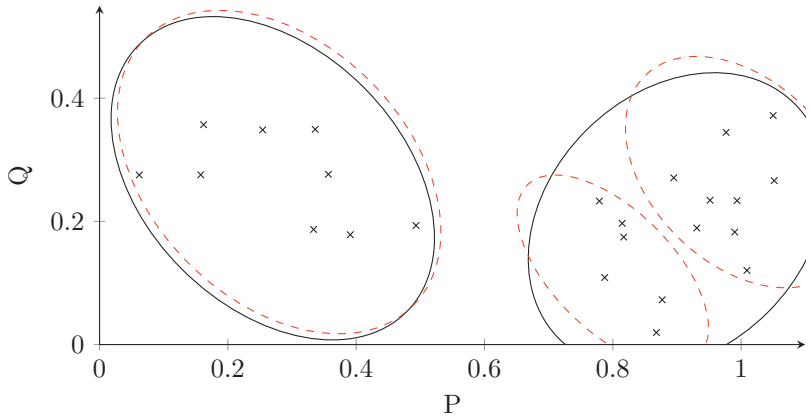


Figure 4.4: Example of application of the k-means algorithm on the data in Figure 4.2. If $k = 2$, the clustering solution is indicated by the black ellipses, while setting $k = 3$ gives the clustering solution shown as dashed red ellipses.

The major drawback of this method is that it solely depend on the user to choose k, and it is sensitive to the initial partition. That is to say, if a bad initial partition is chosen, there is no guarantee that the algorithm converges to the global optimum. However, convergence is guaranteed. One way to partly solve this problem is to start the algorithm with different initial partitions, and see if it converges to the same optimum. If no global optimum is reached, the solution with maximised "between cluster" variation is chosen. The problem of choosing k is addressed later in this chapter.

Self-Organising Map Clustering

The self-organising map (SOM) algorithm was originally designed for projecting high dimensional data onto a two dimensional discrete map, but has later been used as a clustering algorithms as well [49]. Several versions of this algorithm exists as, e.g., both on-line and batch training, and different map structures, can be used. Only the on-line method, which maps high dimensional objects (here the operating scenarios) $\mathbf{x}_i$ onto a two dimensional rectangular discrete map, is discussed here.

The mapping of the operating scenarios onto a two dimensional map is illustrated in Figure 4.5. The two dimensional map $\mathbb{Q}_{q_1,q_2}$ has $k = q_1 \cdot q_2$ nodes, and to each node $j \in \mathbb{Q}_{q_1,q_2}$ there is a prototype $\mathbf{m}_j$ that maps each $x_i \in \mathbb{R}^p$

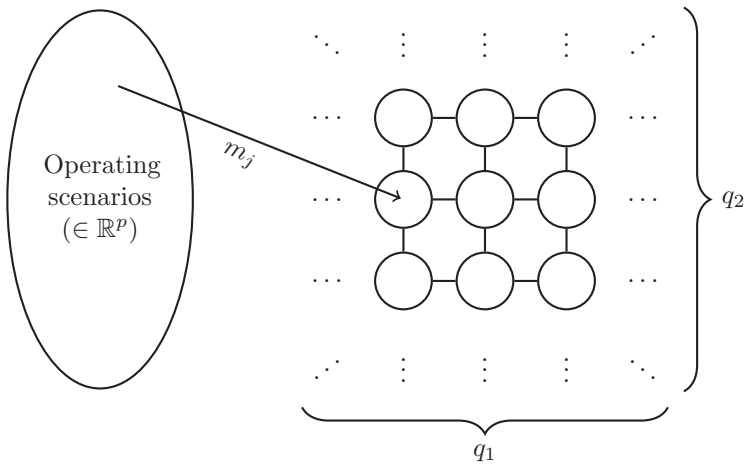


Figure 4.5: Illustration of the SOM process. The prototypes m_j's map the operating scenarios (in the high dimensional space $\mathbb{R}^p$) onto the two dimensional discrete map.

onto $\mathbb{Q}_{q_1,q_2}$. Note that the map $\mathbb{Q}_{q_1,q_2}$ is considered as a discrete structure.

As a clustering algorithm, first an initial prototype is assigned to each node. Then, the prototypes are updated (trained) as given in (4.6).

$$\mathbf{m}_j \leftarrow \mathbf{m}_j + \alpha(\mathbf{x}_i - \mathbf{m}_j) \tag{4.6}$$

This update causes the prototype $\mathbf{m}_j$ to be moved closer to $\mathbf{x}_i$, where α is used to scale the impact on $\mathbf{m}_j$ of this update. For a more thorough discussion of SOM, see [40, 49]. When SOM is used as clustering algorithm in the scenario selection method, it is implemented as method B in [38]. After training, some nodes might be empty. If so, the number of clusters produced by SOM is less than k (the size of the map).

The Infinite Gaussian Mixture Model

The infinite Gaussian mixture model (IGMM) is an advanced clustering algorithm, based on maximising the likelihood of the data sample by using a Gaussian mixture model with a variable number of components [43]. The model is a part of the nonparametric Bayesian domain, e.g., see [54], and a more thorough discussion of the IGMM model can be found in [44].

IGMM was first introduced in [43], which is a detailed discussion of the univariate model. The multivariate extension of the model is briefly discussed here. The full implementation and mathematical derivation of the multivariate IGMM, for scenario selection, is included in appendix A.

The IGMM can be seen as an extension of the finite mixture model, given in (4.7), as k need not be specified, but is rather learned from the data. This is one of the major advantages of this model, compared to the previously described clustering algorithms.

$$p(\mathbf{x}_i|\boldsymbol{\mu}_1, .., \boldsymbol{\mu}_k, \mathbf{S}_1, .., \mathbf{S}_k, \pi_1, .., \pi_k) = \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{S}_j^{-1}), \tag{4.7}$$

where $\pi_j, \boldsymbol{\mu}_j, \mathbf{S}_j$ are the weight, mean, and precision (inverse of the variance) of component j respectively.

The disadvantages of this approach are that the model itself is quite complex, and sampling schemes are needed in order to do draw inference about the parameters. The MCMC scheme, used with this model, is described in appendix A. During the sampling process, (independent) samples of the joint density function, over the operating scenarios, are drawn. Each sample consists of a mixture of a number of multivariate Gaussian components. The number of such components is the number of clusters for the given sample.

The number of components varies between the samples, but after a burn-in period, the number usually lies within a limited interval. This interval indicates how many groups/clusters there is in the sample of operating scenarios. For instance, going back to the data in Figure 4.2, the IGMM might "jump" between the two clustering solutions, with 2 and 3 clusters, in Figure 4.4 during the sampling process. In addition, as several samples are drawn, it is possible to estimate the pairwise probability that two operating scenarios belong to the same cluster.

As the number of clusters varies, the operating scenarios belong to different clusters at different stages throughout the sampling process. For scenario selection, one clustering solution is needed. Based on the sampling process, the solution with the highest posterior likelihood is chosen.

Other Clustering Algorithms

Above are four different clustering algorithms introduced. However, there exist a great variety of other clustering algorithms. Spectral clustering [48] finds clusters by representing data similarity by a similarity graph, and the clusters are

formed by partitioning this graph. Support vector clustering [50], is a method based on advanced techniques from the very popular class of classification algorithms known as support vector machines. For more details and discussions of the vast amount of clustering algorithms, see, e.g., [40, 50, 53].

4.4.3 A Discussion of the Clustering Algorithms

The k-means is quite efficient as it works directly on the data, while building the dendrogram is quite time consuming. On the other hand, the agglomerative algorithm is not dependent on the initial partition, and can handle different dissimilarity measures, as long as they can form a dissimilarity matrix. Cutting the tree/dendrogram at different heights, and by that forming clusters, can be done quite efficiently. If many values of k are to be tested, as is the case when estimating k, this can be more efficient than k-means, as k-means starts from scratch (by generating a new initial partition) for each value of k. The on-line training of SOM is quite efficient, but as for the k-means algorithm, the method starts from scratch for each new map size k. The sampling process of the IGMM is very complex, and thus is this algorithm much slower than the others.

Hierarchical clustering, k-means, SOM, and IGMM all, explicitly or implicitly, put some structure on the data. The k-means tends to form hyperspherical clusters, while this might not (always) be a suitable structure. K-means also tend to produce clusters of approximately equal size. This is not ideal in the context of scenario selection, as ideally, the goal is to search for large clusters of very similar operating scenarios, and let the scenarios with high consequences form their own small clusters. SOM can be seen as a constrained version of the k-means algorithm [40], and often produce similar results as the k-means algorithm.

The agglomerative approach often produces quite different trees (dendrograms) for different linkage criteria. If so, this suggest that this type of structure/hierarchy is not ideal for the data set at hand. If the different linkage criteria produce approximately the same tree, it can indicate that the data fit well into a hierarchy. Agglomerative clustering with complete linkage usually produces compact clusters, which fit well with the objective of the scenario selection method.

The IGMM uses a mixture of Gaussian distributions to search for clusters among the data. If in fact the data are generated from a process that resembles a mixture of Gaussians, this algorithm can work very well. In fact, a mixture of Gaussian distributions can be used to approximate many other distributions as well, thus is this approach suitable for many types of data structures. How-

ever, this approach models the joint density function of the data. For high dimensional data, this is an extremely complex process which requires very long sampling time. There can also be convergence issues during the sampling period.

A quantitative comparison of the clustering algorithms is given later in chapter 5.4. The quantitative comparison is done with respect to how well these algorithms perform when used for scenario selection in the SoS analysis framework in Figure 2.4.

4.5 How Many Clusters?

All the algorithms discussed so far, except IGMM, require that the number of clusters k is specified before a clustering solution is found. This section deals with how the value of k can be estimated, and other methods for verifying the clustering solution (once found).

4.5.1 Estimating k - The Number of Groups

If the goal of the clustering problem is to find say 10 groups, the k-means algorithm seems like a good choice. One could consider running the k-means with some different initial partitions to verify global convergence. On the other hand, if one wants to estimate k from the data, the problem is more complicated.

A common way of estimating the number of clusters k, is through the "within cluster dispersion" W_k. The following definition is taken from [46].

$$D_r = \sum_{i,i' \in C_r} d_{ii'}, \tag{4.8}$$

where C_r is the index of the objects assigned to cluster r, and $d_{ii'}$ is the squared Euclidean distance (which follows from setting $l = 2$ in (4.2)). The "within cluster dispersion" (W_k) is then

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r, \tag{4.9}$$

where $n_r = |C_r|$, the number of objects in cluster r.

As the number of clusters is increased, W_k decreases. If we denote K as the "true" number of clusters, then W_k decreases quickly as k approaches K from below. Objects that do not belong to the same cluster are assigned to different clusters, i.e., clusters containing "very" dissimilar objects are split up. As k

reaches K the decrease in W_k declines, as all objects at this stage are associated with only "natural" neighbours - the curve of W_k markedly flatten out. As it is said in [46] "Statistical folklore has it that the location of such an 'elbow' indicates the appropriate number of clusters". The presence of such an "elbow" however, is often hard (impossible) to locate.

If this approach is used on a data set with a high number of clusters, this is even more complicated as a clustering solution has to be found for each value of k, and then calculate W_k, which gets very computationally intensive. This is not in accordance with the efficiency criterion for the scenario selection method.

There are two major problems with the W_k approach of searching for k. As mentioned, the location of the elbow is difficult to find, and the approach fails in the situation where there is no natural clustering, i.e., $k = 1$. This is for instance the case in a segmentation problem, where there is no natural grouping among the data (operating scenarios). This is some of the motivation behind the "gap statistic" [46], which is an interesting approach to estimating k. However, the method is most suitable for small values of k.

4.5.2 Cluster Verification

Clustering, or unsupervised learning, is essentially learning without a teacher. The goal is to form clusters with similar objects in them, but a general verification of the clustering is not available. This is not the case in supervised learning (e.g., regression models) where, e.g., cross-validation or bootstrapping can be used to estimate the prediction error.

This type of assessment is not available for clustering, and one solely rely on heuristic arguments like W_k and the "tip of the elbow", or the maximized between cluster variation of the k-means. These are logical arguments, partly verified by experience, but without the possibility of any objective assessment.

These heuristics arguments, being the basis for clustering, are one of the reasons there exists such a vast amount of clustering algorithms. As the measure of success is a subjective matter, optimal clustering depends on the user. Additionally, clustering algorithms implicitly put some structure on the data. The true data structure usually deviates from this theoretical structure, and therefore some clustering algorithms are more efficient for some data sets than others. There is no general optimal clustering procedure.

4.6 Estimating the Reliability Indices

For each group of operating scenarios, the medoid is chosen to represent the whole group. The medoid is the operating scenario closest to the centre of the group. The reliability indices of the medoid operating scenario are evaluated as explained in chapter 2.2, and it is assumed that all operating scenarios within each group have the same value of the reliability indices as the medoid.

Within each group, the reliability indices of each operating scenario is set equal to the group representative. The resulting (per operating scenario) indices are marked with a hat as they are only approximations of the reliability indices. For instance, the annual expected interruption duration is denoted $\widehat{U}^a_{y,d}$. To get the annual indices per year y:

$$\widetilde{U}^a_{y,d} = \sum_{i=1}^{n_{os}} \widehat{U_{i,d}} \cdot \frac{h_i}{h_{year}} \quad \text{[h]}, \tag{4.10}$$

$$\widetilde{EENS}^a_{y,d} = \sum_{i=1}^{n_{os}} \widehat{EENS}_{i,d} \cdot \frac{h_i}{h_{year}} \quad \text{[MWh]} \tag{4.11}$$

This is equivalent to what is done in (2.4) and (2.5). These indices are denoted the scenario selection indices. Indices per operating scenario are marked with a hat, while tilde marks annual indices (per modelled hydro inflow year).

4.7 Post Analysis Cluster Verification

The ultimate goal of the scenario selection method is to estimate the annual reliability indices as accurately as possible by analysing as few operating scenarios as possible. That is to say, the scenario selection indices

$$\widetilde{U}^a_{y,d}, \widetilde{U}^a_y, \widetilde{EENS}^a_{y,d} \text{ and } \widetilde{EENS}^a_y$$

should be as close to the the reliability indices resulting from an analysis of all the operating scenarios

$$U^a_{y,d}, U^a_y, EENS^a_{y,d} \text{ and } EENS^a_y,$$

as possible. The indices based on a full analysis are referred to as target values (TVs) or target indices.

For practical applications of the method, a comparison with the target values is not possible, as these never are evaluated. However, in the development of

the scenario selection method, and for benchmarking purposes in this work, the scenario selection indices are compared with the target values.

In this context, the absolute error and the relative error are commonly used. For an annual reliability index F, at delivery point d, the absolute error and the relative error are defined in (4.12) and (4.13) respectively.

$$\text{absolute error} := |\widetilde{F}^a_{y,d} - F^a_{y,d}| \tag{4.12}$$

$$\text{relative error} := \frac{\widetilde{F}^a_{y,d}}{F^a_{y,d}} \tag{4.13}$$

The same errors can be calculated for annual system indices, and the per operating scenario indices.

In publication B and C, it is discussed how the squared error related to total interrupted power per operating scenario can be used as an error measure, and as a post analysis verification method. This approach is not further explored in this thesis, as it gave no definite conclusion.

4.8 Other Remarks

The main challenges with the scenario selection process have so far been discussed. However, there are other problems as well, which can be equally important in certain situations.

4.8.1 Extreme States and Outliers

A major concern in reliability assessment of power systems is the extreme cases - those cases with low probability and very high consequences. These are often hard to detect, and are typically not sufficiently accounted for when only expected values are used as a measure of power system reliability.

In the context of the scenario selection method, the question is if there exist extreme operating scenarios? And potentially how these can be detected by the method. When using the EMPS-NC, the market analysis is done with respect to compliance with the $N-1$ criterion. This criterion is quite restrictive on the market scenarios, and thus essentially results in only "normal" scenarios. An extreme case is in these situations caused by a combination of forced outages, not by an operating scenario itself.

If an alternative criterion is used, which does not put that much restrictions on the power market scenarios, there could be situation where (potentially) the operating scenarios themselves could be a factor leading to extreme states.

4.8.2 Network Topology

In the SoS analysis discussed in chapter 2.3.1, all operating scenarios have the same initial network topology. However, the network topology of the operating scenarios can change if, e.g., maintenance schedules are included in the power market analysis. If this is the case, one of two adjustments are necessary when applying the scenario selection methods, as a change in topology can cause large differences in the results of a contingency analysis, and thus affect what makes operating scenarios similar.

One option is to include the network topology in the feature set. However, this information typically takes the form of a categorical feature, and might not mix that well with the quantitative features used (so far) as input to the scenario selection method. The easiest approach is to split the generated set of operating scenarios into groups of operating scenarios with the same network topology, and apply the scenario selection method on each group independently.

4.8.3 Time Efficiency

As previously discussed, the main objective of the scenario selection method is to reduce the computation time of the contingency analysis in the SAMLREL SoS analysis framework.

The computational savings are in this thesis measured by how many percent of the full set the representative set is, and the exact time savings are not reported. In the next chapter, a reliability assessment of the Western part of Norway is included. In the current version of the SAMREL SoS analysis framework, the contingency analysis of this system (with AC power flow) takes about 5-7 weeks (depending on the analysis depth), while picking a representative set with k-means, agglomerative clustering, or SOM take about 1 minute. Thus, the actual time it takes to pick out the representative set is negligible in this context. However, the IGMM method requires a MCMC simulation time from 12-72 hours (or more) in its current implementation, and is thus far less efficient than the other methods in terms of picking a representative set.

4.8.4 Cluster Representatives

Instead of choosing one operating scenario to represent each group, an alternative is to choose a set of operating scenarios from each group, evaluate the reliability indices for each of these operating scenarios, find a weighted average of these reliability indices, and use this weighted average to represent the re-

liability level of each group. However, this requires a method for picking out the group characteristics, and increases the computational requirements. Experiences have shown that it is sufficient to only use the medoid as a group representatives (which is shown in the next chapter).

Chapter 5

Case Studies

During the development of the scenario selection method, several case studies were conducted to answer the questions raised in the previous chapters, and for verification of the method when applied for data reduction in the SoS analysis. The main results of these case studies are presented in this chapter. Some of the results are taken from publications D and F, while other results are included to supplement the previously published results and to facilitate the discussion in chapter 6.

5.1 Notation

Different feature representations, clustering algorithms, and number of groups in the representative set are compared in the following case studies. The different clustering algorithms, introduced in the previous chapter, are denoted:

- ACCL: Agglomerative Clustering with Complete Linkage
- IGMM: The Infinite Gaussian Mixture Model
- KM: K-Means clustering
- SOM: Self-Organising Map clustering

To explicitly express which feature set and how many groups/clusters there are in the representative set, the following notation is adopted:

$$\text{ACCL}(X_?, k)$$

Here, ACCL refers to the clustering algorithm, $X_?$ refers to the feature set used to represent similarity, while k refers to the number of groups in the representative set. The different feature sets are defined later in chapter 5.3.

For instance, $KM(X_1, 150)$ is a representative set found by the k-means algorithm with 150 groups, and where feature set X_1 is used to represent similarity of the operating scenarios.

The reliability indices based on a full reliability assessment (full SoS analysis), i.e., a contingency and reliability analysis of all operating scenarios generated by the EMPS-NC analysis, are denoted target values (TV). In this, and the remaining chapters, reliability assessment refers to the reliability assessment (incorporating power market scenarios) as defined in Figure 2.4.

5.2 Test Networks

Two networks were used in the case studies - a small test network and the Nordic power system.

5.2.1 The Four-Area Test Network

The Four-Area Test Network is a small power system designed for testing and benchmarking of EMPS-NC. The network is shown in Figure 5.1. The within area voltage levels are 66kV, while transmission lines connecting the areas operate at 132kV.

For this network, EMPS-NC models four operating scenarios per week, for 50 years to hydro-inflow years, which gives a total of 10400 operating scenarios to be analysed in the reliability assessment.

5.2.2 Western Norway

The other power system that was used in these case studies is the Nordic transmission system - Denmark, Finland, Norway, and Sweden. In the model, Sweden is replaced with a 30 bus network equivalent, giving a total of about 1700 buses in the system.

For the Nordic system, EMPS-NC generates five operating scenarios per week, for 50 years of hydro inflow data. Thus, there are a total of 13000 generated operating scenarios.

The reliability assessment is only concerned with a 60 bus subsystem in the Western part of Norway, as indicated bu the red circle in Figure 5.2, but

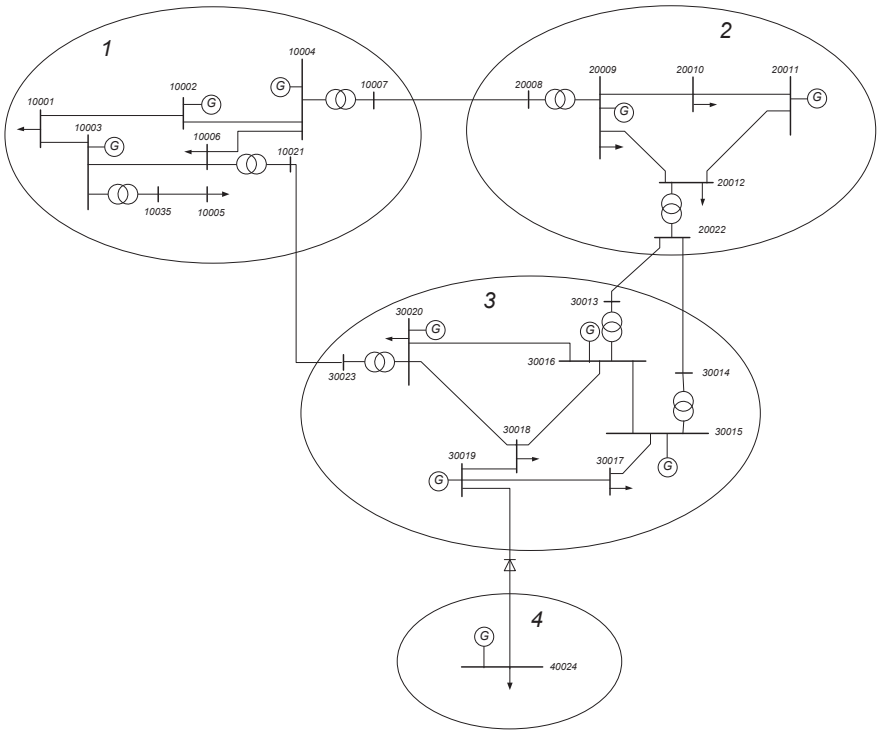


Figure 5.1: The Four-Area Test Network. The arrows indicate delivery points. Generation in area 1-3 is dominated by hydro power, while area 4 represent a mixture of thermal and wind power. The line between area 3 and area 4 is an HVDC connection.

the power flow problem is solved for the whole synchronous areas of Eastern Denmark, Finland, Norway, and Sweden. The peak load in the subsystem is about 2.5 GW, and the voltage levels in the subsystem are 132kV and 300kV.

5.3 Feature Selection

As discussed in the previous chapter, each operating scenario must be represented by some features (data points), before similarity measures and clustering

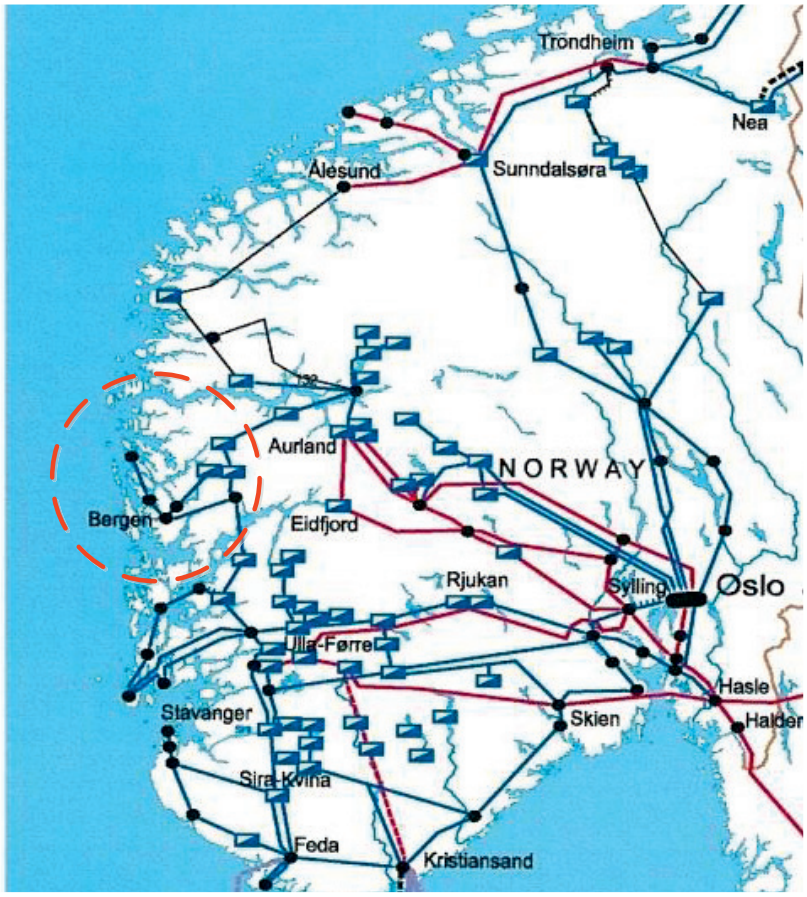


Figure 5.2: Western Norway. The peak load in the subsystem (defined by the red circle) is about 2.5 GW, and the voltage levels in the subsystem are 132kV and 300kV.

algorithms can be used to find clusters/groups of similar operating scenarios.

The objective of the feature selection process, in the context of scenario selection, is to find the set of features that represents what makes operating scenarios similar. Is similarity related to the output from EMPS-NC, the contingency analysis, the topology of the power system, or a combination? And how can this similarity be represented by quantitative features?

Feature selection is in general a case dependent process, and so is the case for the scenario selection method. The results in this section show that the feature selection depends on, e.g., the network under study and the contingency analysis model.

5.3.1 Possible Feature Sets

Below are six different feature sets X_f defined. These sets were used as a basis to find similar operating scenarios, and compared against each other with respect to the performance of the scenario selection method, i.e., how close the scenario selection indices were to the target values. The feature sets are defined such that different information and parameters are used to represent operating scenarios.

X_1 - System Load and Generation Reserve

These two features provide information about the power system state in a compact manner. This feature set is used in [34] for classification of system states in a well-being analysis. Both the active and reactive part of the variables can be included in this set. If similarity of the operating scenarios is related to system load and generation level, this can be a good basis for quantifying similarity. However, these features provide no information about the state of the transmission system, and thus might not be an appropriate basis if there are, e.g., transmission bottlenecks in the system.

X_2 - Delivery Point Loads

To use the delivery point loads as features, give more information about the transmission system state compared to the total system load, as it provides information about where in the system the large loads are. Some loads might be small, or be nearly constant, and thus considered as redundant features which should not be included in the feature set.

X_3 - Power Injections

The power injections are the sum of the load and generation per bus bar P_i, Q_i. If DC (linearised) power flow is used, only the active power injections are relevant. This feature set is used in [35, 38], for classification and clustering of system states respectively. This set provides more detailed information than X_1 and X_2, but might result in a very large number of features, especially for large power systems.

X_4 - Power Flow in System

This is the MVA flow (MW flow if DC power flow is used) trough all transmission components. This requires a solution of the power flow for each operating scenario, which have a computational cost not incurred by the previous feature sets. However, if the transmission system puts limitations on the reliability level of the operating scenarios, this set can be a good basis for finding similar operating scenarios.

X_5 - Power Flow in System in Percent

This is the same set as X_4, but when AC power flow is used, this set use the current flowing through each component to find the loading on that given component.

X_6 - Import/Export Per Area

If the system is split into areas, one can use the import/export per area to measure "how much energy is moved around in the system" in a compact manner. This can be a useful representation of similarity if there are highly meshed areas, with weak links between them. This feature set is used in publication A and C for clustering and classification of operating scenarios.

Remarks

As the feature set should provide information about most of the problems encountered in the contingency analysis, it might be necessary to use a combination of the sets described above as a basis for similarity, to fully capture/represent similarity of the different operating scenarios. The problem of finding optimal feature sets is discussed in [33, 36, 37, 55]. Reference [33] discusses how unavailable transmission capacity can be included in the feature set. However, this is not a relevant feature when clustering operating scenarios generated by EMPS-NC, as EMPS-NC use a constant topology of the transmission network in the SoS analysis discussed here.

After deciding on a specific feature set, correlation analysis, data scaling, and projection methods were used to prepare the data, and increase the performance of the clustering algorithms.

5.3.2 The Four-Area Test Network Results

For the four-area test network, the analysed contingencies in the reliability assessment included all single line and transformer outages and combinations of double line and transformer outages. AC power flow was used, and both the MRM and the CFM were used for the contingency analysis. Agglomerative clustering with complete linkage was used as clustering algorithm, and the number of clusters was fixed at 1000, i.e., ACCL$(\cdot, 1000)$ was used to find the representative set.

Reliability Assessment

In Figure 5.3, the $N-1$ secure and insecure operating scenarios are plotted based on feature set X_6 (import/export per area).

The annual EENS ($EENS_y^a$) was calculated, for each simulated year, as described in (2.5), and is shown as the black curves in Figure 5.4a and Figure 5.5a, for the MRM and the CFM respectively.

Scenario Selection Indices

The annual EENS ($\widetilde{EENS}_y^a$), based on the different feature sets, is shown in Figure 5.4a and Figure 5.5a, for the MRM and the CFM respectively.

Figure 5.4b and Figure 5.5b show $\widetilde{EENS}_y^a / EENS_y^a$ (the relative error per simulated year) for the different consequence models and feature sets.

Comments

For the MRM, feature set X_2 (delivery point loads) is the best, while feature set X_6 (import/export per area) is the best for the CFM. Based on Figure 5.4b and Figure 5.5b, it seems that the MRM is more sensitive to the feature selection process, as the plotted relative errors are large for some feature sets. However, in the cases where the relative errors are very large, the value of the estimated index is fairly low (left area in Figure 5.4a), thus making this type of error measure quite sensitive to (small) variations in the estimated indices. The absolute error is fairly low for these estimated indices.

Based on Figure 5.3, X_6 (import/export per area) seems like a good set for separating $N-1$ secure and insecure states for both the MRM and the CFM. However, for the MRM, this feature set is not a good set for clustering scenarios when the goal is to estimate the annual EENS, as seen in Figure 5.4b. While for the CFM, this is the best feature set as seen in Figure 5.5b.

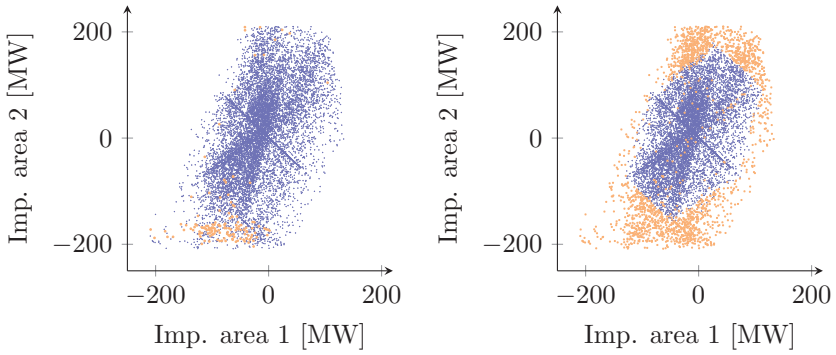


Figure 5.3: $N-1$ secure and insecure scenarios, plotted for feature set X_6 (import/export per area), for the MRM (left) and the CFM (right).

If, e.g., the k-means algorithm is used for comparing the features sets, very similar results apply.

5.3.3 Western Norway Results

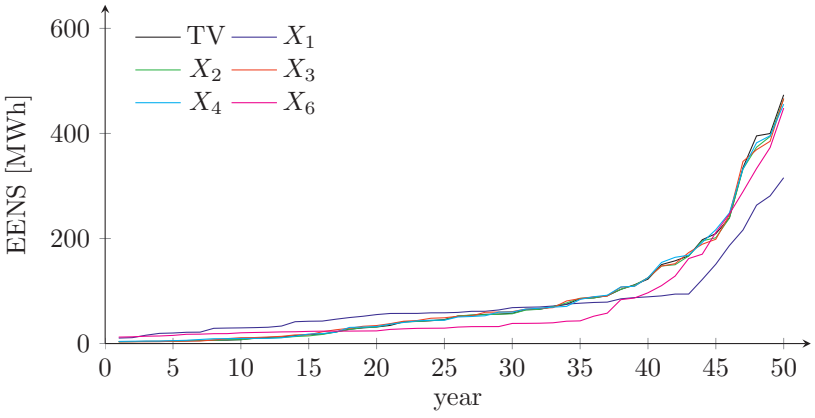
For Western Norway, the analysed contingencies in the reliability assessment included all single line and transformer outages and combinations of double line and transformer outages. DC power flow was used, and both the MRM and the CFM were used for the contingency analysis. Agglomerative clustering with complete linkage was used as clustering algorithm, and the number of clusters was fixed at 1000, i.e., ACCL$(\cdot, 1000)$ was used to find the representative set.

The $EENS_y^a$ is shown as the black curve in Figure 5.6a and Figure 5.7a, for the MRM and the CFM respectively.

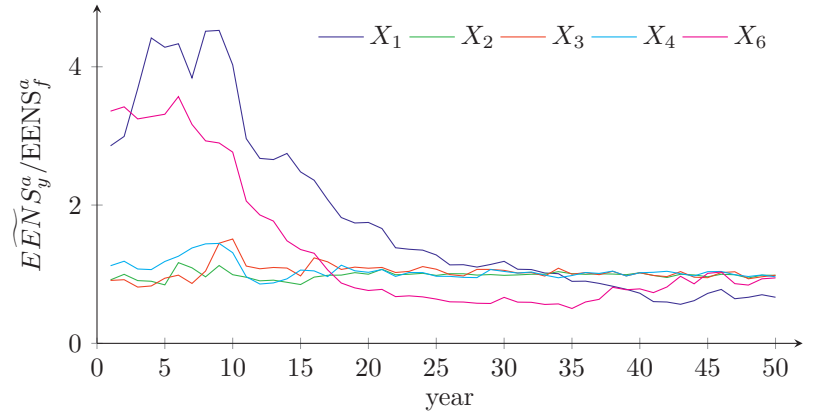
Scenario Selection Indices

The $\widetilde{EENS}_y^a$, based on the different feature sets, is shown in Figure 5.6a and Figure 5.7a, for the MRM and the CFM respectively.

Figure 5.6b and Figure 5.7b show $\widetilde{EENS}_y^a / EENS_y^a$ (the relative error per simulated year) for the different consequence models and feature sets.

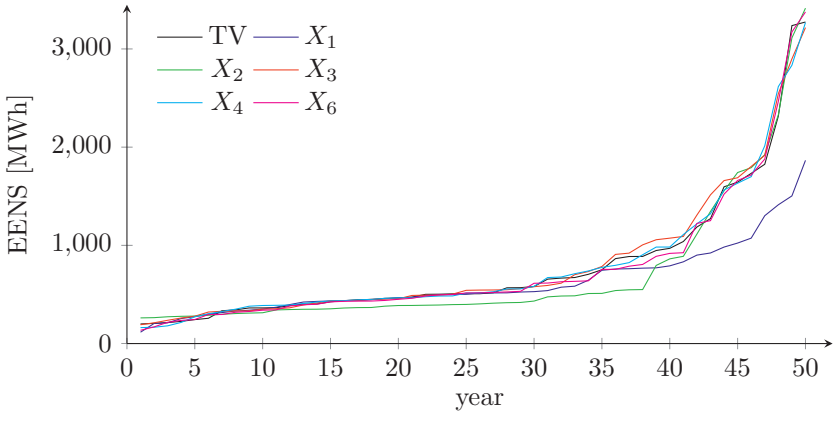


(a) EENS

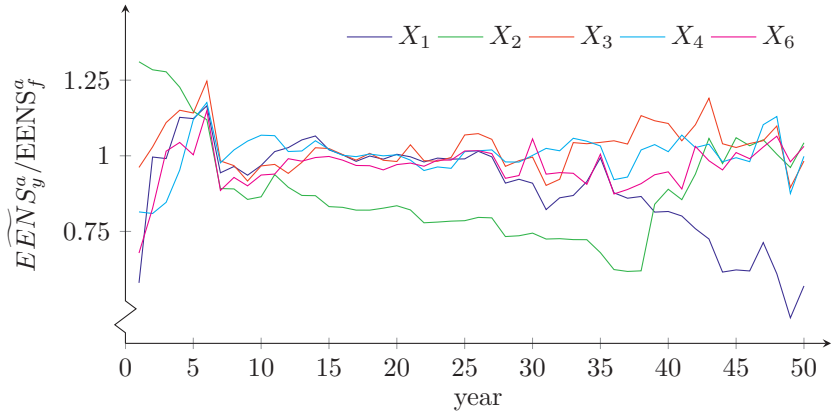


(b) relative error

Figure 5.4: Illustration of the effect of feature selection, for the four-area test network, when ACCL$(\cdot, 1000)$ was used to find the representative set. The MRM was used for the contingency analysis. Feature set X_5 gave the same results as feature set X_4, and is therefore not plotted here.

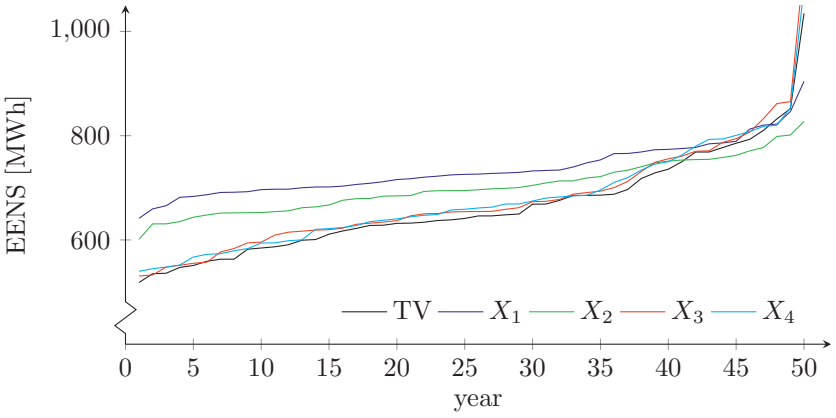


(a) EENS

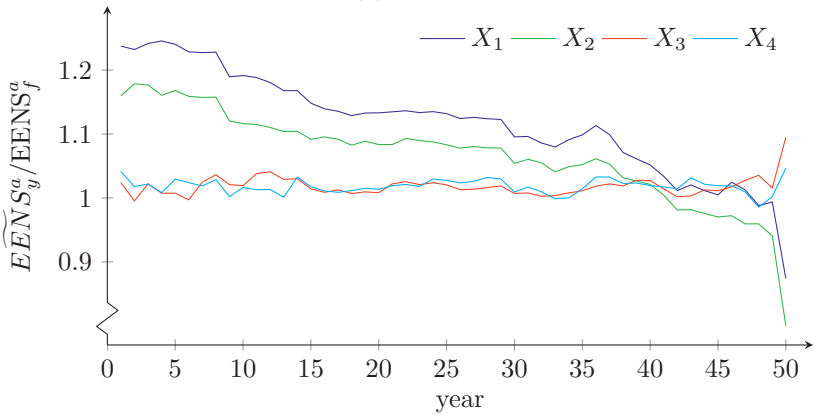


(b) relative error

Figure 5.5: Illustration of the effect of feature selection, for the four-area test network, when ACCL$(\cdot, 1000)$ was used to find the representative set. The CFM was used for the contingency analysis. Feature set X_5 gave the same results as feature set X_4, and is therefore not plotted here.

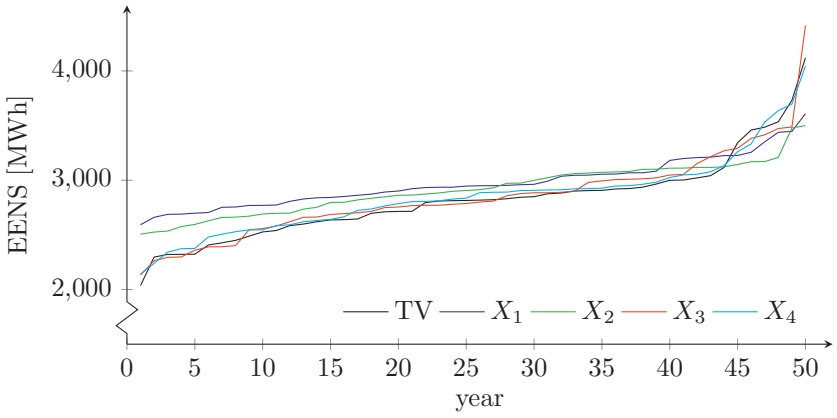


(a) EENS

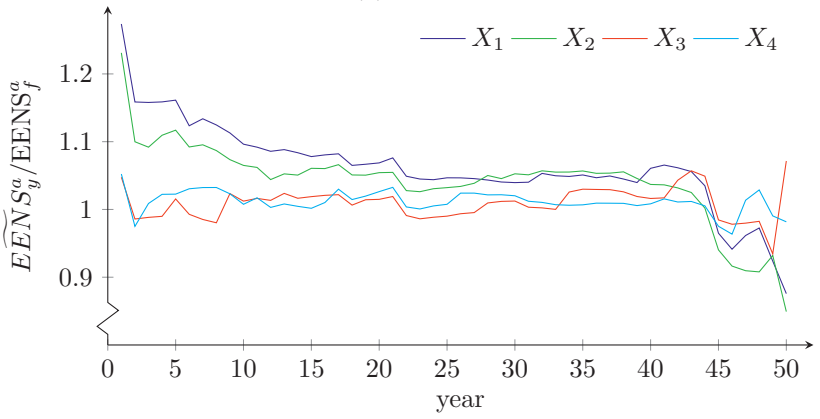


(b) relative error

Figure 5.6: Illustration of the effect of feature selection, for Western Norway, when ACCL$(\cdot, 1000)$ was used to find the representative set. The MRM was used for the contingency analysis. Feature sets X_5 and X_6 are not defined/relevant for this system.



(a) EENS



(b) relative error

Figure 5.7: Illustration of the effect of feature selection, for Western Norway, when ACCL$(\cdot, 1000)$ was used to find the representative set. The CFM was used for the contingency analysis. Feature sets X_5 and X_6 are not defined/relevant for this system.

5.3.4 Comments

Other reliability indices, such as the expected average interruption duration, were also calculated based on the representative sets, and used as a method of comparing the different feature sets. The relative errors of these indices were in the same range as those of the annual EENS, and thus gave very similar results with respect to the performance of the different feature sets.

The results in this section show that the feature selection process depends on the network (The Four-Area Test Network or Western Norway) and the contingency analysis set-up (MRM or CFM). Feature set X_3 (power injections) is the set with the on average best performance.

5.4 Clustering Algorithm

Based on the results of the feature selection study, the (optimal) feature sets were used to compare the different clustering algorithms. The different algorithms were introduced in chapter 4. For both the four-area test network and Western Norway, only up to second order transmission outages were analysed in the contingency analysis, DC power flow was used, and only MRM was used for the contingency analysis in the case studies discussed in this section.

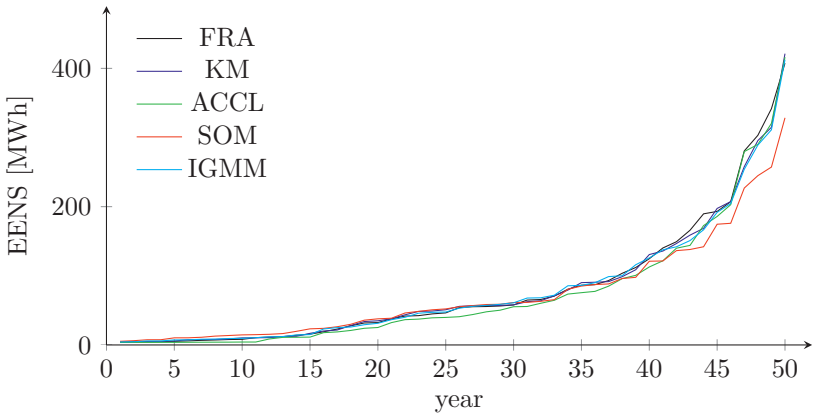
5.4.1 The Four-Area Test Network Results

The annual EENS ($EENS_y^a$) was again calculated based on an analysis of all operating scenarios, and used as the target value for the scenario selection index ($\widetilde{EENS}_y^a$), where feature set X_2 (delivery point loads) from the previous section was used for quantifying similarity.

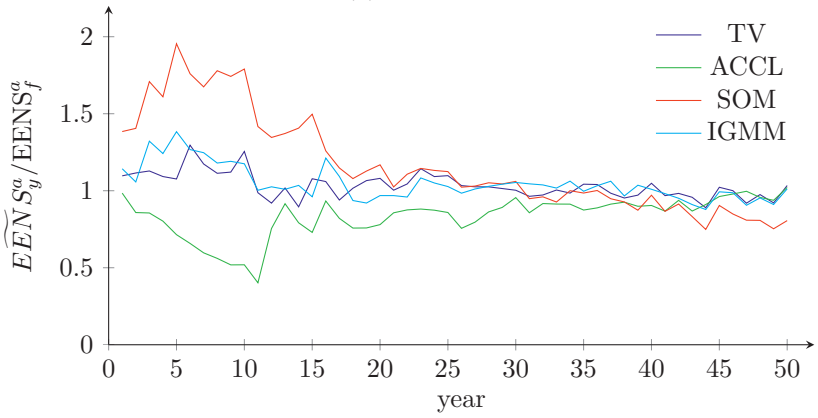
First, IGMM was used to find groups of similar operating scenarios. The solution with the highest posterior likelihood was a mixture with 52 multivariate Gaussian densities, i.e., IGMM produced 52 groups of similar operating scenarios. The scenario selection indices, based on the IGMM solution, is shown as the cyan coloured curve in Figure 5.8. The relative error is also shown.

To make a fair comparison with the other algorithms, KM, ACCL, and SOM were used find representative sets with 52 groups as well, i.e., scenario selection indices were calculated based on KM$(X_2, 52)$, ACCL$(X_2, 52)$, and SOM$(X_2, 52)$. The resulting annual EENS values, and the relative errors, are shown in Figure 5.8.

In Figure 5.9, scenario selection indices, and the relative errors, based on KM$(X_2, 1000)$, ACCL$(X_2, 1000)$, and SOM$(X_2, 1000)$ are shown, as well as the

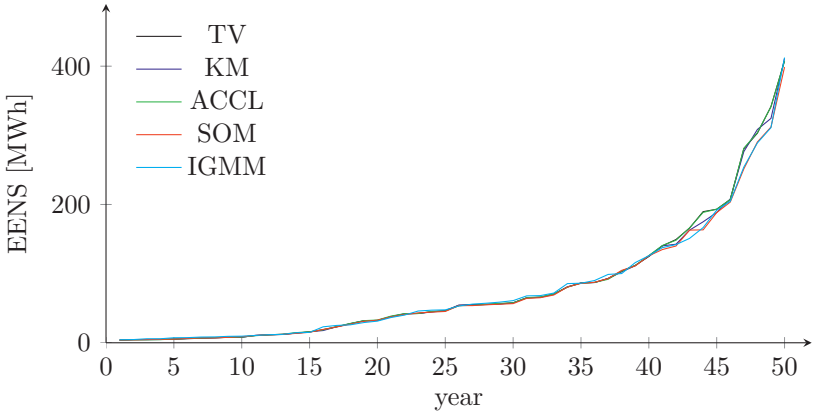


(a) EENS

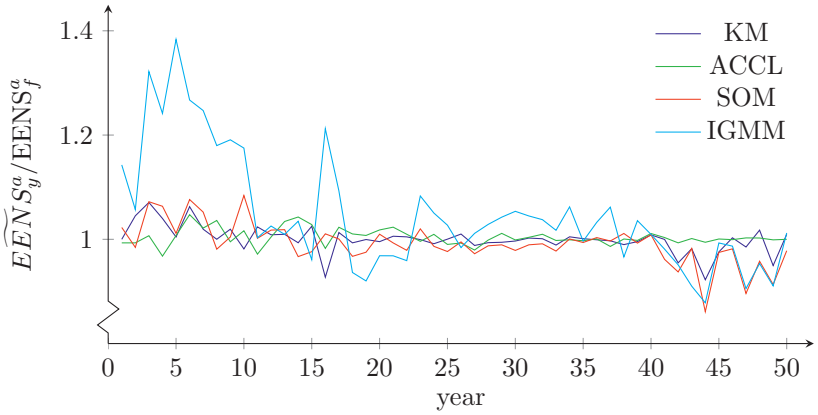


(b) relative error

Figure 5.8: Upper plot: Illustration of the effect of the choice of clustering algorithm, for the four-area test network, when KM, ACCL, SOM and IGMM were used to find representative sets, when $k = 52$, and X_2 was used as feature set. The MRM was used for the contingency analysis. In the lower plot, the corresponding relative errors are shown.



(a) EENS



(b) relative error

Figure 5.9: Upper plot: Illustration of the effect of the choice of clustering algorithm, for the four-area test network, when KM, ACCL, and SOM were used to find representative sets, when $k = 1000$, and X_2 was used as feature set. For the indices based on the IGMM solution, $k = 52$ as in the previous figure. The MRM was used for the contingency analysis. In the lower plot, the corresponding relative errors are shown.

.

scenario selection indices and relative errors based on IGMM($X_2, 52$).

5.4.2 Western Norway Results

Again, the annual EENS ($EENS_y^a$) was calculated based on an analysis of all operating scenarios, and used as the target value for the scenario selection indices ($\widehat{EENS}_y^a$), where feature set X_3 (power injections) from the previous section was used for quantifying similarity in this study of Western Norway.

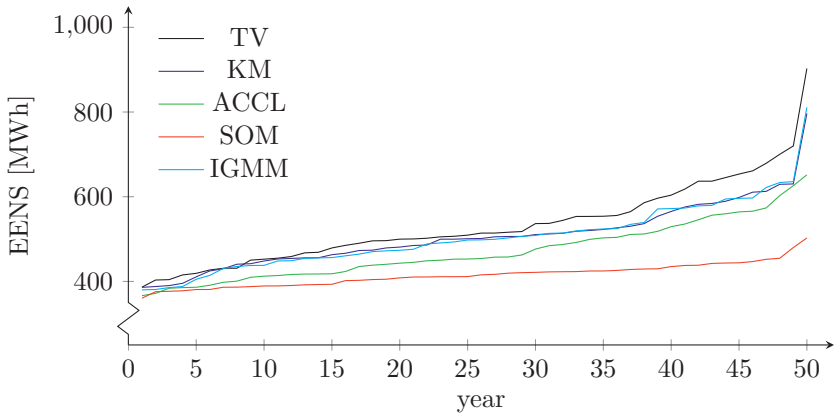
For IGMM, the solution with the highest posterior likelihood gave a solution with 65 different groups. Again, to make a fair comparison with the other algorithms, KM, ACCL, and SOM were used find representative sets with 65 groups as well, i.e., scenario selection indices were calculated based on KM($X_3, 65$), ACCL($X_3, 65$), and SOM($X_3, 65$). The resulting reliability indices, and the relative errors, are shown in Figure 5.10.

In Figure 5.11, scenario selection indices, and the relative errors, based on KM($X_3, 1000$), ACCL($X_3, 1000$), and SOM($X_3, 1000$) are shown, as well as the scenario selection indices and relative errors based on IGMM($X_3, 65$).

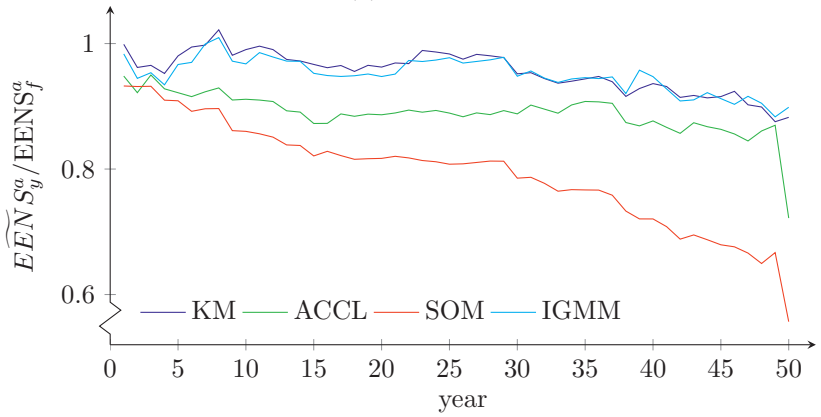
5.4.3 Comments

Based on Figure 5.8 and Figure 5.10, the IGMM model gives fairly good results when finding representative sets with a very small number of groups compared to the full set. However, only using the system annual EENS as a method for comparison is not sufficient, as the delivery point indices are of interest as well. For delivery point indices, the relative errors are much higher, and cluster solutions with more groups are necessary to get results with reasonable accuracy, e.g., the representative set ACCL($X_3, 1000$) produced delivery point indices within a reasonable range of the target values. This is better illustrated in the next case study.

Comparing the results in this case study with the results of feature selection case study, it is evident that to choose a good/correct feature set is much more important than the choice of clustering algorithm, in terms of optimal performance of the scenario selection method.

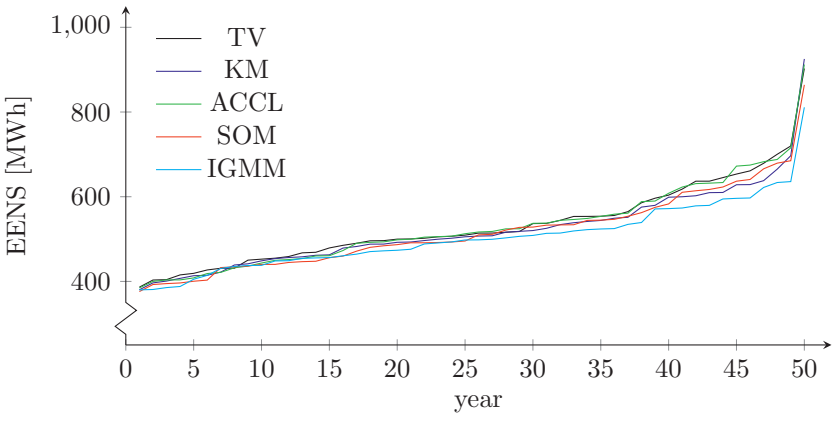


(a) EENS

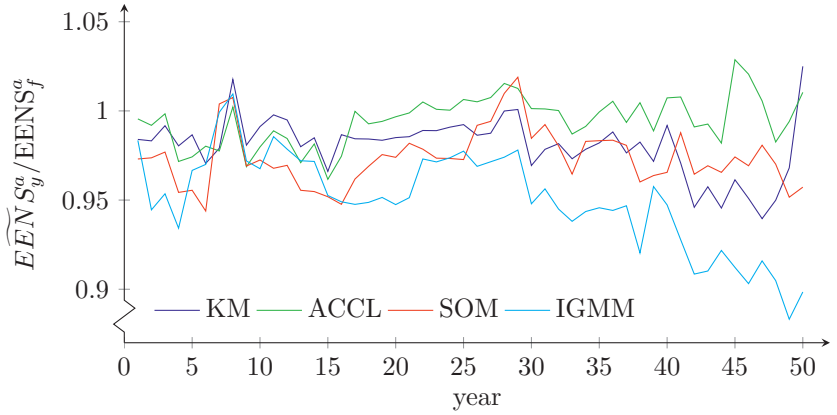


(b) relative error

Figure 5.10: Upper plot: Illustration of the effect of the choice of clustering algorithm, for Western Norway, when KM, ACCL, SOM and IGMM were used to find representative sets, when $k = 65$, and X_3 was used as feature set. The MRM was used for the contingency analysis. In the lower plot, the corresponding relative errors are shown.



(a) EENS



(b) relative error

Figure 5.11: Upper plot: Illustration of the effect of the choice of clustering algorithm, for Western Norway, when KM, ACCL, and SOM were used to find representative sets, when $k = 1000$, and X_3 was used as feature set. For the indices based on the IGMM solution, $k = 65$ as in the previous figure. The MRM was used for the contingency analysis. In the lower plot, the corresponding relative errors are shown.

5.5 Number of Clusters

As discussed in chapter 4.5, KM, ACCL, and SOM, require the number of groups k to be specified before a clustering solution (representative set) is found. For IGMM, the estimation of k is a part of the sampling procedure to build the joint density function over data (the sample of operating scenarios).

In the case studies in this section, X_2 (the power injections) was used as feature set for both the four-area test network and Western Norway, linearised power flow, and MRM were used in the contingency analysis.

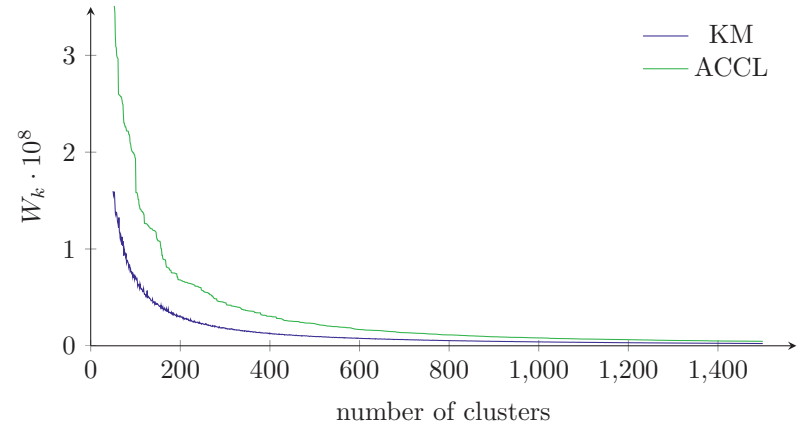
5.5.1 Tip of the Elbow

In Figure 5.12, the W_k (within cluster dispersion) index from (4.9) is plotted for both the four-area test network and Western Norway, for both KM(X_3, k) and ACCL(X_3, k) with a varying number of groups k. There is no indication of a "tip of the elbow" in these figures. Other methods, such as the "gap statistic", were used to attempt to estimate the number of groups k without success.

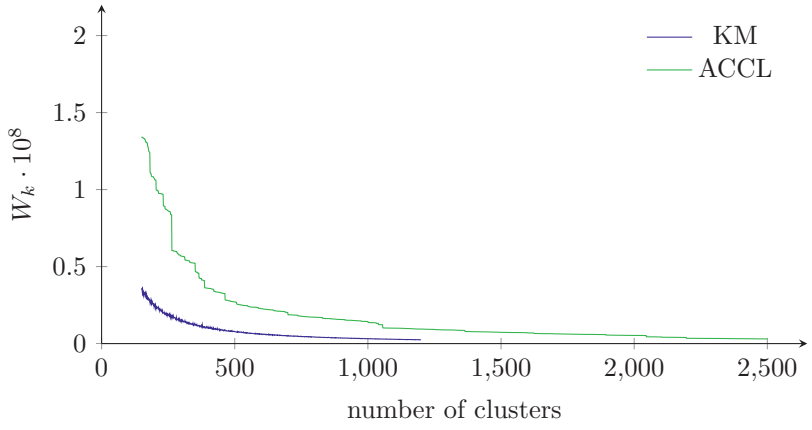
5.5.2 Reliability Indices

As discussed in chapter 4.7, the main objective of the scenario selection method is to recreate the cumulative distribution of the reliability indices, i.e., estimate the annual reliability indices (per modelled hydro-inflow year in EMPS-NC) as accurately as possible. Both system indices and delivery point indices are of interest. The expected interruption duration at a delivery point is in general more difficult to accurately estimate than the expected energy not supplied, at least with methods based on continuous data (such as the scenario selection method) since the expected interruption duration behaves as a stepwise (discrete) function. Only results from an analysis of Western Norway are discussed here, as an analysis of the four-area test network gave very similar results.

In Figure 5.13, the system $\widetilde{U}_y^a$ and $\widetilde{EENS}_y^a$ are plotted for ACCL(X_3, k), for different values of k. In Figure 5.14 and Figure 5.15, $\widetilde{U}_{y,d}^a$ is plotted for two delivery points, denoted delivery point A and B. When $k = 1000$, the absolute error in Figure 5.14 is in the range (0.00,0.13) [h], while in Figure 5.15 the absolute error is in the range (0.00,0.02) [h].



(a) 4-Area Test Network



(b) Western Norway

Figure 5.12: Plot of the W_k index as a function of the number of clusters, for the four-area test network and Western Norway, for ACCL$(X_3, \cdot)$ and KM$(X_3, \cdot)$

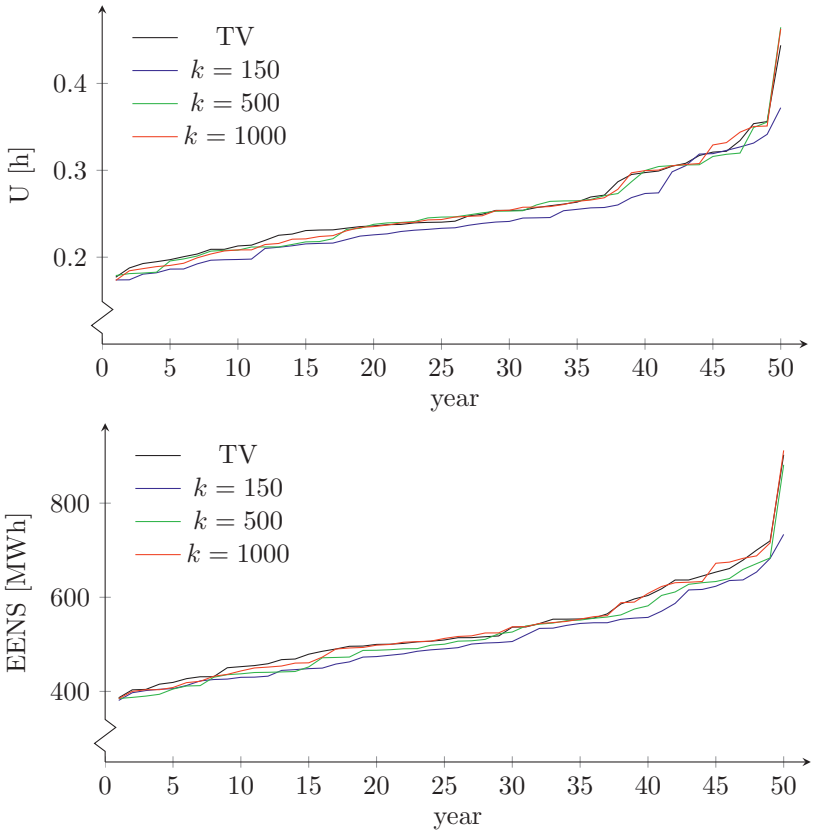


Figure 5.13: Annual average interruption duration and annual system expected energy not supplied for Western Norway, where $\mathrm{ACCL}(X_3, \cdot)$ was used to find the representative set.

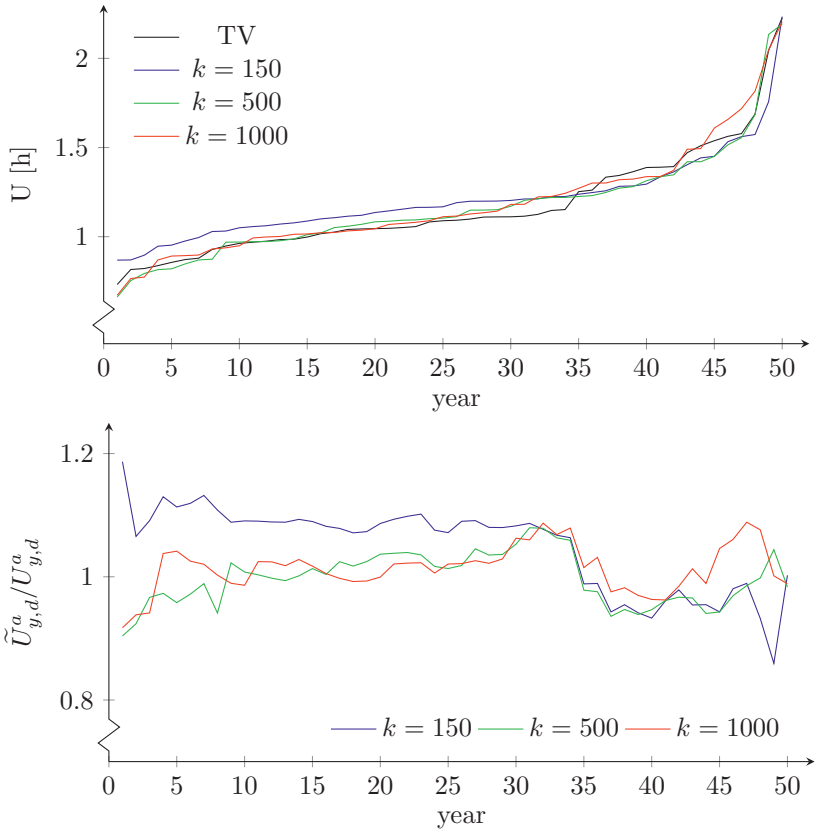


Figure 5.14: Annual expected interruption duration at a delivery point A in Western Norway, where $\mathrm{ACCL}(X_3, \cdot)$ was used to find the representative set.

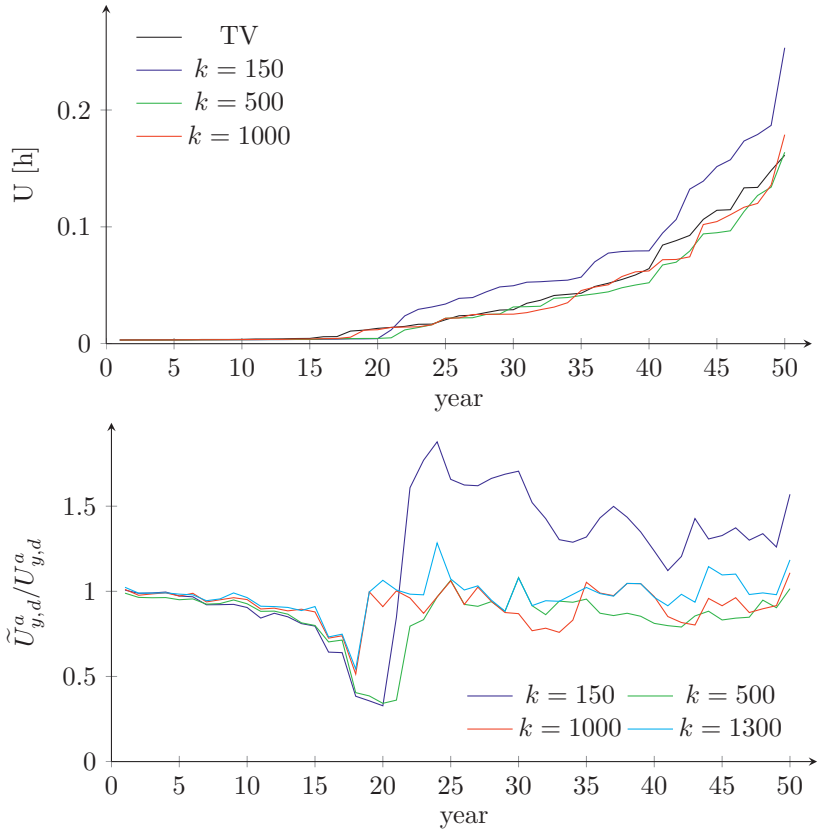


Figure 5.15: Annual expected interruption duration at a delivery point B in Western Norway, where ACCL$(X_3, \cdot)$ was used to find the representative set.

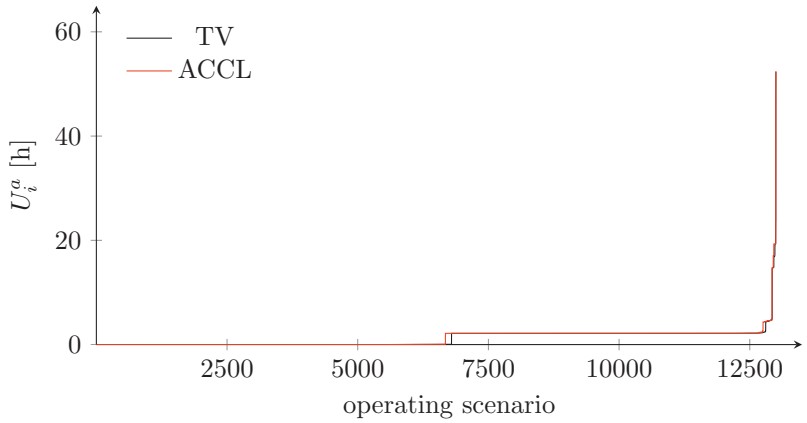


Figure 5.16: Annualised expected interruption duration (per operating scenario) at delivery point B in Western Norway, where ACCL($X_3, 1000$) was used to find the representative set.

Comments

Note that the relative error in Figure 5.15 is larger than the others (Figure 5.13 and Figure 5.14), especially as the index increase from about 0 to higher values around year 15. This high error is mainly caused by two factors.

First, the interruption duration behave in a stepwise manner, as illustrated in Figure 5.16, where the interruption duration per operating scenario for delivery point B is plotted in increasing order. The relative high error in Figure 5.15 is a result of this jump, as the scenario selection method "misplace" this jump.

Second, delivery point B is a fairly reliable delivery point with a fairly low load, thus the feature value for this point is not as dominant a value in the feature set as delivery points with a higher value (load). This means that a small change in the load at this delivery point, which partly causes the jump seen in Figure 5.16, will not be "seen" by the scenario selection method. The implication of this is that increasing the number of groups k only marginally lower the errors here, as this value is not "clearly seen" in the clustering process.

5.5.3 Sample Variation

As discussed in chapter 4.2, the representative set should maintain the sample variation of the full set, but at the same time reduce the number of operating

scenarios. In Figure 5.17, the annual EENS is plotted per operating scenario for the four-area test network, as a function of the system load. This is done for the full set, and for ACCL(X_3, k) with $k = 150, 500, 1000$. The solution with $k = 150$ seems to average out too much of the sample variation, while the solution with $k = 1000$ seems to be an appropriate trade-off between data reduction and maintaining the sample variation.

5.5.4 Comments

When k was set to 1000, all indices in the system were within a reasonable error range, getting a 90% data reduction while most of the information based on a full analysis was retained.

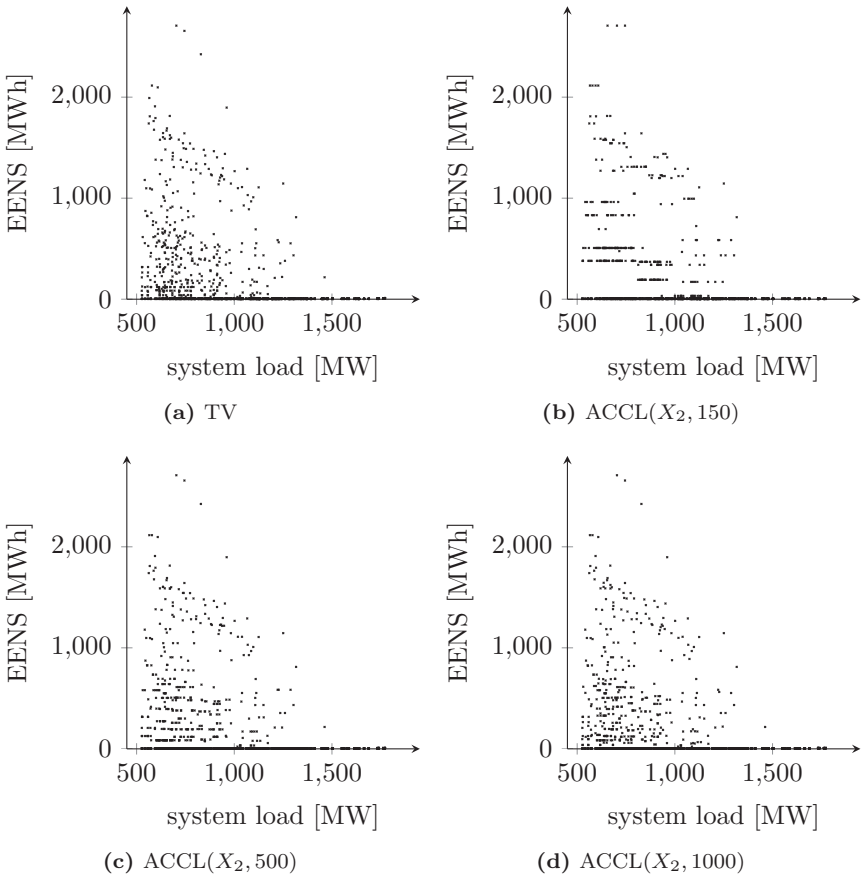
Setting $k \approx 0.1 \cdot n$ is shown to produce scenario selection indices within a 5-10% range of the target values. Take for instance the annualised system EENS per operating scenario, as defined by (2.3). This index is in the range of 50 [MWh/year] to about 5000 [MWh/year] for Western Norway. Considering the large range of this index, and the fact that there is no clear indication of a "natural number of clusters" among the data, some error in the scenario selection indices is to be expected. Thus, a 90% data reduction, with scenario selection indices within a 5-10% range of the target values, has been set as a reasonable and acceptable error.

5.6 Clustering Structure

So far, reliability indices and the W_k index have been used to attempt to determine the appropriate number of groups in the representative set. However, a closer look at the clustering solutions themselves can give a better understanding why the scenario selection method works.

5.6.1 Confusion Plots

To visualise (illustrate) the cluster solution, it is useful to plot the confusion matrix relating to a specific cluster solution. The confusion matrix is a binary matrix, where entry (i, j) is 1 if operating scenario i and j belong to the same cluster, and zero otherwise. The confusion matrices, for Western Norway, is shown for KM$(X_3, 150)$ and ACCL$(X_3, 150)$ in Figure 5.18. Only the part of the matrix corresponding to the first 520 operating scenarios (first two years



(a) TV **(b)** ACCL($X_2, 150$) **(c)** ACCL($X_2, 500$) **(d)** ACCL($X_2, 1000$)

Figure 5.17: Plot of EENS per operating scenario (as a function of system load) based on a full analysis (top left) and ACCL solutions with 150 (top right), 500 (lower left), and 1000 (lower right) operating scenarios in the representative set. It seems that 150 clusters/groups results in too much averaging, as much of the variance in the sample is lost, while 1000 clusters seems to be sufficient to maintain most of the sample variance.

in EMPS-NC analysis) is plotted. There is a clear seasonal trend in how the clusters are formed.

5.6.2 Cluster Structures

Based on the confusion plots in Figure 5.18, it seems that KM and ACCL have the same seasonal trend in how the clusters are formed. To further compare the algorithms, the largest group, and mean and median group sizes, in the representative set, are plotted for the two algorithms as a function of the total number of clusters in Figure 5.19.

From Figure 5.19, one can see that KM produced clusters of approximately the same size, as the maximum, mean, and median cluster size almost coincide. While the ACCL produced a few (very) large clusters and many small ones, as seen in Figure 5.19 as the maximum is much higher than the median, and the mean is higher than the median cluster size for this algorithm.

Note that the largest group size in KM varies, while it is monotonically increasing for ACCL, as k decreases. For KM, a new solution was found every time (for each value of k), while each ACCL solution build on the previous solution as k was decreased.

5.7 "Rule-Based" Scenario Selection

An alternative to using the scenario selection method to pick a subset of scenarios is to pick the subset based on the total system load. Two alternatives were used, where the selection process was done with the goal of finding subsets (representative sets) of size 1000, to ensure that this approach had about the same computational effort as the scenario selection method with $k = 1000$. Only results from an analysis of Western Norway are discussed here.

Worst Case Analysis (WCA)

For each year, the 20 operating scenarios with the highest system load were found. These operating scenarios were analysed, and the annualised EENS per operating scenario was evaluated. The annual EENS was defined as the mean value of the 20 EENS indices. This is plotted as the blue curve in Figure 5.20.

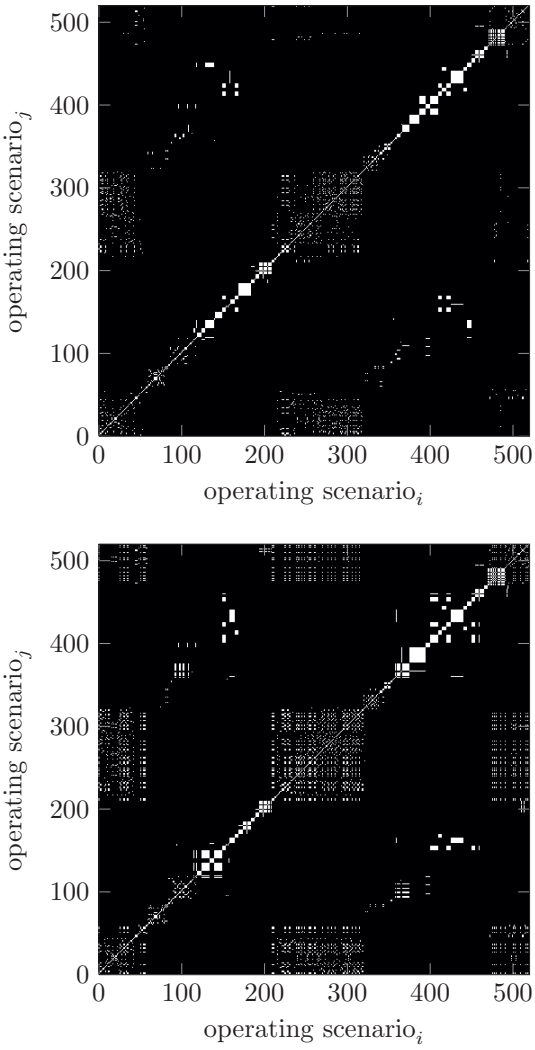
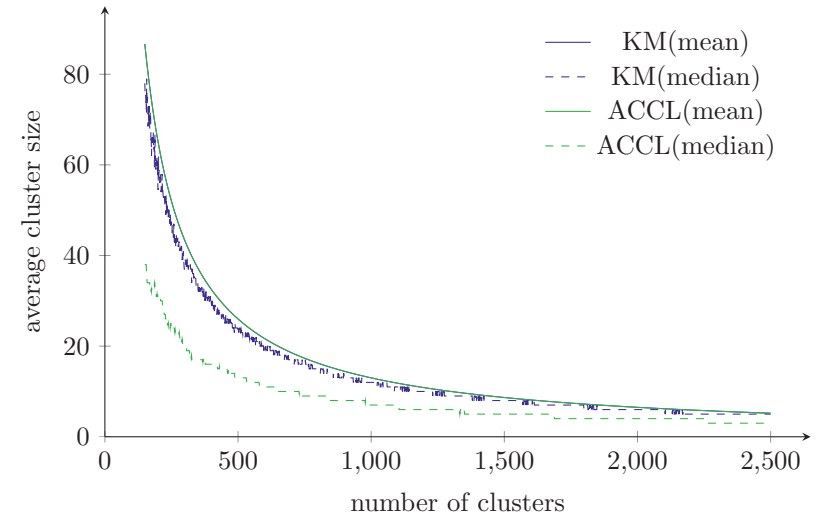
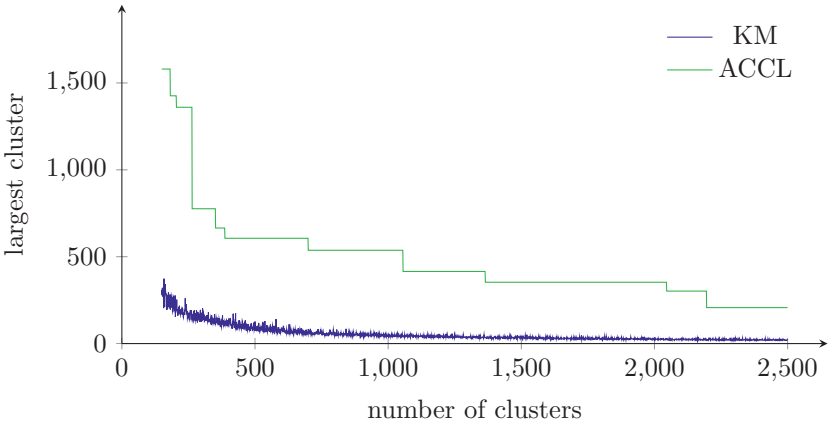


Figure 5.18: Plot of the confusion matrices for Western Norway for K-Means (upper) and ACCL (lower), when the number for groups in the representative set was 150. The clustering solutions look quite similar, and there is a clear seasonal trend in terms of similarity of operating scenarios.



(a) Mean and Median Cluster Size



(b) Largest Cluster Size

Figure 5.19: Upper: The mean and median groups size in the representative set for KM(X_3, k) and ACCL(X_3, k) (as function of the total number of groups in the representative set.), in an analysis of Western Norway. Lower: The maximum group size in the representative set for KM(X_3, k) and ACCL(X_3, k) (as function of the total number of groups in the representative set.)

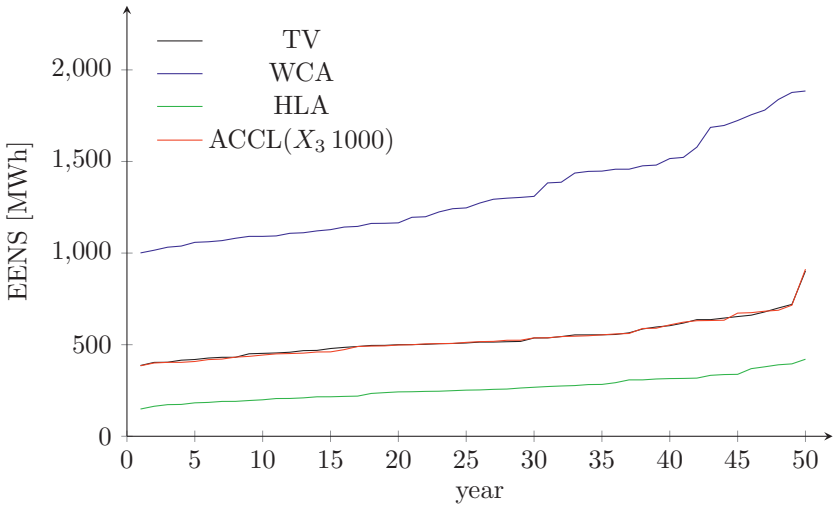


Figure 5.20: Annual EENS, where subsets of operating scenarios, of size 1000, are picked based on the total system load, instead of using the scenario selection method.

High Load Analysis (HLA)

The 1000 operating scenarios with the highest system load were found, and the annualised EENS for each of these operating scenarios were evaluated. For the remaining 12000 scenarios, it was assumed that the contribution to the annual EENS for these operating scenarios was negligible, i.e., the annualised EENS for each of these scenarios was set to zero. Then, then annual EENS was found as described in (4.11). This is plotted as the red curve in Figure 5.20.

Comments

Comparing the results of the scenario selection method and the WCA and HLA methods, in Figure 5.20, it is observed that the scenario selection method results in reliability indices much closer to the target values, than the indices based on the rule-based analysis methods. It is clear that the scenario selection method is a much better method of objectively picking a subset of scenarios than purely basing the selection on the total system load in this type of SoS analysis. As mentioned earlier, the total system load is an important variable in generation adequacy assessment, and these rule-based selection techniques would perform

much better in such a setting.

The scenario selection is marginally slower than such rule-based methods with respect to reducing the scenario set, but the time difference vanishes when compared to the time it takes to complete the contingency analysis.

In publication B, the scenario selection method was compared with another rule-based selection technique. In that case as well, the scenario selection method outperformed the rule-based method.

5.8 Scenario Selection with Hourly Wind Data

In the preceding case studies, the variation in the generation patterns is mainly related to hydro inflow, where four (or five) operating scenarios are used to describe the generation and load profile per week. If wind farms are a part of the generation system, the variation in the generation patterns increases due to (rapidly) changing wind conditions. The wind speed, and thus the power output from the wind farm, is constantly changing, and thus affects the system operation (and reliability level). Wind power is often considered as unscheduled (and non-dispatchable) generation, which makes it less flexible and harder to control than, e.g., thermal and hydro power generation. Typically, in planning studies, an hourly wind model is used to check the impact on the power system when a wind farm is connected to the network.

The same SoS analysis set-up as described earlier is used for the contingency and reliability analysis in this case study. However, in this case study, each single hour in a year is described by an unique operating scenario, i.e., the load and generation profile over one year is composed of 8760 operating scenarios. The scenario selection method is again used to reduce the number of operating scenarios used as input to the contingency and reliability analysis.

5.8.1 The Four-Area Test Network

In this section, only the four-area test network is considered. A wind farm of 100 MW is connected to bus 10004 in area 1 in Figure 5.1. The hourly wind data, used as input to EMPS-NC analysis, are from the COSMO-EU project [56], where the data represent aggregated wind data for Southern Norway. The data are scaled to have a maximum power output of 100 MW.

The process of generating hourly operating scenarios consists of two main steps.

Table 5.1: Annual Reliability Indices

Reliability Index	Full Analysis	Scenario Selection	Error
U [h]	0.68	0.72	6.02 %
EENS [MWh]	12.5	12.8	2.40 %

First, EMPS-NC is used for the market analysis. Again, the analysis is done with respect to four load periods per week. For each load period, the total energy provided by the wind farm is the sum of the hourly wind power output for the hours corresponding to the given load period. This maintains a balance between the energy provided by the wind farm and the hydro power plants in the power market analysis.

The next step is concerned with generating hourly operating scenarios per week. For each hour within the week:

- The wind power production is set equal to that of the wind power time series.
- The load is scaled to match the daily load curves in Norway. For each day within a week, the hourly load is a factor of the mean load of the given week.
- Hydro power is used to cover the part of the load which is not covered by the wind farm.

For one EMPS year, this gives $24 \times 7 \times 52 = 8736$ operating scenarios. The minimal rescheduling model is used in the contingency analysis, and reliability indices are evaluated per operating scenario as described in chapter 2.2.3. Note that the wind power is considered as constant and non-dispatchable per hour, i.e., the wind farm cannot participate in the rescheduling in (2.6) and (2.7).

5.8.2 Results

All 8736 operating scenarios were analysed, and reliability indices evaluated for each of them. The annual values (the sum of the hourly indices) of the system average interruption duration and the system expected energy not supplied are shown in table 5.1. In Figure 5.21, the value of the reliability indices are shown per season (black bars).

For the scenario selection indices, $\mathrm{ACCL}(X_3, 874)$ (X_3: power injections) was used to find the representative set. The corresponding reliability indices

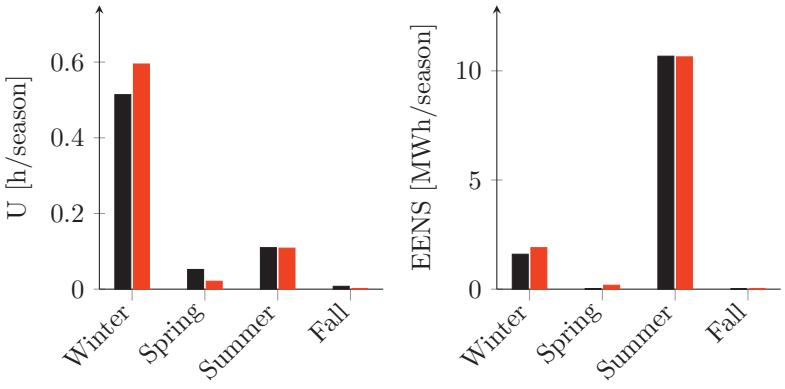


Figure 5.21: Seasonal average interruption duration (left) and expected energy not supplied (right) per season, when a wind farm of 100 MW is connected to bus 10004 in the four-area test network in Figure 5.1. The black bars are values based on a full analysis (target values), while the red bars are the values of the scenario selection indices when ACCL($X_3, 874$) (X_3: power injections) is used to find the representative set.

are reported in table 5.1, and the seasonal values are shown as the red bars in Figure 5.21. The reason for the relatively high EENS during summer is that there are a few large loads that are interrupted during the summer season, while the relatively high average interruption duration during winter is caused by many interruptions which only require minor load shedding to fix the system problems.

To illustrate the effect of connecting the wind farm to this system, the system net load, corresponding to the 500 hours with highest load, is shown in decreasing order in Figure 5.22. The net load, after wind power production is subtracted from the system load (for the same hours), is shown as the red curve in the same figure.

The same reliability analysis as described above was done for the system before the wind farm was connected. The reliability indices for the studies with and without wind are shown in table 5.2. As the scaling of the load, to match the daily load curves, follow the same procedure for each weekday and each day in the weekend, the reliability analysis for the case without wind is only concerned with $48 \times 52 = 2496$ operating scenarios. Applying the scenario

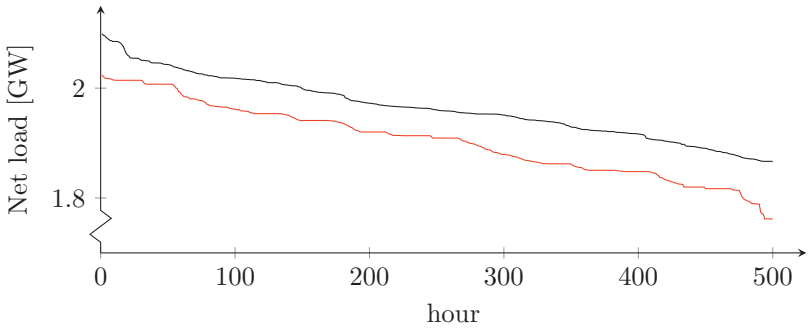


Figure 5.22: Load curve, for the 500 hours with the highest load, without wind (black curve) and the corresponding net load when wind power production is subtracted from the net load (red curve)

Table 5.2: Annual Reliability Indices

Reliability Index	Without Wind	With Wind
U [h]	0.73	0.68
EENS [MWh]	17.2	12.5

selection method on this case gives very similar results (errors) as for the case with wind power.

5.8.3 Comments

The results show that the scenario selection method can be used to find groups of operating scenario (representing an hourly time scale) when wind farms are connected to a power network, and keep the scenario selection indices within a 5-10% error range when only 10% of the hourly operating scenarios are analysed.

In the SoS analysis in this case study, the wind power production is kept constant for each operating scenario, i.e., the uncertainty relating to the hourly wind power forecast is neglected. Wind power is here included in the analysis to illustrate that the scenario selection method can deal with the increased variability in the generation patterns. However, the SoS analysis needs to be further developed/refined to if, e.g., uncertainties in the wind power forecast are to be taken into account.

Chapter 6

Discussion

Based on the results of the case studies, reported in the previous chapter, some general remarks and comments are given given in this chapter (with respect to practical applications of the scenario selection method).

6.1 Similar Operating Scenarios

For each operating scenario, generated in the power market analysis (with EMPS-NC), the power generation is given per module (power generator), and the loads are modelled per bus, in the analysed power system. The system and/or area price predictions are also given per scenario, together with the import/export to neighbouring power networks (through, e.g., HVDC connections). Water-values are given per week, and EMPS-NC keeps track of the hydro reservoir levels, and can also model water spillage (in case of flooding of the reservoirs).

To find similar operating scenarios, the first question that needs to be answered is "What make operating scenarios similar?". Based on the output of EMPS-NC, there are several variables available for representing the operating scenarios, but which of these are relevant in terms of representing similarity of operating scenarios?

6.1.1 The Consequence Analysis

In the feature selection case study in chapter 5.3, MRM and CFM were used for the contingency analysis. The results of that case study show that these two models require different feature sets with respect to optimal performance of the scenario selection method.

In terms of application of the scenario selection method, this means that not only the output of the EMPS-NC analysis is relevant for defining similarity, but also how the operating scenarios themselves are analysed (in the contingency analysis). However, this is not too surprising as, e.g., the two contingency analysis models react differently when violations of the operating criteria occur.

The Minimal Rescheduling Model

In the consequence analysis, the objective function (2.6) (and (2.7)) is defined such that for forced outages which require rescheduling (and load shedding), the cost of the corrective actions are minimised. In practice, this means that the power injections (sum of load and generation per bus) after rescheduling (and load shedding) are as close as possible to the power injections given by the initial operating scenario. Thus, the power injections at each bus in the analysed system are a natural choice of features when the goal is to find similar operating scenarios, the MRM is used for the contingency analysis, and the initial network topology is the same of all analysed operating scenarios.

The Cascading Failure Model

The CFM is based on tripping overloaded components. Thus, this model will either isolate/clear the overload (fault) by tripping a number of components, or result in a complete shut-down of the whole network (which quite often happens with this contingency analysis model).

Thus, the consequences of a violation of the operating criteria are usually severe. For good performance of the scenario selection method, it is important to group together scenarios with similar consequences, i.e., similar amount of load shedding. After the CFM analysis, each load point will either be disconnected or have the same load as in the initial operating scenario, i.e., the load shedding is done in discrete steps. In terms of finding similar operating scenarios, it is more important to know if a certain load point is disconnected or not in the post analysis state than to know the initial load, as disconnecting the loads is the main cause of dissimilarity (variation) between the operating scenarios in this case. Thus, as indicated by the results in chapter 5.3, the initial power flow

in the system is a good basis for similarity, as it indicates how high the load at the transmission system components are, and thus again indicates the post forced outage situation (when CFM is used for the contingency analysis).

6.1.2 Scaling and Weights

As seen in chapter 5.5, the scenario selection method gives better results at the load points with a large load (and high unreliability), compared to the load points with a lower load. The reason, as previously discussed, is that the scenario selection method does not put as much weight on the low load points.

A scaling of the features onto $[0, 1]$ (before applying the clustering algorithm) gives all features the same weight in the clustering process. However, this results in overall worse performance of the scenario selection method, as in fact the larger loads are usually the driving force behind the overall system (un)reliability.

6.1.3 General Comments

The power injections have been used as features, with good results, when the scenario selection method has been used in combination with a rule based contingency analysis model as well, see publication B. The power injections provide information regarding where in the system large loads are, where the generation is located, and thus indicate how power is transferred across the system. The geographical placements of large loads are especially important in terms of transmission system reliability. This feature set is used, with success, when classification and clustering algorithms are used in other power system reliability studies [35, 38].

As mentioned, feature selection is a case dependent process, and ideally the feature selection should be customised to suit the analysis to get optimal results. However, the power injections give good results when used as features for scenario selection, and is the best general recommendation.

To use the power injections as features can lead to an extensive number of features for large systems. Correlation analysis and projection methods, e.g., principal component analysis or multidimensional scaling [40], can in these situations be used to reduce the dimensionality of the problem.

In the power market analysis in this work, the ATCs were set according to the $N - 1$ criterion, which gives quite high congestion costs [23], and it puts quite heavy restrictions on the market clearing process (in EMPS-NC). Thus, the variability in the sample of operating scenarios is limited. If the

ATCs are defined according to some other criteria, e.g., a probabilistic security criterion [57], this can possibly lead to a larger variability in the sample of operating scenarios. If MRM is used for the consequence analysis, the power injections should still be a good feature set, with respect to application of the scenario selection method, as the consequence analysis still aims at minimising the deviation from the initial operating scenario. However, the number of groups k might have to be increased to keep the error in the reliability indices within a reasonable level.

6.2 The Groups of Operating Scenarios

There is a great variety of unsupervised learning algorithms that can be used to cluster the operating scenarios, as discussed in chapter 4.4. Some techniques exist for determining the number of clusters k in a dataset, as discussed in chapter 4.5, but these methods are usually only suitable (and successful) when applied to datasets with a few number of clusters.

Four of the unsupervised learning algorithms were, in the previous chapter, compared with each other, with the feature selection process, and with respect to estimating k. Based on these comparisons, some general remarks, with respect to application of the scenario selection method, are given below:

- Feature selection is much more important than the choice of clustering algorithm. (For instance, compare the results in Figure 5.6 and Figure 5.11)

- None of the algorithms gave any conclusive evidence in favour of a clear clustering structure among the operating scenarios generated by EMPS-NC. For instance, the W_k (within cluster dispersion) index in Figure 5.12 gives no indication of the number of natural groups. In addition, the IGMM results in fairly poor accuracy in the scenario selection indices. If there, in fact, are as few as $50 - 100$ natural groups in the data set (as indicated by the IGMM analysis), this is not a good enough representation to be used as a representative set. (At least when only the medoid of each group is analysed.) Thus, the problem of finding similar operating scenarios is interpreted and treated as a segmentation problem.

- In a segmentation problem, there is no objective method for determining k. However, based on the results in the previous chapter, setting $k \approx 0.1 \cdot n$ (where n is the total number of operating scenarios generated by EMPS-NC) seems to ensure that all indices are within reasonable accuracy. If

only system indices are of interest, a lower value of k might be chosen. To treat the scenario selection problem as a segmentation problem in this way gives good results.

6.2.1 A Segmentation Problem?

It is not too surprising that the problem of finding groups of similar operating scenarios resembles, and most likely is, a segmentation problem, as the power market behaves in a continuous way.

In a segmentation problem, the goal of the clustering algorithm is to group together operating scenarios which are very similar. An agglomerative clustering algorithm, with complete linkage, tends to produce small and compact clusters [40, 53], and is therefore thought of as an appropriate algorithm for this problem. Results from applications of the scenario selection method supports this hypothesis, as agglomerative clustering with complete linkage has been, overall, the algorithm with the best results when applied for scenario selection.

The k-means algorithm tend to produce clusters of equal size, and thus split the sample space into approximately equally heavy regions. This algorithm also performs well as a clustering algorithm in the scenario selection method. The more advanced algorithms, SOM and IGMM, gave no better results than the more basic KM and ACCL algorithms, which support the hypothesis that this is a segmentation problem.

As the clustering problem in the scenario selection process is interpreted as a segmentation problem, it is hard (impossible) to define an objective criteria to determine k. Thus, choosing the number of groups k by setting k at a value of about 10% of the total number of scenarios has been a good compromise between reducing the number of scenarios, while maintaining the variation in the sample of operating scenarios. This reduces the computational requirements of the reliability assessment by about 90%.

As seen in chapter 5.5, setting $k \approx 0.1 \cdot n$ tends to produce scenario selection indices within a 5-10% range of the target values, and this has been referred to as a reasonable error range. The annualised system EENS per operating scenario is in the range of 0 [MWh/year] to about 2500 [MWh/year] for the four-area test network, as seen in Figure 5.17. For Western Norway, the annualised system EENS per operating scenario is in the range of 50 [MWh/year] to about 5000 [MWh/year]. Considering the large range of this index, and the fact that there is no clear clustering structure in the data, some error in the scenario selection indices is to be expected. Thus, a 90% data reduction, with scenario selection indices within a 5-10% range of the target values, is thought of as

a reasonable and acceptable solution. In addition, a 5-10% error is nothing compared to the error when applying the rule-based methods for data reduciton, as seen in Figure 5.20.

6.2.2 Group Structures

The KM tends to produce clusters of equal size, see Figure 5.19. Thus, the sample of operating scenarios is split into groups of approximately equal size. The ACCL tend to produce a few large groups, and many smaller ones, see Figure 5.19.

For scenario selection, if the variation in the reliability indices from scenario to scenario is not too large, it might be most suitable to produce equally sized groups of operating scenarios. However, if some scenarios are very unreliable, and most are fairly reliable, it is desirable to keep the highly unreliable scenarios in smaller groups, and have large groups of reliable operating scenarios. In this case, the ACCL is more likely to keep the unreliable scenarios in their own groups, and produce larger groups of reliable operating scenarios, compared to the KM algorithm.

It remains to be tested how the scenario selection method works in combination with other power market analysis set-ups than the one defined in chapter 2.3.1.

6.2.3 Solution Stability and Reliability Indices

If MRM with AC power flow is used for the contingency analysis, it might not be possible to find a solution of the power flow problem for all contingencies due to convergence problems. In terms of calculating the reliability indices, these convergence problems pose a difficult problem. If it is assumed that non-convergence leads to a total black-out, the consequences of these cases are typically overestimated, and the forced outage combinations leading to non-convergence dominates the reliability indices. In addition, seen from the scenario selection method point of view, a small change in the initial operating scenario might lead to very large changes in the consequences, e.g., a small increase in a load might result in a total black-out. This makes it harder to apply the scenario selection method.

As seen in the previous chapter, the expected interruption duration is harder to estimate with the scenario selection method than the EENS, due to the discrete nature of the expected interruption duration. However, as the value of the expected interruption duration is usually much smaller than the EENS, this

index is more sensitive to the relative error measure than the EENS. The absolute error, when using the scenario selection method to estimate the expected interruption duration, is within a reasonable range.

6.3 Extreme States: Not There?

In power system reliability assessment, the "high impact low probability" (HILP) events are of special interest, as these events have extreme consequences. In the context of scenario selection, the question is if there exist "extreme" operating (power market) scenarios, or if there are cases where an operating scenario in combination with a given contingency event constitutes a HILP event. As long as EMPS-NC, in combination with the $N-1$ criterion, is used for the power market analysis, the operating scenarios themselves are not HILP events, and it is not likely that an operating scenario can be a contributing factor in the constitution of a HILP event. However, if other security criteria are used for the power market analysis, there might be cases where an operating scenario itself is a contributing factor in the constitution of an HILP event. If so, the operating scenario itself should be considered an outlier, and not be a part of any cluster/group in the representative set.

Outlier detection, in the context of scenario selection, is briefly discussed in publication B. Note that from a mathematical point of view, an outlier can also be a highly reliable operating scenario.

6.4 Very Large System Application

In this work, the reliability assessment has been done on a small test network and on a per subsystem basis for a large network (Western Norway). In this context, the scenario selection method works very well. If the goal of the reliability assessment is to analyse a large system as one, say the whole Nordic power system, this possess several challenges in terms of application of the scenario selection method.

First, there is a question whether or not it is possible to find a set of a few operating scenarios which represent the characteristics of the whole Nordic power system. And are the (main) problems encountered in the reliability assessment caused by local (subsystem) effects, or do the problems have a global characteristic? If the problems are mostly local, it is possible to analyse the subsystems individually as is done in this work, and use the scenario selection method as

suggested. However, if there are global effects, the features need to reflect these global problems in order to truly represent similarity of the operating scenarios. Such an analysis (of the whole Nordic power system) is not attempted in this work.

If the SAMREL SoS analysis framework were to be used for an analysis of another network, e.g., Central Europe, the scenario selection method can still be applied. However, this might require some adjustment of the methods as, e.g., similarity of the operating scenarios might change. So far, the scenario selection method has only been applied (and tested) in an analysis of the Nordic power system.

6.5 The "Black-box" Algorithm

Based on the results of the case studies, the best general guidelines for practical application of the scenario selection method are:

- Features: Power injections
- Clustering algorithm: Agglomerative clustering with complete linkage
- Number of groups: 10% of total number of operating scenarios
- Group representatives: The group medoid

Applying the scenario selection method as a black-box algorithm gives good results, as shown in Figure 6.1 where annual EENS is plotted (per hydro inflow year) for Western Norway and delivery point A in Western Norway. However, as for all applications of learning algorithms, (small) adjustments to fit the algorithm to the problem at hand, are necessary for optimal results. For instance, the choice of features and the choice of k (the number of groups) might be changed according to the objective of the analysis.

As discussed in chapter 3, there is no (general) objective method of verifying a cluster solution. Thus, when the scenario selection method is applied in a SoS analysis of a new power system (subsystem), the method should be applied in accordance with the best practice gained through previous applications of the method, i.e., as the "black-box" algorithm given above. However, as more knowledge about the system under study is gained, the scenario selection method should be customised to give optimal results, e.g., new feature and/or a new number of groups in the representative set might be chosen to get better results.

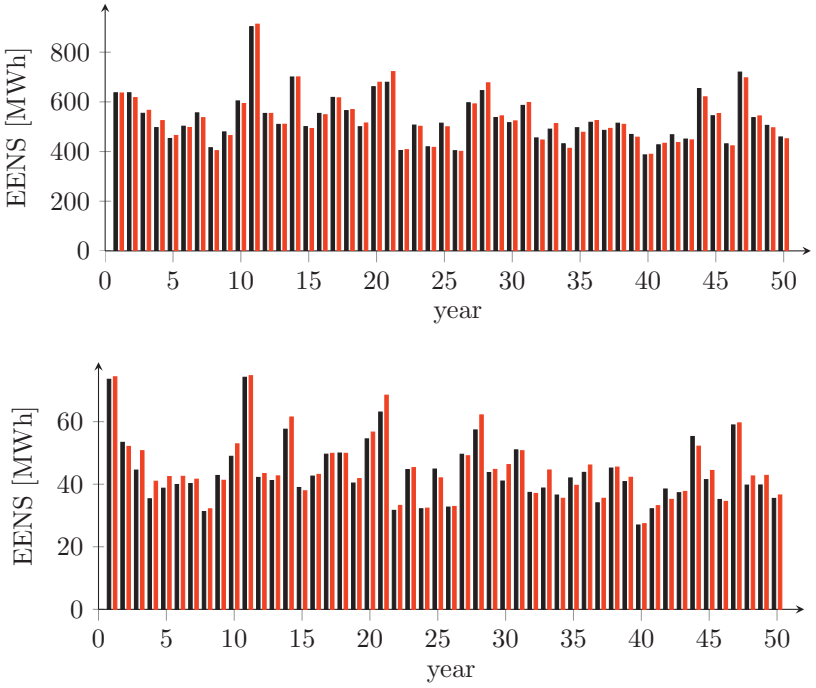


Figure 6.1: Top: Annual EENS for Western Norway. Both target values (black) and "black-box" scenario selection indices (red) are shown. Bottom: Annual EENS for delivery point A in Western Norway.

6.6 Supervised Learning

The application of supervised learning algorithms, in the context of power system reliability assessment, usually takes the form of a classifier. In the SAMREL SoS analysis framework, the classifier can be used to, e.g., classify the operating scenarios as $N-1$ secure or insecure. In terms of reliability indices are the $N-1$ insecure scenarios often the main driving force behind high values of the indices, thus is it important to fully analyse the $N-1$ insecure scenarios. Thus, a classifier can be used to pick out these $N-1$ insecure scenarios. If there are many such insecure operating scenarios, the computational savings are limited, as discussed in chapter 3, and unsupervised learning algorithms are a more efficient data reduction technique in this situation.

Chapter 7

Conclusion and Future Research

This chapter summarises the main findings of the research presented in this thesis, and gives some suggestions for future research.

7.1 Conclusion

The incorporation of operating (power market) scenarios in a reliability assessment has been discussed, both in general and with respect to an analysis of the Nordic power system. This is done by using a power market model to generate future load and generation patterns, and use these as a basis for contingency and reliability analysis.

EMPS-NC is used for the power market analysis, which is a power market model designed for simulation of hydro-thermal power systems, where the market analysis is done by finding the optimal socio-economic dispatch considering, e.g., different hydro inflow scenarios and unit commitment cost. The annual variation in hydro inflow, wind speed, and temperature (in the analysis period) is taken care of by using a sample of 50-75 years as representative for the planning horizon. For each (historic) year, EMPS-NC optimises the operation of the hydro-thermal power system per week by splitting the week into load periods or do the dispatch per hour. The generated load and generation patterns (operating scenarios) represent the daily, weekly, and seasonal variation in load and generation, and are regarded as representative for the load and generation

profile over the whole year.

The operating scenarios are the main input from the power market analysis to the contingency and reliability analysis, and for each operating scenario, the consequences of a set of contingencies are determined. The contingency analysis determines which delivery points experience an interruption during a contingency (if any), and determines the severity of the interruption in the form of, e.g., load shedding at the delivery points. Two different models are used to determine the consequences of the set of contingencies - the minimal rescheduling model and the cascading failure model.

Reliability indices are evaluated per delivery point per operating scenario, by using analytical techniques, based on minimal cuts and approximate techniques. The evaluated indices include the duration of interruptions and the expected energy not supplied. The per operating scenario indices are combined to give annual indices for each of the historic years (included in the EMPS-NC analysis), where system indices are derived from the delivery point indices.

To take all operating scenarios, generated by EMPS-NC, as input to the contingency and reliability analysis results in excessive computation time for the overall (SAMREL) SoS analysis. The scenario selection method is proposed, and used to reduce the overall computation time. The scenario selection method finds groups of similar operating scenarios, and then, for each group, chooses one scenario to represent the group characteristics. The set of chosen scenarios is denoted the representative set, and only this subset of scenarios are used as input to the contingency and reliability analysis. Based on the results of the case studies, the main conclusions (with respect to application of the scenario selection method) are:

- Feature selection, i.e., selection of variables used for quantifying similarity among the operating scenarios, is a case dependent process, and ideally the feature selection should be customised to suit the analysis to get optimal results. However, the power injections give good results when used as features for scenario selection, and is the best general recommendation.

- The problem of finding groups of similar operating scenarios is most likely a segmentation problem. Thus, there is no objective method for determining the number of natural groups k in the set of operating scenarios, and k must be selected based on experience. The results of the case studies indicate that $k \approx 0.1 \cdot n$ is a sensible choice.

- For this segmentation problem, the goal of the clustering algorithm is to group together operating scenarios which are very similar. An agglomer-

ative clustering algorithm, with complete linkage, tends to produce small and compact groups, and is therefore thought of as an appropriate algorithm for this problem. The results of the case studies support this assumption, as agglomerative clustering with complete linkage is, overall, the algorithm with the best results when applied for scenario selection. The k-means algorithm gives similar results as the agglomerative clustering algorithm, while more advanced clustering algorithms perform at the same level, or worse, than the agglomerative clustering algorithm when applied for scenario selection.

- The scenario selection method can reduce the computation time of the SoS analysis by about 90% (by setting $k \approx 0.1 \cdot n$), and simultaneously keep the error in the annual reliability indices within a 5-10% margin. This is thought of as an appropriate trade-off between reduced computation time and maintaining the accuracy of the analysis.

The contingency analysis, in the SAMREL SoS analysis of Western Norway, with AC power flow, takes about 5-7 weeks in its current version (and implementation). Although the analysis is done off-line (and for long-term planning), such a long computation time put severe limitations on the analysis in terms of practical applications. A 90% reduction in computation time means that the analysis can be done in about 2-3 days, which, although still a substantial computation time, is a considerable improvement compared to 6 weeks.

As more details are included in the power market analysis, e.g., hourly load data and/or several wind speed scenarios, the computation time of the SAMREL SoS analysis increases rapidly (as more operating scenarios are generated by EMPS-NC). Thus, a combination of the scenario selection method and a more efficient implementation of the contingency analysis are necessary to keep the computation time of the SoS analysis within reasonable bounds as the number of operating scenarios, generated by EMPS-NC, increases further.

The scenario selection method results in reliability indices much closer to the target values, than indices based on the rule-based analysis methods, i.e., only analysing high load scenarios. Thus, the scenario selection method is a much better method of objectively picking a subset of scenarios than purely basing the selection on the total system load.

7.2 Future Work

In the power market analysis (with EMPS-NC), only ATCs (market clearing restrictions) set according to the $N-1$ criterion is considered. A power market analysis based on, e.g., a probabilistic security criterion, can yield larger variability in the load and generation patterns (as discussed in previous chapters). This might require some adjustments in the application of the scenario selection method, e.g., increasing the number of groups in the representative set to keep in estimate reliability indices within a reasonable range. The effect of changing the market criteria, and/or market model itself, remains to be tested.

The results in chapter 5.8 show that the scenario selection method can handle sets of operating scenarios where part of the generation mix consists of wind power. However, in systems with a high penetration of wind power, the SoS analysis set-up described in this thesis should be adjusted to take into account, among others, the uncertainty of the wind power forecast. In terms of application of the scenario selection method in such an analysis setting, one concern is whether or not two operating scenarios with quite similar initial generation and load patterns actually can be quite dissimilar due to, e.g., low and very high uncertainty in the wind power forecast for the two given operating scenarios. In such a setting, a method of incorporating the wind power uncertainty into the feature set might be necessary.

The guidelines for application of the scenario selection method are not carved in stone and are subject to changes as more experience is gained through practical applications of the method. Future applications of the SAMREL SoS analysis framework, in combination with the scenario selection method, might be, e.g., grid expansion studies and studies of large scale wind power integration.

Figure 7.1 illustrates how operating scenarios are incorporated in the reliability analysis. Market zones and ATCs are included in the power market analysis, and the operating scenarios are used as a basis for a reliability assessment. The dotted path indicates that the results of the reliability assessment ideally should be transferred back to the power market analysis module, and be used to, e.g., indicate how to update the ATCs (if necessary). The reliability assessment will typically be used as an objective method for comparison of the different ATCs and/or market zone definitions. In such a setting, the scenario selection method can be especially useful as the analysis loop in Figure 7.1 has to be done several times (one for each power market analysis). For such an analysis, EMPS-NC will be used to generate operating scenarios for the different market regulations, a reliability assessment done for each EMPS-NC analysis, and the scenario selection method will be used as an objective method for reducing the

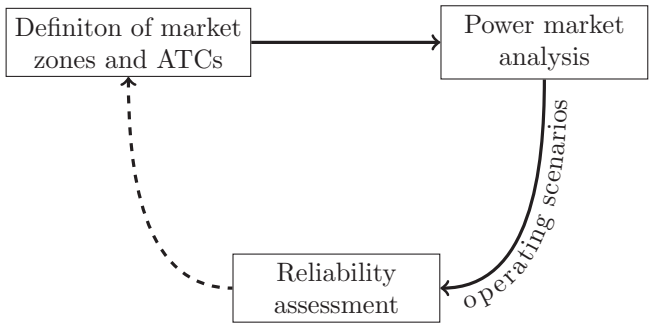


Figure 7.1: Illustration of how operating (power market) scenarios are incorporated in the reliability assessment. Market zones and ATCs are included in the power market analysis, and the operating scenarios are used as a basis for a reliability assessment.

computation time of each analysis loop. The feedback of information from the reliability assessment to the power market model is not considered in this work.

Even though the scenario selection is discussed in the context of a reliability assessment of the Nordic power system, the methodology is applicable in all settings where a (large) set of generation and load patterns is used as input to a reliability assessment. The generation and load patterns can, e.g., be generated by a power market model or be generated from historic data. As discussed, the parameters (e.g., the features and/or number of groups) might have to be changed for optimal results. The application of the scenario selection method in such analysis settings have not been attempted.

7.3 Parallel Computing

In steady-state analysis, the power flow computations for different load and generation patterns and contingencies are independent of each other. Thus, parallel computing can be used to speed up the analysis when, e.g., several contingencies are to be analysed, as these contingencies can be analysed simultaneously (by applying parallel computing) instead of sequentially. This can severely reduce the computation time of the contingency analysis in the SAMREL SoS analysis. However, as of today, it is not possible to run (optimal) power flow calculation in parallel with PSS/E, and thus there is a need for the scenario selection method when this framework is to be used in practice. Convergence problems relating to the AC power flow solution cannot be solved by parallel computing alone.

Bibliography

[1] R. Billinton and R. N. Allan, *Reliability Evaluation of Engineering Systems*, 2nd ed. New York: Plenum, 1992.

[2] ——, *Reliability Evaluation of Power Systems*, 2nd ed. New York: Plenum, 1996.

[3] R. Allan and R. Billinton, "Probabilistic assessment of power systems," *Proc. IEEE*, vol. 88, no. 2, pp. 140–162, Feb. 2000.

[4] G. Kjølle and O. Gjerde, "Integrated approach for security of electricity supply analysis," *Int. J. Syst. Assur. Eng. Manag.*, vol. 1, no. 2, pp. 163–169, Jun. 2010.

[5] O. Gjerde, G. Kjølle, L. Warland, M. Korpås, and G. Warland, "Integration of market and network models for security of electricity supply analysis," SINTEF Energy Research, Trondheim, Norway, Tech. Rep. A-6751, 2009.

[6] G. L. Doorman, K. Uhlen, G. H. Kjølle, and E. S. Huse, "Vulnerability analysis of the Nordic power system," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 402–410, Feb. 2006.

[7] O. Wolfgang, A. Haugstad, B. Mo, A. Gjelsvik, I. Wangensteen, and G. Doorman, "Hydro reservoir handling in Norway before and after deregulation," *Energy*, vol. 34, no. 10, pp. 1642–1651, Oct. 2009.

[8] A. Helseth, G. Warland, and B. Mo, "A hydrothermal market model for simulation of area prices including detailed network analyses," *Euro. Trans. Electr. Power*, 2012.

[9] G. Kjølle and O. Gjerde, "The OPAL methodology for reliability analysis of power systems," SINTEF Energy Research, Trondheim, Norway, Tech. Rep. A-7175, 2012.

[10] R. Billinton and G. Lian, "Composite power system health analysis using a security constrained adequacy evaluation procedure," *IEEE Trans. Power Syst.*, vol. 9, no. 2, pp. 936–941, Feb. 1994.

[11] W. Wangdee and R. Billinton, "Bulk electric system well-being analysis using sequential Monte Carlo simulation," *IEEE Trans. Power Syst.*, vol. 21, no. 1, pp. 188–103, Feb. 2006.

[12] R. Billinton and G. Lian, "Composite power system health analysis using a security constrained adequacy evaluation procedure," *IEEE Trans. Power Syst.*, vol. 9, no. 2, pp. 936–941, May 1994.

[13] A. M. L. da Silva, L. C. de Resende, L. A. da Fonseca Manso, and R. Billinton, "Well-being analysis for composite generation and transmission systems," *IEEE Trans. Power Syst.*, vol. 10, no. 3, pp. 1763–1770, Jun. 2009.

[14] M. T. Schilling, R. Billinton, and M. G. dos Santos, "Bibliography on power systems probabilistic security analysis 1968-2008," *International Journal of Emerging Electric Power Systems*, vol. 19, no. 4, Nov. 2004.

[15] CIGRE Working Group C4.601, "Review of the current status of tools and techniques for risk-based and probabilistic planning in power systems," CIGRE, Tech. Rep. 434, Oct. 2010.

[16] R. Billinton and W. Li, *Reliability Assessment of Electric Power Systems Using Monte Carlo Methods.* New York: Plenum, 1994.

[17] A. Sankarakrishnan and R. Billinton, "Sequential Monte Carlo simulation for composite power system reliability analysis with time varying loads," *IEEE Trans. Power Syst.*, vol. 10, no. 3, pp. 1540–1545, Aug. 1995.

[18] M. V. Pereira and N. J. Balu, "Composite generation/transmission reliability evaluation," *Proc. IEEE*, vol. 80, no. 4, pp. 470–491, Apr. 1992.

[19] R. Allan and R. Billinton, "Power system reliability and its assessment. 2. Composite generation and transmission systems," *Power Engineering Journal*, vol. 6, no. 6, pp. 291–297, Nov. 1992.

[20] A. Rei and M. Schilling, "Reliability assessment of the Brazilian power system using enumeration and Monte Carlo," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1480–1487, Aug. 2008.

[21] K. Samdal, G. Kjølle, O. Gjerde, J. Heggset, and A. Holen, "Requirement specification for reliability analysis in meshed power networks," SINTEF Energy Research, Trondheim, Norway, Tech. Rep. A-6429, 2006.

[22] A. Helseth, G. Warland, and B. Mo, "Long-term hydro-thermal scheduling including network constraints," in *Proc. 7th Int. Conf. on the European Energy Market (EEM)*, Madrid, Spain, Jun. 2010, pp. 1–6.

[23] N. Flatabø, G. Doorman, O. S. Grande, H. Randen, and I. Wangensteen, "Experience with the Nord Pool design and implementation," *IEEE Trans. Power Syst.*, vol. 18, no. 2, pp. 541–547, May 2003.

[24] O. B. Fosso, A. Gjelsvik, A. Haugstad, B. Mo, and I. Wangensteen, "Generation scheduling in a deregulated system. The Norwegian case," *IEEE Trans. Power Syst.*, vol. 14, no. 1, pp. 75–81, Feb. 1999.

[25] R. Billinton, E. Chan, and G. Wacker, "Probability distribution approach to describe customer costs due to electric supply interruptions," *IEE Proc. Generation, Transmission and Distribution*, vol. 141, no. 6, pp. 594–598, Nov. 1994.

[26] R. Ghajar, R. Billinton, and E. Chan, "Distributed nature of residential customer outage costs," *IEEE Trans. Power Syst.*, vol. 11, no. 3, pp. 1236–1244, Aug. 1996.

[27] G. H. Kjølle, K. Samdal, B. Singh, and O. Kvitastein, "Customer costs related to interruptions and voltage problems: Methodology and results," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 1030–1038, Aug. 2008.

[28] A. M. L. da Silva, L. A. da Fonseca Manso, J. C. de Oliveira Mello, and R. Billinton, "Pseudo-chronological simulation for composite reliability analysis with time varying loads," *IEEE Trans. Power Syst.*, vol. 15, no. 1, pp. 73–80, Feb. 2000.

[29] J. C. de Oliveira Mello, M. Pereira, and A. M. L. da Silva, "Evaluation of reliability worth in composite systems based on pseudo-sequential Monte Carlo simulation," *IEEE Trans. Power Syst.*, vol. 9, no. 3, pp. 1318–1326, Aug. 1994.

[30] G. Oliveira, M. Pereira, and S. Cunha, "A technique for reducing computational effort in Monte-Carlo based composite reliability evaluation," *IEEE Trans. Power Syst.*, vol. 4, no. 4, pp. 1309–1315, Nov. 1989.

[31] R. Billinton and A. Jonnavithula, "Variance reduction techniques for use with sequential Monte Carlo simulation in bulk power system reliability evaluation," in *Proc. Canadian Conf. on Elect. and Comp. Eng.*, vol. 1, Calgary, Canada, May 1996, pp. 416–419.

[32] C. Singh and J. Mitra, "Composite system reliability evaluation using state space pruning," *IEEE Trans. Power Syst.*, vol. 12, no. 1, pp. 471–479, Feb. 1997.

[33] A. M. L. da Silva., L. C. de Resende, L. A. da Fonseca Manso, and V. Miranda, "Composite reliability assessment based on Monte Carlo simulation and artificial neural networks," *IEEE Trans. Power Syst.*, vol. 22, no. 3, pp. 1202–1209, Aug. 2007.

[34] ——, "Well-being analysis for composite generation and transmission systems based on pattern recognition techniques," *IET Gener. Trans. Distrib.*, vol. 2, no. 2, pp. 202–208, Mar. 2008.

[35] H. Kim and C. Singh, "Power system probabilistic security assessment using Bayes classifier," *Elect. Power Syst. Res.*, vol. 74, no. 1, pp. 157–165, Apr. 2005.

[36] N. M. Pindoriya, P. J. Jirutitijaroen, P. Jirutitijaroen, and C. Singh, "Composite reliability evaluation using Monte Carlo simulation and least squares support vector classifier," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2483–2490, Nov. 2011.

[37] R. C. Green, L. Wang, and M. Alam, "Composite power system reliability evaluation using support vector machines on a multicore platform," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, California, USA, Jul./Aug. 2011, pp. 2586–2592.

[38] C. Singh, X. Luo, and H. Kim, "Power system adequacy and security calculations using Monte Carlo simulation incorporating intelligent system methodology," in *Proc. 9th Int. Conf. on Probabilistic Methods Applied to Power Systems (PMAPS)*, Stockholm, Sweden, Jun. 2006, pp. 1–9.

[39] X. Luo, C. Singh, and A. Patton, "Power system reliability evaluation using self organizing map," in *Proc IEEE Power Eng. Soc. Winter Meeting*, vol. 2, Jan. 2000, pp. 1103–1108.

[40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements Of Statistical Learning*, 2nd ed. Springer-Verlag, 2008.

[41] C. M. Bishop, *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press, 1995.

[42] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, Inc, 2000.

[43] C. E. Rasmussen, "The infinite gaussian mixture model," in *Proc. Advances in Neural Information Processing Systems 12 (NIPS)*, Denver, Colorado, USA, Nov./Dec. 1999, pp. 554–560.

[44] D. Görür and C. E. Rasmussen, "Dirichlet process Gaussian mixture models: Choice of the base distribution," *Journal of Computer Science and Technology*, vol. 25, pp. 615–626, Jul. 2010.

[45] H. Kile and K. Uhlen, "Averaging operating states with infinite mixtures in reliability analysis of transmission networks," in *Proc. 12th Int. Conf. on Probabilistic Methods Applied to Power Systems (PMAPS)*, Istanbul, Turkey, Jun. 2012, pp. 670–675.

[46] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set with the gap statistic," *Journal of Royal Statistical Society: Series B*, vol. 63, no. 2, pp. 411–423, Jan. 2001.

[47] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[48] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, Dec. 2007.

[49] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Berlin: Springer, 1989.

[50] R. Xu and D. Wunsch, *Clustering.* Oxford Wiley, 2009.

[51] N. Gröwe-Kuska, H. Heitsch, and W. Romisch, "Scenario reduction and scenario tree construction for power management problems," in *Proc. IEEE Power Tech Conf.*, Bologna, Italy, Jun. 2003, pp. 1–7.

[52] X.-B. Li, "Data reduction via adaptive sampling," *Communications in information and systems*, vol. 2, no. 1, pp. 53–68, Jun. 2002.

[53] A. Jain, R. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.

[54] Y. W. Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning.* Springer, 2010.

[55] H. Kile and K. Uhlen, "Data reduction via clustering and averaging for contingency and reliability analysis," *Int J Elec. Power Energy Sys.*, vol. 43, no. 1, pp. 1435–1442, 2012.

[56] T. Aigner and T. Gjendeal, "Modelling wind power production based on numerical predictionmodels and wind speed measurements," in *17th Power Systems Computation Conference*, Stockholm, Sweden, 2011.

[57] K. Uhlen, G. H. Kjølle, G. G. Løvås, and Ø. Breidablikk, "A probabilistic criterion for determination of power transfer limits in a deregulated environment," in *CIGRE Session 2000*, Paris, France, 2000, pp. 1–6.

[58] A. Gelman, J. B. Carlin, and H. S. Stern, *Bayesian Data Analysis.* Chapman & Hall/CRC, 2004.

[59] R. Neal, "Markov Chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

Appendix A

The Infinite Gaussian Mixture Model

The infinite Gaussian mixture model (IGMM) is introduced in [43], which is a detailed derivation and discussion of the univariate IGMM. In this appendix, the multivariate extension of the IGMM is derived, based on the guidelines and suggestions given in [43]. The univariate IGMM follows directly from setting $p = 1$ in the equations in this appendix.

Reference [44] contains a more detailed discussion of the IGMM, where the use of different conjugate distributions are discussed. The parameterisations (of the densities) used in this appendix follow the ones given in [58].

A.1 The General Model

The data are assumed to be multivariate Gaussian distributed according to a mixture model with k components. This may be written:

$$p(\mathbf{y}_i|\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_k, \mathbf{S}_1, ...\mathbf{S}_k, \pi_1, ..., \pi_k) = \sum_{j=1}^{k} \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{S}_j^{-1}). \tag{A.1}$$

Alternatively, if the class label c_i is known for observation $\mathbf{y}_i$,

$$\mathbf{y}_i|c_i \sim \mathcal{N}(\boldsymbol{\mu}_{c_i}, \boldsymbol{S}_{c_i}^{-1}). \tag{A.2}$$

The complete model, with hyperparameters, is shown in Figure A.1.

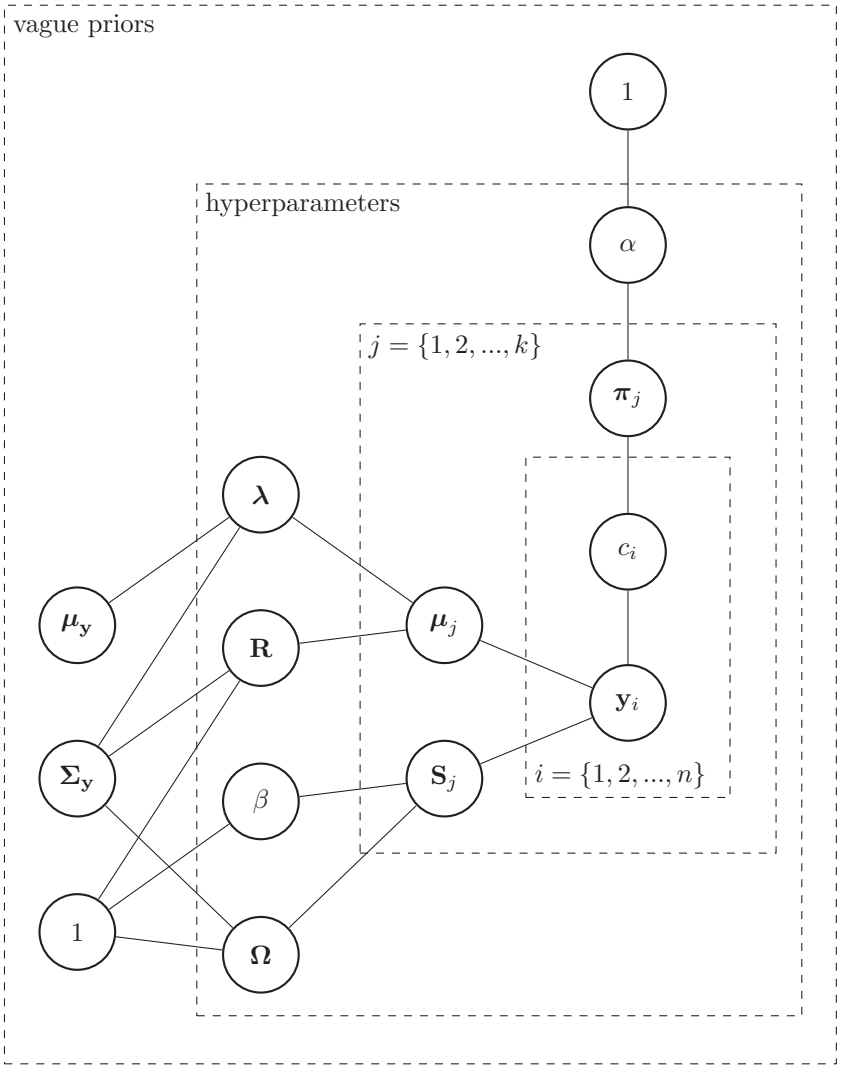


Figure A.1: Illustration of the multivariate infinite Gaussian mixture model.

A.1.1 Component Means

The components means are given Gaussian priors as

$$p(\boldsymbol{\mu}_j|\boldsymbol{\lambda}, \mathbf{R}) \sim \mathcal{N}(\boldsymbol{\lambda}, \mathbf{R}^{-1}). \tag{A.3}$$

The hyperparameters are given vague priors as

$$p(\boldsymbol{\lambda}|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \tag{A.4}$$

and

$$p(\mathbf{R}|\nu_0, \boldsymbol{\Sigma}_y) \sim \mathcal{W}(p,\ (p\boldsymbol{\Sigma}_y)^{-1}). \tag{A.5}$$

where p is the dimension of the observations. To use of the sample mean and sample precision in these equations is equivalent to scaling/normalising the data before the analysis.

The full conditional of $\boldsymbol{\mu}_j$ follows from the prior (A.3) and a likelihood based on (A.1).

$$\begin{aligned}
p(\boldsymbol{\mu}_j|\mathbf{c}, \mathbf{Y}, \mathbf{S}_j, \boldsymbol{\lambda}, \mathbf{R}) &\propto p(\boldsymbol{\mu}_j|\boldsymbol{\lambda}, \mathbf{R}^{-1}) \prod_{i:c_i=j} \exp\left\{\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_j)^T\mathbf{S}_j(\mathbf{y}_i - \boldsymbol{\mu}_j)\right\} \\
&= \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_j - \boldsymbol{\lambda})^T\mathbf{R}(\boldsymbol{\mu}_j - \boldsymbol{\lambda}) - \frac{1}{2}\sum_{i:c_i=j}(\mathbf{y}_i - \boldsymbol{\mu}_j)^T\mathbf{S}_j(\mathbf{y}_i - \boldsymbol{\mu}_j)\right\} \\
&\sim \mathcal{N}\left((\mathbf{R} + n_j\mathbf{S}_j)^{-1}(\mathbf{R}\boldsymbol{\lambda} + n_j\mathbf{S}_j\overline{\mathbf{y}}_j),\ (\mathbf{R} + n_j\mathbf{S}_j)^{-1}\right)
\end{aligned}$$

The full conditionals of the hyperparameters are found by combining the priors (A.4) and (A.5), with a likelihood from (A.3). The full conditional of $\boldsymbol{\lambda}$ then follows as

$$\begin{aligned}
p(\boldsymbol{\lambda}|\boldsymbol{\mu}_1, \ldots\boldsymbol{\mu}_k, \mathbf{R}) &\propto p(\boldsymbol{\lambda}|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \prod_{j=1}^{k} p(\boldsymbol{\mu}_j|\boldsymbol{\lambda}, \mathbf{R}^{-1}) \\
&\sim \mathcal{N}\left((\boldsymbol{\Sigma}_y^{-1} + k\mathbf{R})^{-1}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\mu}_y + \mathbf{R}\sum_{j=1}^{k}\boldsymbol{\mu}_j),\ (\boldsymbol{\Sigma}_y^{-1} + k\mathbf{R})^{-1}\right) \\
&\sim \mathcal{N}\left((\boldsymbol{\Sigma}_y^{-1} + k\mathbf{R})^{-1}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\mu}_y + k\mathbf{R}\overline{\boldsymbol{\mu}}),\ (\boldsymbol{\Sigma}_y^{-1} + k\mathbf{R})^{-1}\right),
\end{aligned}$$

and the full conditional of $\mathbf{R}$ is

$$p(\mathbf{R}|\boldsymbol{\mu}_1, ...\boldsymbol{\mu}_k, \boldsymbol{\lambda}) \sim \mathcal{W}\left(D+k,\ \left[D\boldsymbol{\Sigma}_y + \sum_{j=1}^{k}(\boldsymbol{\mu}_j - \boldsymbol{\lambda})(\boldsymbol{\mu}_j - \boldsymbol{\lambda})^T\right]^{-1}\right),$$

see section A.3 for details.

A.1.2 Component Precision

The component precisions are given Wishart priors, as

$$p(\mathbf{S}_j|\beta, \boldsymbol{\Omega}) \sim \mathcal{W}(\beta, (\beta\boldsymbol{\Omega})^{-1}). \tag{A.6}$$

Again, the hyperparameters are given vague priors. β is given a Gamma prior

$$p((\beta - D + 1)^{-1}) = \mathcal{G}_R(1, 1/D), \tag{A.7}$$

where subscript R means the parametrisation of the Gamma distribution from [43]. This implies that

$$p(\beta - D + 1) \propto (\beta - D + 1)^{-3/2} \exp\left\{\frac{-D}{2(\beta - D + 1)}\right\}. \tag{A.8}$$

$\boldsymbol{\Omega}$ is given a Wishart prior

$$p(\boldsymbol{\Omega}) = \mathcal{W}(D, D^{-1}\boldsymbol{\Sigma}_y). \tag{A.9}$$

The full conditional of $\mathbf{S}_j$ follows from the prior (A.6) and (A.1) as the likelihood.

$$\begin{aligned}
p(\mathbf{S}_j|\mathbf{c}, \mathbf{Y}, \boldsymbol{\mu}_j, \beta, \boldsymbol{\Omega}) &\propto p(\mathbf{S}_j|\beta, \boldsymbol{\Omega}) \prod_{i:c_i=j} |\mathbf{S}_j|^{1/2} \exp\left\{\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_j)^T\mathbf{S}_j(\mathbf{y}_i - \boldsymbol{\mu}_j)\right\} \\
&\propto |\mathbf{S}_j|^{(\beta-k-1)/2} \exp\left\{-\frac{1}{2}tr(\beta\boldsymbol{\Omega}\mathbf{S}_j)\right\} \\
&\quad \times \prod_{i:c_i=j} |\mathbf{S}_j|^{1/2} \exp\left\{\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_j)^T\mathbf{S}_j(\mathbf{y}_i - \boldsymbol{\mu}_j)\right\} \\
&\sim \mathcal{W}\left(\beta + n_j,\ \left[\beta\boldsymbol{\Omega} + \sum_{i:c_i=j}(\mathbf{y}_i - \boldsymbol{\mu}_j)(\mathbf{y}_i - \boldsymbol{\mu}_j)^T\right]^{-1}\right),
\end{aligned}$$

see appendix A.3 for details. To get the full conditionals of $\mathbf{\Omega}$ and β, the likelihood based on (A.6) is used.

$$
\begin{aligned}
L(\mathbf{S}|\mathbf{\Omega},\beta) &= \prod_{j=1}^{k} p(\mathbf{S}_j|\mathbf{\Omega},\beta) \\
&= \prod_{j=1}^{k} \left(2^{\beta d/2} \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left(\frac{\beta+1-i}{2}\right) \right)^{-1} |\beta\mathbf{\Omega}|^{\beta/2} |\mathbf{S}_j|^{(\beta-d-1)/2} \\
&\quad \times \exp\left\{ -\frac{\beta}{2} tr(\mathbf{\Omega}\mathbf{S}_j) \right\} \\
&= \prod_{j=1}^{k} \left(2^{\beta d/2} \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left(\frac{\beta+1-i}{2}\right) \right)^{-1} (\beta^d |\mathbf{\Omega}|)^{\beta/2} |\mathbf{S}_j|^{(\beta-d-1)/2} \\
&\quad \times \exp\left\{ -\frac{\beta}{2} tr(\mathbf{\Omega}\mathbf{S}_j) \right\} \\
&\propto \prod_{j=1}^{k} \left(2^{\beta d/2} \prod_{i=1}^{d} \Gamma\left(\frac{\beta+1-i}{2}\right) \right)^{-1} \beta^{(\beta d/2)} |\mathbf{\Omega}|^{\beta/2} |\mathbf{S}_j|^{\beta/2} \\
&\quad \times \exp\left\{ -\frac{\beta}{2} tr(\mathbf{\Omega}\mathbf{S}_j) \right\}
\end{aligned}
$$

The full conditional of $\mathbf{\Omega}$ is then given by

$$
\begin{aligned}
p(\mathbf{\Omega}|\mathbf{S},\beta) &= p(\mathbf{\Omega}) \prod_{j=1}^{k} p(S_j|\beta,\mathbf{\Omega}) \\
&\propto |\mathbf{\Omega}|^{(D-k-1)/2} \exp\left\{ -\frac{D}{2} tr(\mathbf{\Sigma}_y^{-1}\mathbf{\Omega}) \right\} \prod_{j=1}^{k} |\mathbf{\Omega}|^{\beta/2} \exp\left\{ -\frac{\beta}{2} tr(\mathbf{\Omega}\mathbf{S}_j) \right\} \\
&\sim \mathcal{W}\left(D + k\beta, \left[D\mathbf{\Sigma}_y^{-1} + \beta \sum_{j=1}^{k} \mathbf{S}_j \right]^{-1} \right).
\end{aligned}
$$

The full conditional of β is

$$\begin{aligned}
p(\beta|\mathbf{S},\mathbf{\Omega}) &\propto (\beta-D+1)^{-3/2}\exp\left\{\frac{-D}{2(\beta-D+1)}\right\} \\
&\times\prod_{j=1}^{k}\left(2^{\beta d/2}\prod_{i=1}^{d}\Gamma\left(\frac{\beta+1-i}{2}\right)\right)^{-1}\beta^{(\beta d/2)}|\mathbf{\Omega}|^{\beta/2}|\mathbf{S}_j|^{\beta/2} \\
&\times\exp\left\{-\frac{\beta}{2}tr(\mathbf{\Omega}\mathbf{S}_j)\right\} \\
&\propto (\beta-D+1)^{-3/2}\exp\left\{\frac{-D}{2(\beta-D+1)}\right\} \\
&\times\left(2^{\beta d/2}\prod_{i=1}^{d}\Gamma\left(\frac{\beta+1-i}{2}\right)\right)^{-k}\beta^{(k\beta d/2)}|\mathbf{\Omega}|^{k\beta/2}\left(\prod_{i=1}^{k}|\mathbf{S}_j|^{\beta/2}\right) \\
&\times\exp\left\{-\frac{\beta}{2}\sum_{j=1}^{k}tr(\mathbf{\Omega}\mathbf{S}_j)\right\}
\end{aligned}$$

A.1.3 Class Labels

The Finite Case

The mixing proportions in (A.1) are given a symmetric Dirichlet prior.

$$p(\pi_1,...,\pi_k|\alpha)\sim\mathcal{D}(\alpha/k,...,\alpha/k)=\frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k}\prod_{j=1}^{k}\pi_j^{\alpha/k-1} \tag{A.10}$$

Given $\boldsymbol{\pi}$, the class labels are independent, and distributed according to a multinomial distribution. The joint distribution is then

$$p(c_1,...,c_n|\pi_1,...,\pi_k)=\prod_{i=1}^{n}\prod_{j=1}^{k}\pi_j^{\delta(c_i,j)}=\prod_{j=1}^{k}\pi_j^{n_j}, \tag{A.11}$$

where $\delta(c_i, j) = 1$ if $c_i = j$, and 0 otherwise, and $n_j = \sum_{i=1}^{n} \delta(c_i, j)$. Now it is possible to integrate out the mixing proportions $\boldsymbol{\pi}$.

$$
\begin{aligned}
p(c_1, ..., c_n|\alpha) &= \int p(c_1, ..., c_n, \pi_1, ..., \pi_k|\alpha) d\pi_1 ... d\pi_k \\
&= \int p(c_1, ..., c_n|\pi_1, ..., \pi_k) p(\pi_1, ..., \pi_k|\alpha) d\pi_1 ... d\pi_k \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \int \prod_{j=1}^{k} \pi^{n_j + \alpha/k - 1} d\pi_j \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \int \frac{\Gamma\left(\sum_{j=1}^{k} n_j + \alpha/k\right)}{\Gamma\left(\sum_{j=1}^{k} n_j + \alpha/k\right)} \frac{\prod_{j=1}^{k} \Gamma(n_j + \alpha/k)}{\prod_{j=1}^{k} \Gamma(n_j + \alpha/k)} \prod_{j=1}^{k} \pi^{n_j + \alpha/k - 1} d\pi_j \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \frac{\prod_{j=1}^{k} \Gamma(n_j + \alpha/k)}{\Gamma\left(\sum_{j=1}^{k} n_j + \alpha/k\right)} \left[\int \frac{\Gamma\left(\sum_{j=1}^{k} n_j + \alpha/k\right)}{\prod_{j=1}^{k} \Gamma(n_j + \alpha/k)} \prod_{j=1}^{k} \pi^{n_j + \alpha/k - 1} d\pi_j \right] \\
&= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \frac{\prod_{j=1}^{k} \Gamma(n_j + \alpha/k)}{\Gamma\left(\sum_{j=1}^{k} n_j + \alpha/k\right)} = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \frac{\prod_{j=1}^{k} \Gamma(n_j + \alpha/k)}{\Gamma(n + \alpha)} \\
&= \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} \prod_{j=1}^{k} \frac{\Gamma(n_j + \alpha/k)}{\Gamma(\alpha/k)}
\end{aligned}
$$

The integral in the brackets above is 1, since it is the integral over the Dirichlet distribution with parameters $n_j + \alpha/k$. To get the conditional distribution of one c_i, conditioned on the rest of the c's:

$$
\begin{aligned}
p(c_i = j|\mathbf{c}_{-i}, \alpha) &= \frac{p(\mathbf{c}|\alpha)}{p(\mathbf{c}_{-i}|\alpha)} \\
&= \frac{\frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=1}^{k} \frac{\Gamma(n_j + \alpha/k)}{\Gamma(\alpha/k)}}{\frac{\Gamma(\alpha)}{\Gamma(n-1+\alpha)} \prod_{j=1}^{k} \frac{\Gamma(n_{-i,j} + \alpha/k)}{\Gamma(\alpha/k)}} \\
&= \frac{\Gamma(n - 1 + \alpha)}{\Gamma(n + \alpha)} \prod_{j=1}^{k} \frac{\Gamma(n_j + \alpha/k)}{\Gamma(n_{-i,j} + \alpha/k)} \\
&= \frac{n_{-i,j} + \alpha/k}{n - 1 + \alpha},
\end{aligned}
$$

where it is used that

$$\Gamma(n+1) = n\Gamma(n) \Rightarrow \Gamma(n+\alpha) = (n-1+\alpha)\Gamma(n-1+\alpha).$$

The Infinite Limit

Take the limit as $k \to \infty$, which gives

$$p(c_i = j|\mathbf{c}_{-i}, \alpha) = \frac{n_{-i,j}}{n-1+\alpha} \tag{A.12}$$

$$p(c_i \neq c_{i'} \ \forall i \neq i'|\mathbf{c}_{-i}, \alpha) = \frac{\alpha}{n-1+\alpha}. \tag{A.13}$$

The likelihood of the observation $\mathbf{y}_i$ belonging to component j is

$$L(\mathbf{y}_i|\boldsymbol{\mu}_j, \mathbf{S}_j, \mathbf{c}_{-i}) \propto |\mathbf{S}_j|^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_j)^T \mathbf{S}_j (\mathbf{y}_i - \boldsymbol{\mu}_j)\right\}. \tag{A.14}$$

It follows that the conditional posterior of c_i, for components with $n_{-i,j} > 0$ is

$$p(c_i = j|\mathbf{c}_{-i}, \mathbf{y}_i, \boldsymbol{\mu}_j, \mathbf{S}_j, \alpha) \propto p(c_i = j|\mathbf{c}_{-i}, \alpha) L(\mathbf{y}_i|\boldsymbol{\mu}_j, \mathbf{S}_j, \mathbf{c}_{-i}) \tag{A.15}$$

$$\propto \frac{n_{-i,j}}{n-1+\alpha} |\mathbf{S}_j|^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_j)^T \mathbf{S}_j (\mathbf{y}_i - \boldsymbol{\mu}_j)\right\}, \tag{A.16}$$

and for all other components combined

$$p(c_i \neq c_{i'} \ \forall i \neq i'|\mathbf{c}_{-i}, \mathbf{y}_i, \boldsymbol{\lambda}, \mathbf{R}, \beta, \boldsymbol{\Omega}, \alpha) \propto \tag{A.17}$$

$$p(c_i \neq c_{i'} \ \forall i \neq i'|\mathbf{c}_{-i}, \alpha) \int p(\mathbf{y}_i|\boldsymbol{\mu}, \mathbf{S}) p(\boldsymbol{\mu}, \mathbf{S}|\boldsymbol{\lambda}, \mathbf{R}, \beta, \boldsymbol{\Omega}) d\boldsymbol{\mu} d\mathbf{S}, \tag{A.18}$$

the average is taken over all the inactive components, as $\boldsymbol{\mu}$ and $\mathbf{S}$ are integrated out. Since all these components have the same parameters, through their hyperparameters, it is not necessary to differentiate between them.

Prior and Posterior

The inverse of α is given a vague gamma prior, which means that

$$p(\alpha) \propto \alpha^{-3/2} \exp\left\{\frac{-1}{2\alpha}\right\} \tag{A.19}$$

To get the conditional posterior of α, a likelihood of the $\mathbf{c}$ given α is needed. This is found by a closer look at (A.12), and using some properties of the Dirichlet

process. Observe one and one c, conditional on the previously observed c's, then the following likelihood is obtained

$$\begin{aligned} p(\mathbf{c}|\alpha) &= p(c_1|\alpha)p(c_2|c_1,\alpha)\cdots p(c_n|\mathbf{c}_{n-1},\alpha) \\ &= \frac{\alpha^k \prod_{j_1}^{k}(n_j-1)!}{\alpha(1+\alpha)\cdots(n-1+\alpha)} \propto \frac{\alpha^k\Gamma(\alpha)}{\Gamma(n+\alpha)}. \end{aligned}$$

The full conditional of α thus becomes

$$p(\alpha|k,n) \propto \alpha^{-3/2}\exp\left\{\frac{-1}{2\alpha}\right\}\frac{\alpha^k\Gamma(\alpha)}{\Gamma(n+\alpha)} = \frac{\alpha^{k-3/2}\exp\left\{\frac{-1}{2\alpha}\right\}\Gamma(\alpha)}{\Gamma(n+\alpha)}. \quad \text{(A.20)}$$

Notice that it only depends on $\mathbf{c}$ trough the number of observations n, and the number of active mixture components k.

A.2 Markov Chain Monte-Carlo Sampling

A.2.1 General Procedures

Assume that the Markov Chain currently is in state t, which means that the parameters have the following current values:

$$\begin{aligned} &\mathbf{c}^t = [c_1^t, ..., c_n^t] \\ &\boldsymbol{\mu}^t = [\boldsymbol{\mu}_1^t, ..., \boldsymbol{\mu}_k^t] \\ &\mathbf{S}^t = [\mathbf{S}_1^t, ..., \mathbf{S}_k^t] \\ &\boldsymbol{\lambda}^t \\ &\mathbf{R}^t \\ &\beta^t \\ &\boldsymbol{\Omega}^t \\ &\alpha^t. \end{aligned}$$

To get to the next state, $t+1$, the following updates are done:

- Sample new $\mathbf{c}^{t+1}$. Different approaches are be described below.
- Update the component parameters based on $\mathbf{c}^{t+1}$ and the observations, and draw new values of $\boldsymbol{\mu}^{t+1}$ and $\mathbf{S}^{t+1}$ from their full conditional distributions given earlier.

- Update the hyperparameters, and draw new values for these according to their conditional distributions: $\boldsymbol{\lambda}^{t+1}, \mathbf{R}^{t+1}, \beta^{t+1}, \boldsymbol{\Omega}^{t+1}$ and α^{t+1}. The β and α parameters require Metropolis-Hastings sampling, or equivalent approach, as they have no closed form of their conditionals.

A.2.2 Sampling Class Labels

The methods used here are based on [59]. The full conditional of c_i is determined by (A.15) and (A.17), i.e.,

$$p(c_i = j|\mathbf{c}_{-i},\mathbf{y}_i, \boldsymbol{\mu}_j, \mathbf{S}_j, \alpha) \tag{A.21}$$

$$= \begin{cases} b\dfrac{n_{-i,j}}{n-1+\alpha}|\mathbf{S}_j|^{1/2}\exp\left\{-\dfrac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_j)^T\mathbf{S}_j(\mathbf{y}_i-\boldsymbol{\mu}_j)\right\} \\ b\dfrac{\alpha}{n-1+\alpha}\displaystyle\int p(\mathbf{y}_i|\boldsymbol{\mu}_j,\mathbf{S}_j)p(\boldsymbol{\mu}_j,\mathbf{S}_j|\boldsymbol{\lambda},\mathbf{R},\beta,\boldsymbol{\Omega})d\boldsymbol{\mu}_j d\mathbf{S}_j \end{cases}, \tag{A.22}$$

where b is some normalising constant. Minus the logarithm of the likelihood of y_i belonging to component j will be useful in the sampling of class labels, this can be written

$$-\frac{1}{2}\ln|\mathbf{S}_j| + \frac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_j)\mathbf{S}_j(\mathbf{y}_i-\boldsymbol{\mu}_j)^T. \tag{A.23}$$

Here follows a description of a two algorithms for sampling class labels.

Algorithm 4 in [59]

Use the conditional prior of c as the proposal distribution.

$$p(c_i = j|\mathbf{c}_{-i},\alpha) = \begin{cases} \dfrac{n_{-i,j}}{n-1+\alpha} \\ \dfrac{\alpha}{n-1+\alpha} \end{cases}$$

This will then cancel out in the acceptance probability, which now is given as

$$\alpha(c_i', c_i^t) = \min\left\{1, \frac{\pi(c_i'|\theta_{c_i'})}{\pi(c_i^t|\theta_{c_i^t})}\right\}. \tag{A.24}$$

where

$$\pi(c_i|\theta_{c_i}) = |\mathbf{S}_{c_i}|^{1/2}\exp\left\{-\frac{1}{2}(\mathbf{y}_i-\boldsymbol{\mu}_{c_i})^T\mathbf{S}_{c_i}(\mathbf{y}_i-\boldsymbol{\mu}_{c_i})\right\} \tag{A.25}$$

if c_i is equal to an existing component. If c_i suggests a new component, it is not necessary to evaluate the integral in (A.21), but rather draw new $\boldsymbol{\mu}_{c_i}$ and $\mathbf{S}_{c_i}$ from their priors, and evaluate the likelihood in (A.25), with these new values inserted.

Algorithm 8 in [59]

This is a different approach since it uses auxiliary variables, and then applies Gibbs sampling.

To be able to use Gibbs sampling, m new components are "invented". The parameters of the new components are drawn from the priors, or the "old" empty component parameters can be used. Then, the integral above in (A.21) is evaluated, and Gibbs sampling is used.

A.2.3 Sampling β

The full conditional of β is given by

$$\begin{aligned} p(\beta|\mathbf{S},\boldsymbol{\Omega}) \propto\ & (\beta - D + 1)^{-3/2} \exp\left\{\frac{-D}{2(\beta - D + 1)}\right\} \\ & \times \left(2^{\beta d/2} \prod_{i=1}^{d} \Gamma\left(\frac{\beta + 1 - i}{2}\right)\right)^{-k} \beta^{(k\beta d/2)} |\boldsymbol{\Omega}|^{k\beta/2} \left(\prod_{i=1}^{k} |\mathbf{S}_j|^{\beta/2}\right) \\ & \times \exp\left\{-\frac{\beta}{2} \sum_{j=1}^{k} tr(\boldsymbol{\Omega}\mathbf{S}_j)\right\} \end{aligned}$$

Again, the minus logarithm transformation of the expression is applied

$$\begin{aligned} & \frac{3}{2}\ln(\beta - D + 1) + \frac{D}{2(\beta - D + 1)} \\ & + \frac{k\beta d}{2}\ln(2) + k\sum_{i=1}^{d} \ln\Gamma\left(\frac{\beta + 1 - i}{2}\right) - \frac{k\beta d}{2}\ln(\beta) \\ & - \frac{k\beta}{2}\ln|\boldsymbol{\Omega}| - \frac{\beta}{2}\sum_{j=1}^{k}\ln|\mathbf{S}_j| + \frac{\beta}{2}\sum_{j=1}^{k} tr(\boldsymbol{\Omega}\mathbf{S}_j) \end{aligned}$$

Now, standard Metropolis-Hastings to update β is used.

A.2.4 Sampling α

The full conditional of α is given as

$$p(\alpha|k,n) \propto \alpha^{-3/2} \exp\left\{\frac{-1}{2\alpha}\right\} \frac{\alpha^k \Gamma(\alpha)}{\Gamma(n+\alpha)} = \frac{\alpha^{k-3/2} \exp\left\{\frac{-1}{2\alpha}\right\} \Gamma(\alpha)}{\Gamma(n+\alpha)}.$$

Minus the logarithm of this expression is

$$-(k-\frac{3}{2})\ln(\alpha) + \frac{1}{2\alpha} - \ln\Gamma(\alpha) + \ln\Gamma(n+\alpha). \tag{A.26}$$

If a normal proposal distribution centred at α^t is used, the result is

$$\alpha(\alpha', \alpha^t) = \min\left\{1, \frac{\pi(\alpha')}{\pi(\alpha^t)}\right\}. \tag{A.27}$$

where the likelihood $\pi(\cdot)$ has the form as given earlier.

A.3 Conjugate Distributions

A.3.1 Gaussian Distributed Data

When the data are Gaussian distributed, with mean $\boldsymbol{\mu}$, and precision $\boldsymbol{\Lambda}$

$$\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), \tag{A.28}$$

the likelihood of the data is

$$L(\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto |\boldsymbol{\Lambda}|^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{y}_i - \boldsymbol{\mu})\right). \tag{A.29}$$

Conjugate Prior to the Mean

A conjugate prior for $\boldsymbol{\mu}$ is the Gaussian distribution with mean $\boldsymbol{\mu}_0$ and precision $\boldsymbol{\Lambda}_0$

$$p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}). \tag{A.30}$$

It follows that the conditional posterior of $\boldsymbol{\mu}$ is

$$\begin{aligned} p(\boldsymbol{\mu}|\mathbf{Y}, \boldsymbol{\Lambda}, \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) &\propto p(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}) L(\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{y}_i - \boldsymbol{\mu})\right\} \\ &\sim \mathcal{N}\left((\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda})^{-1})(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0 + n\boldsymbol{\Lambda}\overline{\mathbf{y}}),\ (\boldsymbol{\Lambda}_0 + n\boldsymbol{\Lambda})^{-1}\right) \end{aligned}$$

Conjugate Prior for the Precision

A conjugate prior to the precision is the Wishart distribution with

$$p(\mathbf{\Lambda}|\nu_0, \mathbf{S}_0) = \mathcal{W}(\nu, \mathbf{S}_0), \tag{A.31}$$

It follows that the conditional posterior of $\mathbf{\Lambda}$ is

$$\begin{aligned}
p(\mathbf{\Lambda}|\mathbf{Y}, \boldsymbol{\mu}, \nu_0, \mathbf{S}_0) &\propto p(\mathbf{\Lambda}|\nu_0, \mathbf{S}_0) L(\mathbf{Y}|\boldsymbol{\mu}, \mathbf{\Lambda}) \\
&\propto |\mathbf{\Lambda}|^{(\nu_0-k-1)/2} \\
&\quad \times \exp\left\{-\frac{1}{2} tr(\mathbf{S}_0^{-1}\mathbf{\Lambda})\right\} |\mathbf{\Lambda}|^{n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^T \mathbf{\Lambda}(\mathbf{y}_i - \boldsymbol{\mu})\right\} \\
&= |\mathbf{\Lambda}|^{((n+\nu_0)-k-1)/2} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[tr(\mathbf{S}_0^{-1}\mathbf{\Lambda}) + \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^T \mathbf{\Lambda}(\mathbf{y}_i - \boldsymbol{\mu})\right]\right\} \\
&= |\mathbf{\Lambda}|^{((n+\nu_0)-k-1)/2} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[tr(\mathbf{S}_0^{-1}\mathbf{\Lambda}) + tr\left(\mathbf{\Lambda}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right)\right]\right\} \\
&= |\mathbf{\Lambda}|^{((n+\nu_0)-k-1)/2} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[tr(\mathbf{S}_0^{-1}\mathbf{\Lambda}) + tr\left(\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\mathbf{\Lambda}\right)\right]\right\} \\
&= |\mathbf{\Lambda}|^{((n+\nu_0)-k-1)/2} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[tr\left(\left(\mathbf{S}_0^{-1} + \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right)\mathbf{\Lambda}\right)\right]\right\} \\
&\sim \mathcal{W}\left(n + \nu_0, \left[\mathbf{S}_0^{-1} + \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right]^{-1}\right)
\end{aligned}$$

Were the relation (see, e.g. Gelman (2004))

$$\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})^T \mathbf{\Lambda}(\mathbf{y}_i - \boldsymbol{\mu}) = tr\left(\mathbf{\Lambda}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T\right),$$

is used, and for matrices with appropriate dimensions are:

$$\begin{aligned} tr(A) &= \sum_{i=1}^{k} a_{ii} \\ tr(AB) &= tr(BA) \\ tr(A) + tr(B) &= tr(A + B) \\ tr(AC + BC) &= tr((A + B)C) \end{aligned}$$

A.3.2 Wishart Distributed Data

When the data are Wishart distributed, with ν degrees of freedom, and a symmetric positive definite scale matrix $\mathbf{S}$.

$$\mathbf{W}_i \sim \mathcal{W}(\nu, \mathbf{S}^{-1}), \tag{A.32}$$

the likelihood takes the following form

$$L(\mathbf{W}|\nu, \mathbf{S}) \propto \prod_{i=1}^{n} |\mathbf{S}^{-1}|^{-\nu/2} |\mathbf{W}_i|^{(\nu-d-1)/2} \exp\left\{-\frac{1}{2} tr(\mathbf{S}\mathbf{W}_i)\right\}$$

Conjugate Prior to the Scale Matrix S

A conjugate prior to the scale matrix $\mathbf{S}$, is a Wishart distribution, with parameters $\nu_0, \mathbf{S}_0$

$$p(\mathbf{S}|\nu_0, \mathbf{S}_0) \sim \mathcal{W}(\nu_0, \mathbf{S}_0) \tag{A.33}$$

Then the conditional posterior of $\mathbf{S}$ is

$$
\begin{aligned}
p(\mathbf{S}|\mathbf{W},\nu) &\propto |\mathbf{S}|^{(\nu_0-k-1)/2}\ \exp\left\{-\frac{1}{2}tr(\mathbf{S}_0^{-1}\mathbf{S})\right\} \\
&\quad \times \prod_{i=1}^{n} |\mathbf{S}|^{\nu/2}|\mathbf{W}_i|^{(\nu-k-1)/2}\ \exp\left\{-\frac{1}{2}tr(\mathbf{S}\mathbf{W}_i)\right\} \\
&\propto |\mathbf{S}|^{(\nu_0-k-1)/2}\ \exp\left\{-\frac{1}{2}tr(\mathbf{S}_0^{-1}\mathbf{S})\right\} \prod_{i=1}^{n} |\mathbf{S}|^{\nu/2}\ \exp\left\{-\frac{1}{2}tr(\mathbf{S}\mathbf{W}_i)\right\} \\
&\propto |\mathbf{S}|^{(\nu_0+n\nu-k-1)/2}\ \exp\left\{-\frac{1}{2}tr(\mathbf{S}_0^{-1}\mathbf{S}) - \frac{1}{2}\sum_{i=1}^{n} tr(\mathbf{S}\mathbf{W}_i)\right\} \\
&\propto |\mathbf{S}|^{(\nu_0+n\nu-k-1)/2}\ \exp\left\{-\frac{1}{2}tr(\mathbf{S}_0^{-1}\mathbf{S}) - \frac{1}{2}tr\left(\left[\sum_{i=1}^{n}\mathbf{W}_i\right]\mathbf{S}\right)\right\} \\
&\sim \mathcal{W}\left(\nu_0+n\nu, \left[\mathbf{S}_0^{-1}+\sum_{i=1}^{n}\mathbf{W}_i\right]^{-1}\right)
\end{aligned}
$$

Appendix B

Failure Rates and Mean Outage Times

The failure rates and mean outage times are based on data[1] collected by Statnett (Norwegian TSO) in the period 1996-2005. Failure rates for transmission lines are given per km per year.

Component Type	Failure Rate [1/year]	Mean Outage Time [h]
Transmission line 420 kV	0.0008	28.53
Transmission line 300-220 kV	0.0007	61.58
Transmission line 132 kV	0.0019	91.400
Transmission line 110-33 kV	0.0045	16.70
Transformer 420 kV	0.0083	536.80
Transformer 300-220 kV	0.0122	1000.20
Transformer 132 kV	0.0036	367.63
Transformer 110-33 kV	0.0057	30.80
Generator 420-132 kV, $>$ 150 MVA	0.5665	20.88
Generator 420-132 kV, 150-100 MVA	0.1836	60.05
Generator 420-132 kV, 100-50 MVA	0.1611	124.88
Generator 420-132 kV, $<$ 50 MVA	0.1150	70.33
Generator 110-33 kV, 0-120 MVA	0.0988	70.33

[1] www.statnett.no/no/Kraftsystemet/Systemansvaret-FoS/Feilstatistikk/