SELECTION OF SIMPLIFIED MODELS AND PARAMETER ESTIMATION USING LIMITED DATA

by

Shaohua Wu

A thesis submitted to the Department of Chemical Engineering

in conformity with the requirements for the degree of

Doctor of Philosophy

Queen's University

Kingston, Ontario, Canada

(December, 2009)

Abstract

Due to difficulties associated with formulating complex models and obtaining reliable estimates of unknown model parameters, modellers often use simplified models (SMs) that are structurally imperfect and that contain a smaller number of parameters. The objectives of this research are: 1) to develop practical and easy-to-use strategies to help modellers select the best SM from a set of candidate models, and 2) to assist modellers in deciding which parameters in complex models should be estimated, and which should be fixed at initial values. The aim is to select models and parameters so that the best possible predictions can be obtained using the available data and the modeller's engineering and scientific knowledge.

This research summarizes the extensive qualitative and quantitative results in the statistics literature regarding the use of SMs. Mean-squared error (MSE) is used to judge the quality of model predictions obtained from different candidate models, and a confidence-interval approach is developed to assess the uncertainties associated with whether a SM or the corresponding extended model will give better predictions. Nine commonly-applied model-selection criteria (MSC) are reviewed and analyzed for their propensities of preferring SMs. It is shown that there exist preferential orderings for many MSC that are independent of model structure and the particular data set. A new MSE-based MSC is developed using univariate linear statistical models. The effectiveness of this criterion for selecting dynamic nonlinear multivariate models is demonstrated both theoretically and empirically. The proposed criterion is then applied for determining the optimal number of parameters to estimate in complex models, based on ranked parameter lists obtained from estimability analysis. This approach makes use of the modeller's prior knowledge about precision of initial parameter values and is less computationally expensive than comparable methods in the literature.

Co-Authorship

The research in Chapter 2 has been published in the *Canadian Journal of Chemical Engineering*. Chapters 3 and 4 were submitted to the *Canadian Journal of Chemical Engineering* in September 2009. A journal article based on research results in Chapter 5 is in preparation and will be submitted in the future. I prepared the first and subsequent drafts for all of these coauthored manuscripts, performed all of the calculations and simulations, and generated all of the Figures and Tables. Drs. Kim McAuley and Thomas Harris were co-authors of the journal articles. They helped to formulate research objectives, they provided technical advice and they edited the thesis and suggested revisions.

Acknowledgements

First, I would like to express my sincerest gratitude to my supervisors, Drs. Kim McAuley and Thomas Harris, for their critical assessments and helpful suggestions for my research, and for their kind support of my life and my family in Canada. I would also like to thank Dr. James McLellan for the many interesting discussions and helpful suggestions on my graduate studies, my research and my work as a teaching assistant.

My parents, my wife, and my kids deserve the most credit for all of their love and support. I would not be in the position I am without the opportunities and encouragement they gave me.

I would also like to thank Eugene Theobald and Weihua Li from Shell Global Solutions (US), who offered me the precious opportunity of an internship to practice my knowledge, to improve my problem-solving abilities, and to enhance my communication skills. Their support and guidance have been of great help to me. The successful completion of this internship sponsored by Shell and MITACS was one of the greatest accomplishments I achieved during my graduate studies.

Many thanks go out to all of my former and current lab-mates, especially Saeed Varziri and Duncan Thompson, who help me both in research and in my life in Canada. I would also like to thank my many friends, too many to mention individually, who have made my stay in Kingston a pleasurable journey.

Finally, I would like to thank MITACS, Government of Ontario (OGSST), Queen's University, and various industrial companies (Cybernetica, DuPont, Hatch, Matrikon, SAS, Shell) for their financial support of my research.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Modelling Dilemmas: Complex Models and Limited Data

A mathematical model, by definition, is a representation, in mathematical terms, of *certain aspects* of a nonmathematical system (Aris, 1999). Here "certain aspects" means that only part of the nonmathematical situation is being represented, depending on the intended model-use, the modeller's knowledge about the underlying process, the data at hand, and the assumptions made. Researchers and engineers develop empirical models based on past data and experience, or phenomenological models based on material and energy balances and constitutive equations. Often phenomenological models contain empirical expressions, either due to a lack of detailed understanding of the underlying phenomena or the desire to keep the structure of the model reasonably simple. In chemical engineering, mathematical modelling plays an important role, and models are used for simulating, designing, controlling and optimizing many different types of processes (e.g., Perregaard, 1993; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Golbert and Lewin, 2004; Lv et al., 2004; Maria, 2004; Mchaweh et al., 2004; Chang et al., 2005; Romdhane and Tizaoui, 2005).

In many modelling situations, modellers have sufficient scientific knowledge to derive complex models, which can be expected to match the underlying behaviour of the process very well. Unfortunately, the data available for parameter estimation are often not sufficiently informative to provide precise estimates of all unknown model parameters (e.g., Perregaard, 1993; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Lv et al., 2004; Maria, 2004, 2006; Mchaweh et al., 2004; Chang et al., 2005; Romdhane and Tizaoui, 2005; Wang et al., 2007). For

complex models with many parameters, the resulting parameter estimates and model predictions may exhibit high variability, and decisions that are made using these models or their parameter estimates may be incorrect. In some situations, it is impossible to estimate all of the unknown parameters due to the model structure or due to insufficient information from the measured responses (Grewal and Glover, 1976; Eisenfeld, 1986; Vajda et al., 1989; Ljung and Glad, 1994; Dotsch and Van Den Hof, 1996; Walter and Pronzato, 1996; Ben-Zvi et al., 2003, 2004a, 2004b, 2006; Raue et al., 2009). Because of the many difficulties associated with formulating complex models and obtaining good parameter estimates, modellers often use simplified models (SMs), which are known to be structurally imperfect and contain only a reduced set of parameters (Hiskens, 2001; Kou et al., 2005a, 2005b; Anh et al., 2006; Bastogne et al., 2007; Sin and Vanrolleghem, 2007; Thompson et al., 2009). In these less-than-ideal situations, they must make decisions about which parameters to estimate, and which to hold constant or to remove via model simplification, so that the best possible predictions can be obtained.

1.2 Simplified Models: Ockham's Razor

Since "all models are wrong, but some are useful" (Box, 1979), the job of a modeller is to find a good "wrong" model for the intended purpose, so that reliable decisions can be made based on the resulting parameter estimates and model predictions. This good model should be as simple as is reasonable, following the advice from Ockham's Razor (from the 14th century), which has been translated as "It is vain to do by more what can be done by fewer" (Brooks and Tobias, 1996).

SMs are developed when: 1) the modeller lacks an understanding of some underlying phenomena in the complex process being modelled, 2) when fast on-line calculations are

required; or 3) when there are insufficient data to adequately tune and validate a more complex model. There are many reasons to choose a SM with fewer parameters and terms than the correctly-structured complex model (Neto and Cotta, 1992; Talukdar and Basu, 1995; Gray, 1997; Zhang, 1997; Ismail, 2004; Jaree et al., 2004; Forney et al., 2005; Marquardt, 2005). SMs often have reduced input and computational requirements, making them more portable and less expensive to use and maintain (Innis and Rexstad, 1983; Rexstad and Innis, 1985; Brooks and Tobias, 1996). The fewer unknown parameters contained in the SM can be more readily and precisely estimated than the numerous parameters in the corresponding complex model. The practical advantages of a parsimonious model often overshadow concerns over the correctness of the model structure. Sometimes SMs can be expected to give better predictions, in the sense of mean-squared prediction error, than the correctly-structured model (Rao, 1971; Hocking 1976; Wu et al., 2007), especially when the data are limited (e.g., data can be noisy or strongly correlated, or obtained using a small range of independent-variable settings from poorly designed experiments). These conditions are unfavourable for obtaining precise parameter estimates in any model. In this thesis, it is shown that when modellers are faced with uninformative and noisy data from poorly-designed experiments, they should not try to estimate too many model parameters. Rather, they should confine themselves to fitting only a few key parameters that appear in the most important parts of their models. Unfortunately, an over-simplified model may fail to account for important phenomena, resulting in poor operating decisions and ill-conceived process designs.

Engineers and scientists struggle to make models with an appropriate level of detail, either by making simplifying assumptions during the model formulation step or by simplifying a complex model after it has been derived (e.g., Bagajewicz and Cabrera, 2003; Golbert and Lewin, 2004;

Mchaweh et al., 2004; Chang et al., 2005; Romdhane and Tizaoui, 2005). Modellers need simple and reliable tools to determine the appropriateness of their modelling assumptions, or the optimal number of parameters and terms to estimate in their models.

In the literature, various model-selection criteria (MSC) have been developed to aid the modeller in selecting an appropriate model (e.g. Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Lanterman (2001), Rao and Wu (2001), Burnham and Anderson (2002), Stoica and Selen (2004), Konishi and Kitagawa (2008) and the many references contained therein). Some modellers develop a set of candidate models, which are postulated to give good approximations to the underlying behaviour of the process (Burnham and Anderson, 2002). Among these candidate models, the "best" one is selected based on some measure of model goodness. Akaike (1973) said that modelling and model selection are essentially concerned with the "art of approximation". Model simplification consists of techniques for reducing the complexity of an existing complex model to find a good approximation of the underlying process without losing much accuracy (Obinata and Anderson, 2001).

In chemical engineering applications, MSC have been studied and applied by modellers to solve practical problems (Schaper et al., 1994; Kendi and Doyle, 1996; Li et al., 2002; Chetouani, 2007; Toher et al., 2007; Choi et al., 2008; Scherr et al., 2008), but practical issues involving model selection using limited data are far from settled. Lanterman (2001) states: "Unfortunately, model order estimation remains a subject of tremendous controversy; there is little agreement on what the 'best' approach is, and indeed little agreement on if there is, in fact, such a thing as a 'best' approach". In a similar vein, Shibata (1989) notes that consistency in selecting the true system is only meaningful when the true system is simple and when one of the candidate models

can describe the system without error. He also notes that properly selecting the true model does not always lead to the best parameter estimates.

The research in this thesis is related to the literature on parameter subset selection in complex models, which usually contain too many unknown parameters for reliable estimation using the available data. Some parameters may have little effect on the model predictions, and the effect of some parameters can be highly correlated with the effect of others. Several different strategies have been developed to select sets of influential parameters for estimation (e.g., Velez-Reyes and Verghese (1995), Sandink et al. (2001), Degenring et al. (2004), Lund and Foss (2008)). Yao et al. (2003) proposed an orthogonalization algorithm to rank model parameters from most estimable to least estimable. This deflation algorithm accounts for both the magnitude of parameter influences and for correlation among the effect of various parameters during the ranking procedure. Thompson et al. (2009) used this orthogonalization algorithm and brute-force cross-validation method (Stone, 1974) to determine the optimal number of parameters to estimate in a model that predicts molecular weight distributions of ethylene/hexene copolymers produced by a Ziegler-Natta catalyst. A detailed review of parameter subset selection strategies is provided in Chapter 5. Note that, current strategies either involve subjective decisions in the selection of the parameter subset, or require extensive computations, especially with complex models with many unknown parameters. These deficiencies may result in unreliable model predictions and in incorrect engineering decisions based on these predictions.

1.3 Thesis Outline

In this thesis, Chapter 2 summarizes the literature concerned with using and selecting SMs. A confidence-interval approach is developed to assess the uncertainties associated with whether a

SM or the corresponding extended model (EM) will provide lower-MSE model predictions at the design points used in parameter estimation. It is shown that, when SMs are preferable due to limited data, decisions concerned with whether the SM or the EM will give better predictions are often uncertain. One short-coming of the confidence-interval approach developed in Chapter 2 is that it can only be used for comparing two nested models, where the SM is a simplified version of the more complex EM. However, in many practical situations, modellers often consider a set of candidate SMs that may or may not be nested with each other. The research in Chapter 3 reviews nine MSC that are commonly-used to compare and select non-nested SMs. Interesting connections between the various criteria and their sampling properties are derived, enabling the ranking of the relative propensities of the criteria for preferring SMs rather than the corresponding correctly-structured EM. It is shown that there exist preferential orderings for many of the MSC that are independent of the model structure and the available data set. It is shown that MSC with strong tendencies to guard against overfitting have high tendencies to select SMs when data are limited.

In Chapter 4, a new MSC is proposed for selecting the best model (i.e., with the lowest expected MSE for predictions) from a set of candidate models that includes the correctly-structured EM and several SMs. This criterion, which accounts for bias due to imperfect model structure and for variance in model predictions arising from noisy data, is first developed using single-response linear statistical models. The proposed criterion is then extended for selecting multi-response nonlinear models. Its effectiveness is demonstrated theoretically and by using Monte-Carlo simulations.

Chapter 5 reviews current strategies in the literature for selecting parameters to estimate in complex models. It is shown that the MSE-based criterion developed in Chapter 4 can be used, in

conjunction with the estimability analysis procedure of Yao et al. (2003), to help modellers determine the optimal number subset of parameters to estimate using the available data. Application of the methodology is illustrated using a DAE model formulated by the Dow Chemical Company (Biegler et al., 1986). The results obtained, which are consistent with those obtained using brute-force cross-validation method, indicate that the proposed strategy is effective and computationally reasonable.

Note that this thesis has been prepared using a manuscript format, so there is some inevitable duplication of ideas, particularly in the Chapter introductions, and some nomenclature is not consistent throughout the entire thesis. I apologize to the reader for any inconvenience.

1.4 References

Akaike, H. "Information Theory and an Extension of the Maximum Likelihood Principle," 2. Int. S. Inf. Theor., Ed. PETROV, B. N. and CSAKI F., pp. 267-281. Budapest: Akademia Kiado (1973).

Anh, D. T., M. P. Bonnet, G. Vachaud, C. V. Minh, N. Prieur, L. V. Duc and L. L. Anh, "Biochemical Modeling of the Nhue River (Hanoi, Vietnam) Practical Identifiability Analysis and Parameters Estimation," Ecol. Model., 193, 182-204 (2006).

Aris, R. "Mathematical Modeling: A Chemical Engineer's Perspective," Academic Press, NY, p.3 (1999).

Bagajewicz, M. J. and E. Cabrera, "Data Reconciliation in Gas Pipeline Systems," Ind. Eng. Chem. Res. 42(22), 5596-5606 (2003).

Bastogne, T., M. Thomassin and J. Masse, "Selection and Identification of Physical Parameters from Passive Observations. Application to a Winding Process," Control Eng. Pract., 15, 1051-1061 (2007).

Ben-Zvi, A., P. J. McLellan and K. B. McAuley, "Identifiability of Linear Time-Invariant Differential-Algebraic Systems. I. The Generalized Markov Parameter Approach," Ind. Eng. Chem. Res., 42, 6607-6618 (2003).

Ben-Zvi, A., K. McAuley and J. McLellan, "Identifiability Study of a Liquid-Liquid Phase-Transfer Catalyzed Reaction System," AIChE J., 50(10), 2493-2501 (2004a).

Ben-Zvi, P. J. McLellan and K. B. McAuley, "Identifiability of Linear Time-Invariant Differential-Algebraic Systems. II. The Differential-Algebraic Approach," Ind. Eng. Chem. Eng., 43, 1251-1259 (2004b).

Ben-Zvi, A., P. J. McLellan and K. B. McAuley, "Identifiability of Non-Linear Differential Algebraic Systems via a Linearization Appraoch," Can. J. Chem. Eng., 84, 590-596 (2006).

Bielger, L. T., J. J. Damiano and G. E. Blau, "Nonlinear Parameter Estimation – A Case Study Comparison," AIChE J., 32(1), 29-45 (1986).

Box, G.E.P., "Robustness in the Strategy of Scientific Model Building," in "Robustness in Statistics," R.L. Launer and G.N. Wilkinson, Academic Press, NY, pp. 201-236 (1979).

Brooks, R. J. and A. M. Tobias, "Choosing the Best Model: Level of Detail, Complexity, and Model Performance", Math. Comput. Modell. 24(4), 1-14 (1996).

Burnham, K. P. and D. R. Anderson, "Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach," 2nd Edn. Springer, NY, pp. 2, 49-148 (2002).

Chang, S., T. D. Waite and A. G. Fane, "A Simplified Model for Trace Organics Removal by Continuous Flow PAC Adsorption/Submerged Membrane Processes," J. Membrane Sci. 253(1-2), 81-87 (2005).

Chetouani, Y., "Modeling and Prediction of the Dynamic Behavior in a Reactor-Exchanger Uisng NARMAX Neural Structure," Chem. Eng. Commun. 194(5), 691-705 (2007).

Choi S. W., J. Morris and I. B. Lee, "Dynamic Model-based Batch Process Monitoring," Chem. Eng. Sci. 63(3), 622-636 (2008).

Degenring, D., C. Froemel, G. Dikta and R. Takors, "Sensitivity Analysis for the Reduction of Complex Metabolism Models," J. Process Contr., 14, 729-745 (2004).

Dotsch, H. G. M. and P. M. J. Van Den Hof, "Test for Local Structural Identifiability of High-order Non-linearly Parametrized State Space Models," Automatica, 32(6), 875-883 (1996).

Eisenfeld, J., "A Simple Solution to the Compartmental Structural-Identifiability Problem," Math. Biosci., 79, 209-220 (1986).

Forney, L. J., J. M. Brown, B. M. Kadlubowski and J. T. Sommerfeld, "Simplified Model for Oxygen Transport with Reaction in a Polymer-Electrolyte Fuel Cell," Can. J. Chem. Eng. 83(3), 500-507 (2005).

Golbert, J. and D. R. Lewin, "Model-Based Control of Fuel Cells: (1) Regulatory Control", J. Power Sources 135(1-2), 135-151 (2004).

Gray, M. R. "Through a Glass, Darkly: Kinetics and Reactors for Complex Mixtures," Can. J. Chem. Eng. 75(3), 481-493 (1997).

Grewal, M. S. and K. Glover, "Identifiability of Linear and Nonlinear Dynamical Systems," IEEE T. Automat. Contr., 21(6), 833-836 (1976).

Hiskens, I. A., "Nonlinear Dynamic Model Evaluation from Disturbance Measurements," IEEE T. Power Syst., 16(4), 702-710 (2001).

Hocking, R. R. "Analysis and Selection of Variables in Linear Regression," Biometrics 32(1), 1-49 (1976).

Innis, G. and E. Rexstad, "Simulation Model Simplification Techniques", Simulation 41(1), 7-15 (1983).

Ismail, A. S., "Holdup Profile in Multistage Stirred-Cell Liquid-Liquid Extraction Column Using Population Balance Model," Can. J. Chem. Eng. 82(5), 1037-1043 (2004).

Jaree, A., R. R. Hudgins, H. Budman, P. L. Silveston, M. Menzinger, "Numerical Investigation of Resonance Behaviour of a Tubular Packed-Bed Reactor with Non-uniform Activity," Can. J. Chem. Eng. 82(2), 387-391 (2004).

Kendi, T. A. and F. J. Doyle, "Nonlinear Control of a Fluidized Bed Reactor using Approximate Feedback Linearization," Ind. Eng. Chem. Res. 35(3), 746-757 (1996).

Konishi, S. and G. Kitagawa, "Information Criteria and Statistical Modeling," Springer, NY, pp. 29-254 (2008).

Kou B., K. B. McAuley, C. C. Hsu, D. W. Bacon and K. Z. Yao, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene Homopolymerization with Supported Metallocene Catalyst," Ind. Eng. Chem. Res. 44, 2428-2442 (2005a).

Kou B., K. B. McAuley, J. C. C. Hsu and D. W. Bacon, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene/Hexene Copolymerization with Metallocene Catalyst," Macromol. Mater. Eng., 290, 537-557 (2005b).

Lanterman, A. D. "Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection," Int. Stat. Rev. 69(2), 185-212 (2001).

Li B. B., J. Morris and E. B. Marin, "Model Selection for Partial Least Sqaures Regression," Chemom. Intell. Lab. Syst., 64(1), 79-89 (2002).

Linhart, H. and W. Zucchini, "Model Selection," John Wiley and Sons, NY, pp. 1-38 (1986).

Ljung, L. and T. Glad, "On Global Identifiability for Arbitrary Model Parametrizations," Automatica, 30(2), 265-276 (1994).

Lund, B. F. and B. A. Foss, "Parameter Ranking by Orthogonalization – Applied to Nonlinear Mechanistic Models," Automatica, 44(1), 278-281 (2008).

Lv, P., J. Chang, T. Wang, C. Wu and N. Tsubaki, "A Kinetic Study on Biomass Fast Catalytic Pyrolysis", Energy Fuels 18(6), 1865-1869 (2004).

Maria, G. "A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems," Chem. Biochem. Eng. Q. 18(3), 195-222 (2004).

Maria, G. "Application of Lumping Analysis in Modeling the Living Systems - a Trade-off between Simplicity and Model Quality," Chem. Biochem. Eng. Q. 20(4), 353-373 (2006).

Marquardt, W., "Model-Based Experimental Analysis of Kinetic Phenomena in Multi-Phase Reactive Systems," Chem. Eng. Res. Des. 83(A6), 561-573 (2005).

Mchaweh, A., A. Alsaygh, K. Nasrifar and M. Moshfeghian, "A Simplified Method for Calculating Saturated Liquid Densities," Fluid Phase Equilib. 224(2), 157-167 (2004).

McQuarrie, A. D. R. and C. L. Tsai, "Regression and Time Series Model Selection," World Scientific, Singapore, pp. 15-87 (1998).

Neto, F. S. and R. M. Cotta, "Lumped-Differential Analysis of Concurrent Flow Double-Pipe Heat-Exchanger," Can. J. Chem. Eng. 70(3), 592-595 (1992).

Obinata G. and D. O. Anderson, "Model Reduction for Control System Design," Springer-Verlag, London, pp.1-59 (2001).

Perregaard, J., "Model Simplification and Reduction for Simulation and Optimization of Chemical Processes,". Comput. Chem. Eng. 17(5-6), 465-483 (1993).

Rao, P. "Some Notes on Misspecification in Multiple Regressions," Am. Stat. 25(5), 37-39 (1971).

Rao, C. R. and Y. Wu, "On Model Selection (with Discussion)," in "Model Selection," (Ed. by P. Lahiri), IMS Lecture Notes – Monograph Series 38, 1-64 (2001).

Raue, A., C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmuller and J. Timmer, "Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood," Systems Biol., 25(15), 1923-1929 (2009).

Rexstad, E. and G. S. Innis, "Model Simplification – 3 Applications", Ecological Modeling, 27(1-2), 1-13 (1985).

Romdhane, M. and C. Tizaoui, "The Kinetic Modeling of a Steam Distillation Unit for the Extraction of Aniseed (Pimpinelia Anisum) Essential Oil," J. Chem. Technol. Biot. 80(7), 759-766 (2005).

Sandink, C. A., K. B. McAuley and P. J. McLellan, "Selection of Parameters for Updating in On-line Models," Ind. Eng. Chem. Res., 40, 3936-3950 (2001).

Schaper C. D., W. E. Larimore, D. E. Seborg and D. A. Mellichamp, "Identification of Chemical Processes Using Canonical Variate Analysis," Comput. Chem. Eng., 18(1), 55-69 (1994).

Scherr, F. F., A. K. Sarmah, H. J. Di and K. C. Cameron, "Modeling Degradation and Metabolite Formation Kinetics of Estrone-3-sulfate in Agricultural Soils," Environ. Sci. Technol. 42(22), 8388-8394 (2008).

Shibata, R., "Statistical Aspects of Model Selection," In "From Data to Model," J. C. Willems, Springer-Verlag, NY, pp. 215-240 (1989).

Sin, G. and P. A. Vanrolleghem, "Extensions to Modeling Aerobic Carbon Degradation using Combined Respirometric-Titrimetric Measurements in View of Activated Sludge Model Calibration," Water Res., 41, 3345-3358 (2007).

Stoica, P. and Y. Selen, "Model-Order Selection: A Review of Information Criterion Rules," IEEE Signal Proc. Mag. 21(4), 36-47 (2004).

Stone, M., "Cross-Validation Choice and Assessment of Statistical Predictions," J. Roy. Stat. Soc. B., 36(2), 111-147 (1974).

Talukdar, J. and P. Basu, "A Simplified Model of Nitric-Oxide Emission from a Circulating Fluidized-Bed Combustor," Can. J. Chem. Eng. 73(5), 635-643 (1995).

Thompson D. E., K. B. McAuley and P. J. McLellan, "Parameter Estimation in a Simplified MWD Model for HDPE Produced by a Ziegler-Natta Catalyst," Macromol. React. Eng. 3, 160-177 (2009).

Toher D., G. Downey and T. B. Murphy, "A Comparison of Model-based and Regression Classification Techniques applied to Near Infrared Spectroscopic Data in Food Authentication Studies," Chemom. Intell. Lab. Syst. 89(2), 102-115 (2007).

Vajda, S., H. Rabitz, E. Walter and Y. Lecourtier, "Qualitative and Quantitative Identifiability Analysis of Nonlinear Chemical Kinetic Models," Chem. Eng. Commun., 83, 191-219 (1989).

Velez-Reyes, M. and G. C. Verghese, "Subset Selection in Identification, and Application to Speed and Parameter Estimation for Induction Machines," Proc. 4th IEEE Conf. Control Appl., 991-997 (1995).

Walter, E. and L. Pronzato, "On the Identifiability and Distinguishability of Nonlinear Parametric Models," Math. Comput. Simulat. 42, 125-134 (1996).

Wang, F. Y., Z. H. Zhu, P. Massarotto and V. Rudolph, "A Simplified Dynamic Model for Accelerated Methane Residual Recovery from Coals," Chem. Eng. Sci. 62(12), 3268-3275 (2007).

Wu, S., T. J. Harris and K. B. McAuley, "The Use of Simplified or Misspecified Models: Linear Case," Can. J. Chem. Eng. 85, 386-398 (2007).

Yao, K. Z., B. M. Shaw, K. B. McAuley and D. W. Bacon, "Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design," Polym. React. Eng., 11(3), 563-588 (2003).

Yoshida, H., Y. Takahashi and M. Terashima, "A Simplified Reaction Model for Production of Oil, Amino Acids, and Organic Acids from Fish Meat by Hydrolysis under Sub-Critical and Supercritical Conditions," J. Chem. Eng. JPN. 36(4), 441-448 (2003).

Zhang, P. "Comment on 'An Asymptotic Theory for Linear Model Selection'," Stat. Sinica, 7, 254-258 (1997).

Chapter 2

The Use of Simplified or Misspecified Models: Linear Case[1]

2.1 Summary

Simplified models have many appealing properties and sometimes give better parameter estimates and model predictions, in sense of mean-squared error, than extended models, especially when the data are not informative. This Chapter summarizes extensive quantitative and qualitative results in the statistics literature concerned with using simplified or misspecified models. Based on confidence intervals and hypothesis tests, a practical strategy is developed to help modeller decide whether a simplified model should be used, and the difficulty in making such a decision is discussed. This analysis also evaluates several methods for statistical inference based on simplified or misspecified models.

2.2 Introduction

Chemical engineers develop simplified models (SMs) and use them for simulating, designing, controlling and optimizing many different types of processes (Perregaard, 1993; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Golbert and Lewin, 2004; Lv et al., 2004; Maria, 2004; Mchaweh et al., 2004; Chang et al., 2005; Romdhane and Tizaoui, 2005; Brendel et al., 2006). SMs are developed when the modeller lacks an understanding of some of the underlying

[1] The research work summarized in this Chapter has been published in *Canadian Journal of Chemical Engineering* in 2007. Drs. Thomas Harris and Kim McAuley were co-authors of this journal article. Note that this thesis has been prepared using a manuscript format, so some nomenclature used is not consistent throughout the entire thesis. Please refer to Section 2.9 for the nomenclature used in this Chapter.

phenomena in complex processes, or there are insufficient data to adequately calibrate or validate an extended model (EM). An extended model is a mathematical description of the process that contains sufficient phenomenological detail to provide good predictions over the range of conditions of interest, including those experimental conditions where data may not have been collected. SMs often have reduced input and computational requirements compared with EMs, making them more portable and less expensive to use and maintain (Innis and Rexstad, 1983; Rexstad and Innis, 1985; Brooks and Tobias, 1996). SMs usually contain fewer unknown parameters, which are more readily and precisely estimated than the numerous parameters in the EM. Problems of parameter estimability in EMs are more pronounced when limited experimental data are available. However, SM may fail to account for important phenomena, resulting in poor operating decisions and ill-conceived process designs. Although formal techniques have been developed for model simplification (Innis and Rexstad, 1983; Rexstad and Innis, 1985; Maria, 2004; Kou et al., 2005a, 2005b; Sun and Hahn, 2006), finding the right balance between simplicity and complexity usually involves more engineering judgment than science (Brooks and Tobias, 1996).

There has been considerable research in the statistics literature on the statistical consequences of using simplified or misspecified models (Box and Draper, 1959; Freund et al., 1961; Goldberg, 1961; Goldberg and Jochems, 1961; Kabe, 1963; Wallace, 1964; Toro-Vizcarrondo and Wallace, 1968; Rao, 1971; Rosenberg and Levy, 1972; Hocking, 1976; White, 1981, 1982; Miller, 1990; Golden, 1995; Bera, 2000; Rao and Wu, 2001; Waldorp et al., 2005; O'Brien et al., 2006; Waldorp et al., 2006), particularly for models that are linear in the parameters. Three general situations have been considered: 1) input variables that belong in the true model are mistakenly omitted (this is sometimes known as under-modelling); 2) variables that have no real influence on

the output variables are mistakenly included (this is sometimes known as overfitting); and 3) errors are made in the distributional assumptions of the stochastic or random component of the model (Rao, 1971; Hocking, 1976; Seber and Wild, 2003). Issues related to 3) are more difficult to generalize and are not discussed further in this Chapter.

Important quantitative and qualitative results have been derived to compare the parameter estimates and model predictions from misspecified models with those from correctly-structured models (Box and Draper, 1959; Freund et al., 1961; Goldberg, 1961; Goldberg and Jochems, 1961; Kabe, 1963; Wallace, 1964; Rao, 1971; Rosenberg and Levy, 1972; Hocking, 1976; Abdullaev and Geidarov, 1985; Miller, 1990). While it might seem that the use of misspecified models will always lead to inferior model predictions and parameter estimates that are biased, this is not always true (Rao, 1971; Hocking, 1976; Waldorp et al., 2006).

Like many chemical engineers, we are particularly interested in developing phenomenological models based on material and energy balances and constitutive equations. The usual objective of the statistical approach to model building (for either empirical or mechanistic models) is to develop models that are of sufficient complexity so that the model passes statistical adequacy tests (Montgomery and Runger, 2003). When this objective has been achieved, probability statements can be assigned to the precision of the estimated parameters and model predictions (Draper and Smith, 1998; Montgomery et al., 2001). There has been far less research on providing similar information for simplified or misspecified models (White, 1981, 1982; Golden, 1995; Bera, 2000). Recent research in the use of more computationally intensive methods for statistical analysis of misspecified models (e.g. nonparametric bootstrapping) has revived interest in this topic (Davison and Hinkley, 1997; Aerts and Claeskens, 2001; Velilla, 2001; Fushiki, 2005; Waldorp et al., 2006).

This Chapter: 1) summarizes the quantitative and qualitative results in the literature concerned with using simplified or misspecified models; 2) provides new insights into the conditions under which simplified or misspecified models give superior predictions compared with the correctly-structured EM; and 3) evaluates methods that can be used for statistical inference for models that are simplified or misspecified. A new practical strategy, based on confidence intervals and hypothesis tests, is developed to help modellers decide whether a SM will give better predictions than the EM. However, there are considerable challenges in making such a decision. The resulting confidence intervals are quite large and the statistical tests, while exact in their construction, have poor discrimination properties for the alternative hypothesis that the SM is better or that the EM is better.

This research focuses on models that are linear in the parameters. While this choice might at first seem restrictive because chemical engineers tend to use nonlinear models, we note that, the statistical analysis of nonlinear models usually involves a linearization of the model around the nominal parameter values (Seber and Wild, 2003). Thus, the results in this Chapter can assist model developers in the analysis of phenomenological models that are nonlinear in the parameters.

This Chapter is organized as follows. A general description of model misspecification is given in section 2.3. In section 2.4, misspecification in models that are linear in the parameters is analyzed theoretically. There are extensive results in the literature for this topic, but an important unresolved issue is the lack of practical tests for determining whether a SM will give better predictions than the correctly-structured EM. The new practical strategy that uses the model structure and the data available for parameter estimation is provided in section 2.5. This is followed in section 2.6 by an analysis of constructive numerical methods that can be used to

make statistical statements regarding the uncertainty in parameters and model predictions when the SM is used. In section 2.7, analytical results and Monte Carlo simulations from a simple example are used to provide insights into the most important results from the previous sections. This Chapter concludes with a brief discussion on the applicability of these methods to models that are nonlinear in the parameters.

2.3 Misspecified Models – The General Case

It is assumed that a process can be correctly-described by

$$y_i = f(x_i, \beta) + \varepsilon_i \ \ (i = 1,2, \dots, n) \tag{2.1}$$

where x_i is a k dimensional vector of explanatory variables for the i^{th} observation, β is an m dimensional vector of model parameter, and random variables ε_i is independently and identically distributed with mean 0 and variance σ^2 .

Usually, a modeller supposes that the observed response variable can be represented as

$$y_i = g(z_i, \theta) + e_i \ \ (i = 1,2, \dots, n) \tag{2.2}$$

where z_i is a vector of explanatory variables, θ is a vector of parameters and $g(z_i, \theta)$ is the function that the modeller believes (or hopes) relates (z_i, θ) to the response. The term e_i encompasses the stochastic component and any deterministic part that is not captured by the model. The functional form of $g(z_i, \theta)$ may be specified from a fundamental understanding of the process or a desire to find a purely empirical representation. In either case, the parameters are often estimated as the solution to the least-squares problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, S(\theta) = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \big(y_i - g(z_i, \theta)\big)^2 \tag{2.3}$$

When the parameters enter the model nonlinearly, a nonlinear optimization algorithm is required to determine $\hat{\theta}$. When the parameters enter linearly, ordinary least-squares (OLS) is commonly used (Montgomery and Runger, 2003).

A model-building strategy is typically iterative. A model form is postulated, and the parameters are estimated. The residuals $\hat{e}_i = y_i - g(z_i, \hat{\theta})$ are then examined for systematic patterns that suggest omitted variables. The model may be "pruned" by eliminating parameters that are not statistically different from zero. While no assumptions on the probability structure of the stochastic components are required to determine the least-squares estimates, the validity of the statistical analysis requires several assumptions (Montgomery et al., 2001). Usually, it is assumed that the model structure is correct ($g(z_i, \theta) = f(x_i, \beta)$), that the explanatory variables are deterministic, and that ε_i, in addition to being independently and identically distributed, follows a Normal distribution. When these assumptions are satisfied, there is a rich body of knowledge related to model building and statistical assessment of the model goodness (Draper and Smith, 1998; Montgomery et al., 2001).

However, there are many instances when the modeller deliberately chooses a structural form that does not match the true process. In these instances, the model may be acceptable (in sense of an intended end-use), but may fail a statistical test for adequacy (Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Golbert and Lewin, 2004; Chang et al., 2005). This is particularly true when fundamental models are used. In the case of empirical models, it may happen that some of the predictor variables are deleted from the model because they are inaccessible, in which case the model is incorrect. In this analysis, these structurally imperfect models are referred to as simplified or misspecified models. Several interesting questions arise:

1) Can misspecified models give better parameter estimates and model predictions than the correctly-structured extended model?
2) What statistical methods can be used to analyze these models and to make statements about the quality of their predictions?

There is a rich literature that addresses question 1) (Rao, 1971; Hocking, 1976; Miller, 1990). Not surprisingly, almost all of this work relates to models that are linear in parameters. In these instances, closed-form solutions can be obtained after a definition for "better" is specified. Some results are also available to address question 2) (Golden, 1995; Bera, 2000; Waldorp et al., 2005; Waldorp et al., 2006).

2.4 Misspecified Models – The Linear Case

Assume that the true process is described by

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{2.4}$$

where there are p unknwon parameters in β_1, q unknown parameters in β_2, and n data points available for parameter estimation. In this analysis, this correctly-structured model is referred to as the extended model (EM).

The following assumptions are usually made (Beck and Arnold, 1977):

1) X_1 and X_2 have full column rank, and are deterministic;
2) The stochastic component ε is independently and identically distributed with zero mean and constant variance σ^2.

Define $\beta = (\beta_1^T \quad \beta_2^T)^T$ and $X = (X_1 \quad X_2)$. Then the OLS estimates are given by

$$\hat{\beta}_E = (X^T X)^{-1} X^T Y \tag{2.5}$$

where the subscript "*E*" indicates the use of the correctly-structured EM.

When the OLS assumptions are satisfied,

$$E(\hat{\beta}_E) = \beta$$
$$Cov(\hat{\beta}_E) = \sigma^2 (X^T X)^{-1} \tag{2.6}$$

where $E(\cdot)$ and $Cov(\cdot)$ denote the statistical expectation and covariance of $(\cdot)$.

It is convenient to write Eqn. (2.6) in the form of $\hat{\beta}_E \sim (\beta, \sigma^2 (X^T X)^{-1})$. The notation $(\cdot) \sim (\mu, \Sigma)$ denotes that $\mu = E(\cdot)$ and $\Sigma = Cov(\cdot)$. The variance σ^2 can be estimated via

$$s_E^2 = \frac{S(\hat{\beta}_E)}{n-m} = \frac{(Y - X\hat{\beta}_E)^T (Y - X\hat{\beta}_E)}{n-m} \tag{2.7}$$

where $m = p + q$ is the total number of parameters in the EM.

Since the model is correctly-structured, OLS estimation provides the Best Linear Unbiased Estimates (BLUE) for the model parameters, in the sense that the OLS parameter estimates have the smallest variance among all unbiased estimators (Beck and Arnold, 1977). Furthermore, any linear combination of the parameter estimates of the form $a^T \hat{\beta}_E$ also has the smallest variance among all unbiased estimators of $a^T \beta$, where a is a column vector of length m.

There are many ways in which a model that is used to represent a process can be misspecified (Rao, 1971; Hocking, 1976), including:

1) failure to include some or all of the explanatory variables (under-modelling);
2) inclusion of "extraneous" variables in the model (overfitting).

Even in the linear case, the true process may be complex, encompassing features such as a heteroskedastic structure for the stochastic components, or an error-in-variables structure for the explanatory variables (Beck and Arnold, 1977). The purpose of regression diagnostics is to evaluate model adequacy and reveal inadequacies.

The analysis in this Chapter only focuses on the consequences of under-modelling. In under-modelling, the analyst believes, or decides to use, a model of the form

$$Y = X_1\beta_1 + e \tag{2.8}$$

where $e = X_2\beta_2 + \varepsilon$ is the stochastic component combined with any model mismatch. The model in Eqn. (2.8) is denoted as the SM.

For the SM, only the parameters associated with the explanatory variables in X_1 are estimated by minimizing the objective function

$$S(\beta_1) = (Y - X_1\beta_1)^T(Y - X_1\beta_1) \tag{2.9}$$

with respect to β_1, resulting in the OLS estimates $\hat{\beta}_{1S}$ as

$$\hat{\beta}_{1S} = (X_1^T X_1)^{-1} X_1^T Y \tag{2.10}$$

where the subscript "*S*" indicates the use of a SM. It is readily verified that (Draper and Smith, 1998)

$$\begin{aligned} E(\hat{\beta}_{1S}) &= \beta_1 + A_1\beta_2 \\ Cov(\hat{\beta}_{1S}) &= \sigma^2(X_1^T X_1)^{-1} \end{aligned} \tag{2.11}$$

where $A_1 = (X_1^T X_1)^{-1} X_1^T X_2$ is the projection of X_1 on X_2. These parameter estimates are generally biased, in that $E(\hat{\beta}_{1S}) \neq \beta_1$, unless $A_1\beta_2 = 0$, which only occurs when $\beta_2 = 0$, or when X_1 and X_2 are orthogonal. The modeller may use the residuals from the SM to estimate the noise variance σ^2,

$$s_S^2 = \frac{S(\hat{\beta}_{1S})}{n-p} = \frac{(Y - X_1\hat{\beta}_{1S})^T(Y - X_1\hat{\beta}_{1S})}{n-p} \tag{2.12}$$

ignoring the influence of any model misspecification. The value of s_S^2 is often used to construct confidence intervals for parameter estimates and model predictions and to conduct model

adequacy tests. However, since s_S^2 is a biased estimator of σ^2 (Hocking, 1976), statistical tests that rely on s_S^2 can be misleading.

When misspecified models are analyzed (and the modeller believes that the model is misspecified), it is common for the quality of parameter estimates (and model predictions) to be assessed using mean-squared error (MSE) or mean-squared error matrix (MSEM) (Lowerre, 1974; Gunst and Mason, 1977; Price, 1982; Toutenburg and Trenkler, 1990), which account for both bias and variance. The MSEM and MSE for a parameter estimate $\hat{\beta}$ are defined as

$$MSEM(\hat{\beta}) = E\left((\hat{\beta} - \beta)(\hat{\beta} - \beta)^T\right) = Cov(\hat{\beta}) + \Delta\Delta^T \tag{2.13}$$

and

$$MSE(\hat{\beta}) = E\left((\hat{\beta} - \beta)^T(\hat{\beta} - \beta)\right) = tr\left(MSEM(\hat{\beta})\right) \tag{2.14}$$

where $\Delta = E(\hat{\beta}) - \beta$ is the bias, and $tr(\cdot)$ denotes the trace of the quantity $(\cdot)$.

Freund et al. (1961), Goldberg (1961), Goldberg and Jochems (1961), Kabe (1963), Wallace (1964), Toro-Vizcarrondo and Wallace (1968), Rao (1971), Hocking (1976) and Miller (1990) provide important results concerned with the MSE of parameter estimates and model predictions when a SM structure is selected.

2.4.1 Qualitative Statements Regarding Misspecification

The following results are known for misspecified models in the linear case (Rao, 1971; Hocking, 1976)

1) The omission of a variable from the EM (the truth) introduces bias and decreases the variance in all of the parameter estimates (and model predictions) obtained using the SM;

2) The MSEs of all parameter estimates (and model predictions) are decreased when a single variable in the EM is deleted whose corresponding true parameter value is smaller in magnitude than the theoretical standard deviation of its least-squares estimate;
3) The estimate of the noise variance (Eqn. (2.12)) obtained from the SM residuals is upwardly biased;
4) The inclusion of an irrelevant variable in the model increases the variance and MSEs of all of the parameter estimates (and model predictions).

2.4.2 Quantitative Statements Regarding Misspecification

To enable a comparison of the properties of parameter estimates from the SM and the EM, it is helpful to partition the expected values of the parameter estimates and their covariance matrix from the EM as follows:

$$\begin{pmatrix}\hat{\beta}_{1E}\\ \hat{\beta}_{2E}\end{pmatrix} \sim \left(\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}, \sigma^2 \begin{pmatrix}\Gamma & \Psi\\ \Psi^{\mathrm{T}} & \Omega\end{pmatrix}\right) \tag{2.15}$$

where

$$\begin{pmatrix}\Gamma & \Psi\\ \Psi^{\mathrm{T}} & \Omega\end{pmatrix} = \begin{pmatrix}X_1^T X_1 & X_1^T X_2\\ X_2^T X_1 & X_2^T X_2\end{pmatrix}^{-1} \tag{2.16}$$

The elements of the composite covariance matrix can be obtained from sub-matrices using standard results from linear algebra (Beck and Arnold, 1977)

$$\begin{aligned}\Gamma &= (X_1^T(I_n - P_2)X_1)^{-1} = (X_1^T X_1)^{-1} + A_1(X_2^T(I_n - P_1)X_2)^{-1}A_1^T\\ \Omega &= (X_2^T(I_n - P_1)X_2)^{-1} = (X_2^T X_2)^{-1} + A_2(X_1^T(I_n - P_2)X_1)^{-1}A_2^T\end{aligned} \tag{2.17}$$

where $P_1 = X_1(X_1^T X_1)^{-1}X_1^T$, $P_2 = X_2(X_2^T X_2)^{-1}X_2^T$, $A_2 = (X_2^T X_2)^{-1}X_2^T X_1$. A_2 is the projection of X_2 on X_1.

Based on the above theoretical results, comparisons between the SM and the EM can be made, both for parameter estimates and for model predictions (or other linear combinations of the parameters).

2.4.3 Comparison of Parameter Estimates

Table 2.1 summarizes the expected values, covariance matrices and MSEM of parameter estimates obtained using the SM and the EM. The expected noise variance estimates that would be obtained are given in the final row of the table.

Table 2.1: Comparison of parameter estimates and variance estimates from EM and SM

	Simplified Model (SM)	**Extended Model (EM)**
$E(\hat{\beta}_1)$	$\beta_1 + A_1\beta_2$	β_1
$E(\hat{\beta}_2)$	———	β_2
$Cov(\hat{\beta}_1)$	$\sigma^2(X_1^T X_1)^{-1}$	$\sigma^2(X_1^T X_1)^{-1} + \sigma^2 A_1(X_2^T(I_n - P_1)X_2)^{-1}A_1^T$
$Cov(\hat{\beta}_2)$	———	$\sigma^2(X_2^T(I_n - P_1)X_2)^{-1}$
$MSEM(\hat{\beta}_1)$	$\sigma^2(X_1^T X_1)^{-1} + A_1\beta_2\beta_2^T A_2^T$	$\sigma^2(X_1^T X_1)^{-1} + \sigma^2 A_1(X_2^T(I_n - P_1)X_2)^{-1}A_1^T$
$E(s^2)$	$\sigma^2 + \frac{\beta_2^T X_2^T(I_n - P_1)X_2\beta_2}{n-p}$	σ^2

Mean-Based Comparison of Parameter Estimates

The parameter estimates from the correctly specified EM are unbiased. The parameter estimates from the SM are biased except when $A_1 = 0$, which requires that X_2 is orthogonal to X_1, a situation not likely to be encountered in practice.

Variance-Based Comparison of Parameter Estimates

From Table 2.1, the difference between covariance matrices for $\hat{\beta}_{1E}$ and $\hat{\beta}_{1S}$ is

$$Cov(\hat{\beta}_{1E}) - Cov(\hat{\beta}_{1S}) = \sigma^2 A_1(X_2^T(I_n - P_1)X_2)^{-1}A_1^T = \sigma^2 A_1 \Omega A_1^T \qquad (2.18)$$

Since $Cov(\hat{\beta}_{2E}) = \sigma^2\Omega$, therefore Ω is positive definite, and the above difference will be positive semi-definite (Zhang, 1999). Let β_{1i} be the i^{th} element in β_1 ($i = 1,2,\dots,p$). Using properties of positive semi-definite matrices (Zhang, 1999), the following inequalities are satisfied

$$\begin{aligned} &\text{Individual Parameters:} && Var(\hat{\beta}_{1iS}) \le Var(\hat{\beta}_{1iE}) \\ &\text{Total Variance:} && tr\left(Cov(\hat{\beta}_{1S})\right) \le tr\left(Cov(\hat{\beta}_{1E})\right) \\ &\text{Generalized Variance (determinant):} && \left|Cov(\hat{\beta}_{1S})\right| \le \left|Cov(\hat{\beta}_{1E})\right| \end{aligned} \tag{2.19}$$

Note that, the parameter estimates from the SM are biased, so the Gauss-Markov Theorem does not apply (Beck and Arnold, 1977).

In the case of overfitting, the SM would be correctly-specified and the EM would contain redundant parameters, since the true values of parameters in β_2 are zero. Both the SM and the EM would lead to unbiased parameter estimates because they are both correctly-structured (Rao, 1971). However, the inclusion of the extraneous parameters results in less precision in the parameter estimates and in the predictions made using these parameters. In this scenario, the true covariance matrix of the parameters would be given by the entry in Table 2.1 under the column "Simplified Model (SM)," and the covariance matrix of the parameters for overfitting would be given by the entry in Table 2.1 under the column "Extended Model (EM)."

MSEM-Based Comparison of Parameter Estimates

From Table 2.1, the MSEM difference for $\hat{\beta}_{1E}$ and $\hat{\beta}_{1S}$ is

$$MSEM(\hat{\beta}_{1E}) - MSEM(\hat{\beta}_{1S}) = A_1(\sigma^2(X_2^T(I_n - P_1)X_2)^{-1} - \beta_2\beta_2^T)A_1^T \tag{2.20}$$

This difference is positive semi-definite if

$$\sigma^2(X_2^T(I_n - P_1)X_2)^{-1} - \beta_2\beta_2^T \ge 0 \tag{2.21}$$

A necessary and sufficient condition of inequality (2.21) is that (Wang and Chow, 1994)

$$\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2 \leq \sigma^2 \tag{2.22}$$

This inequality holds when the SM gives better (in sense of smaller MSEM) parameter estimates than the EM.

Inequality (2.21) has several appealing interpretations. First, the last entry in Table 2.1 can be re-arranged as

$$\frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{\sigma^2 (n-p)} = \frac{E(s_S^2) - \sigma^2}{\sigma^2} \tag{2.23}$$

The right-hand side (RHS) of Eqn. (2.23) is the fractional increase in the expected noise variance estimate that arises when the SM is used. If inequality (2.22) holds, then

$$\frac{E(s_S^2) - \sigma^2}{\sigma^2} \leq \frac{1}{n-p} \tag{2.24}$$

Inequality (2.24) provides an upper bound on the bias of the noise variance estimate obtained from the SM when inequality (2.22) is satisfied (i.e., SM parameter estimates are better than the EM estimates).

Rearranging inequality (2.22), a critical ratio R_C can be defined as

$$R_C = \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{q\sigma^2} \tag{2.25}$$

where q is the number of parameters contained in β_2. Inequality (2.22) is then equivalent to

$$R_C \leq \frac{1}{q} \tag{2.26}$$

Examination of Eqn. (2.25) reveals that R_C becomes smaller (implying the SM tends to be preferred to the EM) in the following situations

1) when there are high noise levels, i.e., σ^2 is large;
2) when the true absolute values of parameters in β_2 are small;

3) when there are high correlations among input variable settings,

   i.e., the trace of $X_2^T(I_n - P_1)X_2$ is small;

4) when there are a limited number of experiments or a limited range of input conditions,

   i.e., the trace of $X_2^T(I_n - P_1)X_2$ is small.

Proof of statements 1) and 3) above can be found in Abdullaev and Geidarov (1985).

2.4.4 Comparison of Model Predictions

Model developers and users often care more about model predictions or other linear combinations of parameters than the parameter estimates themselves. Imagine that the model parameters have been estimated using a matrix of input variable settings, $X \in \mathbb{R}^{n\times m}$, and then model predictions are made at other input settings, $Z \in \mathbb{R}^{w\times m}$, where w is the total number of prediction points. Z can be partitioned in the same way as X into $Z_1 \in \mathbb{R}^{w\times p}$ and $Z_2 \in \mathbb{R}^{w\times q}$ corresponding to the partitioned parameters in Eqn. (2.15). Two types of model predictions can be made

$$\begin{aligned} &\text{SM Predictions:} && \hat{Y}_S = Z_1\hat{\beta}_{1S} \\ &\text{EM Predictions:} && \hat{Y}_E = Z\hat{\beta}_E = Z_1\hat{\beta}_{1E} + Z_2\hat{\beta}_{2E} \end{aligned} \tag{2.27}$$

Expressions for expected values, covariance matrices and the MSEM for these predictions are shown for two cases: $Z = X$ (Table 2.2) and $Z \neq X$ (Table 2.3). There is no extrapolation or interpolation in the first case, because predictions are made at the same experimental conditions under which the data were collected. The second case corresponds to a new set of conditions for which model predictions are desired.

Mean-Based Comparison of Model Predictions

As seen in the first row of Table 2.2 and Table 2.3, only the correctly-structured EM provides unbiased model predictions.

Variance-Based Comparison of Model Predictions

From Table 2.3, in general, the difference between covariance matrices for $\hat{Y}_E$ and $\hat{Y}_S$ is

$$Cov(\hat{Y}_E) - Cov(\hat{Y}_S) = \sigma^2(Z_1A_1 - Z_2)\Omega(Z_1A_1 - Z_2)^T \quad (2.28)$$

Since Ω is positive definite, the above difference is positive semi-definite, which means predictions from the SM cannot be more variable than those from the EM.

Table 2.2: Model predictions when $\boldsymbol{Z = X}$.

	SM Predictions $\hat{Y}_S$	EM Predictions $\hat{Y}_E$
$E(\hat{Y})$	$X_1\beta_1 + X_1A_1\beta_2$	$X_1\beta_1 + X_2\beta_2$
$Cov(\hat{Y})$	σ^2P_1	$\sigma^2P_1 + \sigma^2(I_n - P_1)X_2\Omega X_2^T(I_n - P_1)$
$MSEM(\hat{Y})$	$\sigma^2P_1 + (I_n - P_1)X_2\beta_2\beta_2^TX_2^T(I_n - P_1)$	$\sigma^2P_1 + \sigma^2(I_n - P_1)X_2\Omega X_2^T(I_n - P_1)$

Table 2.3: Model predictions when $\boldsymbol{Z \neq X}$.

	SM Predictions $\hat{Y}_S$	EM Predictions $\hat{Y}_E$
$E(\hat{Y})$	$Z_1\beta_1 + Z_1A_1\beta_2$	$Z_1\beta_1 + Z_2\beta_2$
$Cov(\hat{Y})$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T$ $+\sigma^2(I_n - P_1)X_2\Omega X_2^T(I_n - P_1)$
$MSEM(\hat{Y})$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T$ $+(I_n - P_1)X_2\beta_2\beta_2^TX_2^T(I_n - P_1)$	$\sigma^2Z_1(X_1^TX_1)^{-1}Z_1^T$ $+\sigma^2(I_n - P_1)X_2\Omega X_2^T(I_n - P_1)$

MSEM-Based Comparison of Model Predictions

The elements of the MSEM contain the covariance for the model predictions, plus the squared bias. From Table 2.3, the difference between the MSEM for $\hat{Y}_E$ and $\hat{Y}_S$ is

$$MSEM(\hat{Y}_E) - MSEM(\hat{Y}_S)$$
$$= (Z_1A_1 - Z_2)(\sigma^2(X_2^T(I_n - P_1)X_2)^{-1} - \beta_2\beta_2^T)(Z_1A_1 - Z_2)^T \tag{2.29}$$

Following the same argument as for inequalities (2.21), (2.22) and (2.26), if $R_C \leq 1/q$, then $MSEM(\hat{Y}_S) \leq MSEM(\hat{Y}_E)$. This means that when few data points are available or the data are noisy or when there are high correlations between X_1 and X_2, the SM can be expected to provide better predictions than the properly-specified EM.

A special case is that, when $Z = X$, the MSEM difference becomes

$$MSEM(\hat{Y}_E) - MSEM(\hat{Y}_S)$$
$$= (I_n - P_1)X_2(\sigma^2(X_2^T(I_n - P_1)X_2)^{-1} - \beta_2\beta_2^T)X_2^T(I_n - P_1) \tag{2.30}$$

and the MSE difference is

$$MSE(\hat{Y}_E) - MSE(\hat{Y}_S)$$
$$= tr\left((I_n - P_1)X_2(\sigma^2(X_2^T(I_n - P_1)X_2)^{-1} - \beta_2\beta_2^T)X_2^T(I_n - P_1)\right) \tag{2.31}$$
$$= \sigma^2 q - \beta_2^T X_2^T(I_n - P_1)X_2\beta_2$$

In this special case, the SM is better if and only if

$$R_C = \frac{\beta_2^T X_2^T(I_n - P_1)X_2\beta_2}{q\sigma^2} \leq 1 \tag{2.32}$$

Note that Inequality (2.32) is a necessary and sufficient condition for $MSE(\hat{Y}_S) \leq MSE(\hat{Y}_E)$ and is less restrictive than inequality (2.26). However, it only holds when model predictions are made using exactly the same input-variable settings as those used in parameter estimation. If parameter estimation or extrapolation is the main purpose, inequality (2.26) is more appropriate (Hocking, 1976). When there is only one variable in X_2 ($q = 1$), inequalities (2.26) and (2.32) become the same.

In summary, the literature on misspecified linear models provides information about the conditions under which a modeller can expect to get improved parameter estimates and model predictions when a SM is used. Unfortunately, the conditions in inequalities (2.26) and (2.32) depend on the true values of the parameters in β_2 and on the true noise variance σ^2. In practical applications, the modeller does not know these true values, but can obtain estimated values, $\hat{\beta}_{2E}$ and s_E^2, from the data, assuming the correctly-structured EM is available.

2.5 Strategy for Assessing Uncertainty about which Model is Better

If estimated parameter values and noise variances from the EM are available, then an estimate of R_C is

$$\hat{R}_C = \frac{\hat{\beta}_{2E}^T X_2^T (I_n - P_1) X_2 \hat{\beta}_{2E}}{q s_E^2} \tag{2.33}$$

If the stochastic component ε in model (2.4) is also assumed to be Normally distributed, then based on partial F tests (Montgomery et al., 2001), $\hat{R}_C$ follows a noncentral F distribution $F_{q,n-m}(\delta)$ with q and $n-m$ degrees of freedom. The noncentrality parameter δ is

$$\delta = \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{\sigma^2} = q R_C \tag{2.34}$$

When $\beta_2 = 0$, $\hat{R}_C$ follows the more widely-known central F distribution $F_{q,n-m}$, which is often used to test the null hypothesis that $\beta_2 = 0$ (Montgomery et al., 2001). Note that $\hat{\beta}_{2E}$ and s_E^2 are obtained by fitting parameters in the EM, so the analyst must have access to an EM that she believes to be well-structured to compute $\hat{R}_C$ in Eqn. (2.33).

The construction of confidence intervals and hypothesis tests for many standard statistics that follow standard distributions (such as the Normal distribution, Student's t distribution, central χ^2

distribution and central F distribution) is well established in introductory statistics textbooks. However, the construction of confidence intervals and hypothesis tests for R_C is complicated by the fact that $\hat{R}_C$ follows a noncentral F distribution $F_{q,n-m}(\delta)$ with unknown noncentrality parameter δ. As a result, an iterative algorithm is required to find appropriate confidence intervals. The following steps (Steiger, 2004) can be used to calculate (δ_L, δ_U), the exact two-sided $100(1-\alpha)\%$ confidence interval for the noncentrality parameter δ.

1) Calculate the cumulative probability p_C corresponding to $\hat{R}_C$ using the central F distribution with q and $n-m$ degrees of freedom. If p_C is less than $\alpha/2$, then both δ_L and δ_U are zero. The reason for this conclusion is that $\delta \geq 0$ by definition, and, if a one-sided hypothesis test is performed at the $100(\alpha/2)\%$ significance level, we could reject the null hypothesis that δ is zero or larger than zero. If p_C is less than $1-\alpha/2$, $\delta_L = 0$ and δ_U is calculated using Step 3. Otherwise, calculate δ_L and δ_U using Step 2 and Step 3, respectively.
2) To calculate the lower limit, δ_L, iterate on the noncentrality parameter, so that the $1-\alpha/2$ cumulative probability point (the critical value) of a noncentral F distribution with q and $n-m$ degrees of freedom equals $\hat{R}_C$. This value is unique.
3) To calculate the upper limit, δ_U, iterate on the noncentrality parameter, so that the $\alpha/2$ cumulative probability point (the critical value) of a noncentral F distribution with q and $n-m$ degrees of freedom equals $\hat{R}_C$. This value is unique.

Since $\delta = qR_C$, the two-sided $100(1-\alpha)\%$ confidence interval for R_C is $(\delta_L/q\,, \delta_U/q)$. This confidence interval for R_C contains all values of the null hypothesis that would not be rejected at the $100(1-\alpha)\%$ confidence level when testing the alternative hypothesis that $R_C \neq k$. Two values of k are of interest: $k = 1/q$ is used to test whether inequality (2.26) is satisfied

and $k = 1$ is used to test inequality (2.32). If $k < \delta_L/q$, we can be $100(1-\alpha/2)\%$ certain that $R_C > k$ and that the EM is better than the SM. If $k > \delta_U/q$, we can be $100(1-\alpha/2)\%$ certain that $R_C < k$ and that the SM is better than the EM. In the special case when only one parameter in the EM was not included in the SM ($q = 1$), inequalities (2.26) and (2.32) become the same. The confidence intervals are readily computed using the cumulative noncentral *F* distribution function in Matlab™ or other statistical software package.

2.6 Confidence Intervals for Parameter Estimates and Model Predictions from the SM

In situations where the modeller has decided to use the misspecified SM, it is desirable to obtain appropriate confidence intervals for the parameter estimates and model predictions. Such confidence intervals rely on good estimates of variance-covariance matrices. In the literature, three methods have been proposed for estimating variance-covariance matrices for parameter estimates: 1) the conventional method that assumes the SM is correctly-structured (Montgomery et al., 2001); 2) the sandwich estimator (White, 1981; Seber and Wild, 2003; Waldorp et al., 2005; Waldorp et al., 2006); and 3) nonparametric bootstrapping (Efron and Tibshirani, 1994; Montgomery et al., 2001; Martinez and Martinez, 2002; Good, 2006; Waldorp et al., 2006). The last two methods have been recommended for use under model misspecification (Waldorp et al., 2006).

In the conventional method, variance-covariance matrix of $\hat{\beta}_{1S}$ can be estimated by

$$\hat{\Sigma}_{CONV} = s^2(X_1^T X_1)^{-1} \tag{2.35}$$

where s^2 is an estimate of the noise variance. s^2 could be determined from: 1) replicate runs (pooled variances) if replicates are available; 2) the EM (s_E^2), if the correctly-structured extended

model is available; 3) the SM (s_S^2); and 4) nonparametric bootstrapping (s_B^2). The conventional method for estimating the variance-covariance matrix for the parameter estimates requires the assumption that the SM is correctly-structured ($\beta_2 = 0$). The estimated variance-covariance matrix, and the conventional confidence bounds that are determined from its diagonal elements, rely on the goodness of the noise variance estimate and the appropriateness of the assumption that $\beta_2 = 0$.

White (1981) proposed that the sandwich estimator, which is robust to model misspecification, should be used to estimate the variance-covariance matrix for parameter estimates when a misspecified model is used. The sandwich estimator is derived directly from the data without the requirement for a correctly-specified EM or a separate noise variance estimate from replicate experiments. The variance-covariance matrix of parameter estimates is defined as

$$\hat{\Sigma}_{SANW} = (X_1^T X_1)^{-1} \left(\sum_{i=1}^{n} \tilde{x}_i^T \tilde{x}_i \hat{e}_i^2 \right) (X_1^T X_1)^{-1} \tag{2.36}$$

where $\tilde{x}_i$ is the i^{th} row of X_1 and $\hat{e}_i$ is the i^{th} residual from the SM ($i = 1,2, \dots, n$).

The sandwich estimator is a consistent estimator for the true variance-covariance matrix of parameter estimates even when the model is misspecified (White, 1981). The sandwich estimator has been investigated for models that are nonlinear in the parameters (Donaldson and Schnabel, 1987), and has been used to obtain good estimate of the Cramér-Rao bound for the use of the Wald test in situations when the model is misspecified (Waldorp et al., 2005). Additionally, it is shown that sandwich estimator is robust against an incorrect assumption on the noise covariance (Waldorp et al., 2005; Waldorp et al., 2006).

Nonparametric bootstrapping is a computationally intensive procedure commonly-used in situations when no analytical methods are available for determining confidence intervals and

when the sample is representative of the population (Montgomery et al., 2001; Martinez and Martinez, 2002). The bootstrapping algorithm proceeds as follows

1) Resample the original data (X_1, Y) B times with replacement, and for each re-sampled data set, estimate β_1 and the noise variance σ^2;
2) The final noise variance estimate is obtained as the average of the B individual noise variance estimates (Good, 2006),

$$s_B^2 = \frac{1}{B}\sum_{i=1}^{B} s_*^{2,i} \tag{2.37}$$

where $s_*^{2,i}$ is the i^{th} noise variance estimate ($i = 1,2, \dots, B$), and is calculated as

$$s_*^{2,i} = \frac{1}{n-p}\left(Y^i - X_1^i\hat{\beta}_{1*}^i\right)^T\left(Y^i - X_1^i\hat{\beta}_{1*}^i\right) \tag{2.38}$$

$\left(X_1^i, Y^i\right)$ is the i^{th} re-sampled pair, and $\hat{\beta}_{1*}^i$ is the i^{th} estimate of β_1 based on $\left(X_1^i, Y^i\right)$.

3) The variance-covariance matrix for the parameter estimates is obtained directly from the B sets of parameter estimates as

$$\hat{\Sigma}_{BOOT} = \frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\beta}_{1*}^i - \bar{\beta}_{1*}\right)\left(\hat{\beta}_{1*}^i - \bar{\beta}_{1*}\right)^T \tag{2.39}$$

where $\bar{\beta}_{1*}$ is the average of B estimates of β_1 (Waldorp et al., 2006).

Bootstrapping can be used to estimate the bias and to directly construct the confidence intervals for parameter estimates and model predictions. A detailed description of different algorithms and the limitations of each algorithm can be found in Efron and Tibshirani (1994). Matlab™ codes are available from Martinez and Martinez (2002). Chernick (1999) also described some examples from the literature when the bootstrapping method should not be used. Waldorp et al. (2006) showed that the nonparametric bootstrapping give better results than the conventional methods, especially when the noise is correlated.

2.7 Illustrative Example

2.7.1 Theoretical Results

The theoretical analysis provided in the previous section is illustrated using a very simple example with three parameters in the SM ($p = 3$) and one additional parameter ($q = 1$) in the EM. Let the design matrix for the SM be $X_1 = (X_{11}\ X_{12}\ X_{13})$, an orthogonal matrix (obtained from a designed experiment) containing entries of $\pm r$. X_{1i} is the i^{th} column in X_1 ($i = 1,2,3$). It is of interest in this analysis whether or not the additional parameter in β_2 contained in the EM should be estimated. Assume that the vector of experimental settings corresponding to β_2 is correlated with X_{11}, the first column of X_1, which makes it difficult to obtain an independent estimate of β_2. In this example, the amount of correlation between X_{11} and X_2 is adjusted using a design factor λ, where $0 \leq \lambda \leq 1$. Let $X_2 = \lambda X_{11} + \sqrt{1-\lambda^2}W$, where W is a $n \times 1$ vector with entries of $\pm r$, and is orthogonal to all columns in X_1. The following analysis will consider different experimental designs, corresponding to different values of λ. When $\lambda = 0$, X_{11} and X_2 are uncorrelated, and when $\lambda = 1$, $X_{11} = X_2$. This example is selected because it can readily be used to study the influence of various factors (e.g. correlation in the experimental design, input range and number of data points) on whether the SM or the EM will give better predictions.

The EM is described by

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\ &= X_{11}\beta_{11} + X_{12}\beta_{12} + X_{13}\beta_{13} + X_2\beta_2 + \varepsilon \end{aligned} \tag{2.40}$$

where β_{1i} is the i^{th} parameter in β_1, and ε is independently and identically distributed with mean 0 and variance σ^2 following a Normal distribution.

When $n = 16$ data points are used and the input range for all independent variables in X_1 and W is $r = 1$, the input settings are

$$X_1 = (X_{11} \quad X_{12} \quad X_{13}) = \begin{pmatrix} 1 & -1 & -1 \\ -1 & -1 & -1 \\ 1 & 1 & -1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & -1 & -1 \\ 1 & 1 & -1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix} \qquad W = \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{2.41}$$

$$X_2 = \lambda X_{11} + \sqrt{1 - \lambda^2} W$$

Based on these input settings, the covariance matrix of the parameter estimates obtained using the EM is

$$Cov(\hat{\beta}_E) = Cov \begin{pmatrix} \hat{\beta}_{11E} \\ \hat{\beta}_{12E} \\ \hat{\beta}_{13E} \\ \hat{\beta}_{2E} \end{pmatrix} = \frac{\sigma^2}{nr^2} \begin{pmatrix} \frac{1}{1-\lambda^2} & 0 & 0 & \frac{-\lambda}{1-\lambda^2} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{-\lambda}{1-\lambda^2} & 0 & 0 & \frac{1}{1-\lambda^2} \end{pmatrix} \tag{2.42}$$

As $\lambda \to 1$, $\sigma^2/\left(nr^2(1-\lambda^2)\right)$, which is the variance of $\hat{\beta}_{11E}$ and $\hat{\beta}_{2E}$, increases dramatically, and the correlation between $\hat{\beta}_{11E}$ and $\hat{\beta}_{2E}$ approaches -1. However, since X_2 is orthogonal to both X_{12} and X_{13}, the variances of $\hat{\beta}_{12E}$ and $\hat{\beta}_{13E}$ are not affected by λ.

The SM is described as

$$\begin{aligned} Y &= X_1\beta_1 + e \\ &= X_{11}\beta_{11} + X_{12}\beta_{12} + X_{13}\beta_{13} + e \end{aligned} \tag{2.43}$$

The expected values of the parameter estimates (from Eqn. (2.10)) obtained using the SM are

$$E(\hat{\beta}_{1S}) = E\begin{pmatrix} \hat{\beta}_{11S} \\ \hat{\beta}_{12S} \\ \hat{\beta}_{13S} \end{pmatrix} = \begin{pmatrix} \beta_{11} + \lambda\beta_2 \\ \beta_{12} \\ \beta_{13} \end{pmatrix} \tag{2.44}$$

It is seen that $\hat{\beta}_{11S}$ will be biased unless $\lambda = 0$ or $\beta_2 = 0$. $\hat{\beta}_{12S}$ and $\hat{\beta}_{13S}$ are unbiased. The covariance matrix of $\hat{\beta}_{1S}$ is

$$Cov(\hat{\beta}_{1S}) = Cov\begin{pmatrix} \hat{\beta}_{11S} \\ \hat{\beta}_{12S} \\ \hat{\beta}_{13S} \end{pmatrix} = \frac{\sigma^2}{nr^2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.45}$$

Compared with Eqn. (2.42)

$$Var(\hat{\beta}_{11S}) \leq Var(\hat{\beta}_{11E}) \tag{2.46}$$

These two variances are equal only when there is no correlation between X_{11} and X_2 ($\lambda = 0$).

Based on the results summarized in Table 2.1 and Table 2.2, the MSE of parameter estimates and model predictions made at the design points obtained using the SM and the EM are

$$\begin{aligned} MSE(\hat{\beta}_{1S}) &= \frac{\sigma^2 p}{nr^2} + \beta_2^2\lambda^2 \\ MSE(\hat{\beta}_{1E}) &= \frac{\sigma^2 p}{nr^2} + \frac{\lambda^2\sigma^2}{nr^2(1-\lambda^2)} \\ MSE(\hat{Y}_S) &= \sigma^2 p + nr^2(1-\lambda^2)\beta_2^2 \\ MSE(\hat{Y}_E) &= \sigma^2(p+1) \end{aligned} \tag{2.47}$$

From the MSE expressions given in Eqn. (2.47),

$$R_C = \frac{nr^2(1-\lambda^2)\beta_2^2}{\sigma^2} \tag{2.48}$$

The combination of $(\sigma^2, n, r, \lambda, \beta_2)$ that satisfies inequalities (2.26) and (2.32) are very clear. R_C is small when: 1) σ^2 is large (high noise levels); 2) n is small (few data points); 3) r is small (small range of input settings); 4) $\lambda \to 1$ (strong correlation among input variables in the SM with the remaining variables in the EM); and 5) β_2^2 is small (small absolute values for the excluded parameters). This example will be used in Monte Carlo simulations described below.

2.7.2 Monte Carlo Simulations

In this section, Monte Carlo simulations are performed to illustrate the theoretical analysis in the previous sections. Note that, since there is only one additional parameter in the EM ($q = 1$), inequalities (2.26) and (2.32) become the same. We consider a set of experiments with $n = 16$ data points, and the input range $r = 1$, as described in Eqn. (2.41) of earlier. Let the true parameter values be

$$\beta^T = (\beta_{11} \quad \beta_{12} \quad \beta_{13} \quad \beta_2)^T = (1 \quad -1 \quad 1 \quad -1)^T \tag{2.49}$$

and the noise variance $\sigma^2 = 1$. For this example, R_C is

$$R_C = \frac{nr^2(1-\lambda^2)\beta_2^2}{\sigma^2} = 16(1-\lambda^2) \tag{2.50}$$

Based on Eqn. (2.50), the range of λ values that ensures $R_C \leq 1$ is $0.968 \leq \lambda \leq 1$. If λ is in this interval, the SM is better than the EM in the sense of MSE for parameter estimates and for model predictions. To numerically demonstrate several of the concepts, let $\lambda = 0.99$, which gives $R_C = 0.3184$.

Comparison of Parameter Estimates and Model Predictions from the SM and the EM

A total of 1000 simulated data sets were generated using different random noise sequences. Figure 2.1 compares parameter estimates from the SM and the EM using boxplots.

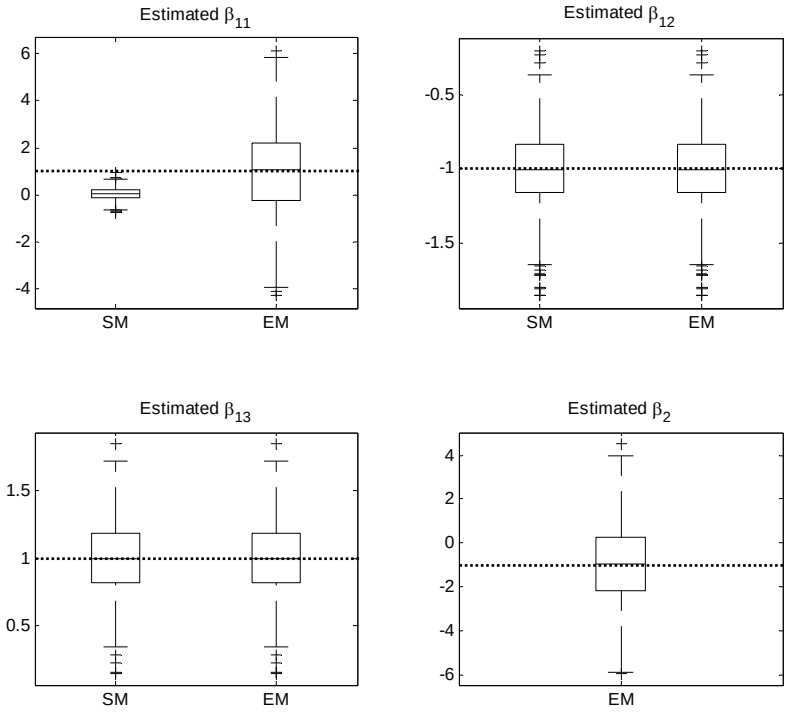

Figure 2.1: Boxplot comparison of parameter estimates from the SM and the EM. ····· shows the true parameter values used in the simulation

Figure 2.1 shows that $\hat{\beta}_{11S}$ has much smaller variance than $\hat{\beta}_{11E}$ due to the strong correlation between X_2 and X_{11}, but $\hat{\beta}_{11S}$ is biased. Point estimates of parameters β_{12} and β_{13} from both models are the same, because X_2 is orthogonal to X_{12} and X_{13}.

Figure 2.2 shows the model predictions made at the 1st, 5th, 11th and 16th observation points (model predictions made at other observation points have similar behaviour). Model predictions from the EM have larger variances than those from the SM, which are biased. As expected, predictions from the SM are better, on the average, than those from the EM.

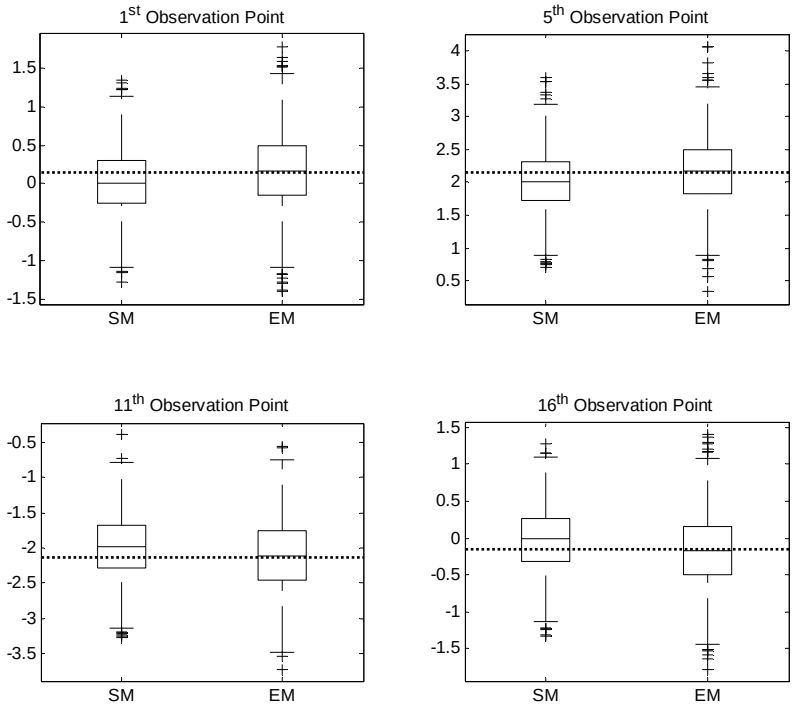

Figure 2.2: Boxplot comparison of model predictions from the SM and the EM. ····· shows the noise-free response at given observation point

Deciding Whether to Use the SM or the EM

To illustrate the construction of confidence intervals for R_C using available data, two cases are considered (i.e., $\hat{R}_C = 0.6585$ and $\hat{R}_C = 1.5821$ could be obtained from different simulated data set). Recall that the true value is $R_C = 0.3184$. The first value, $\hat{R}_C = 0.6585$ corresponds to the *median* value from the probability density function for $\hat{R}_C$ ((which is distributed as $F_{1,12}(\delta)$ with $\delta = qR_C = 0.3184$). The second value, $\hat{R}_C = 1.5821$, corresponds to the *mean* value of the same distribution. The large difference between median and mean indicates that the distribution is right-skewed.

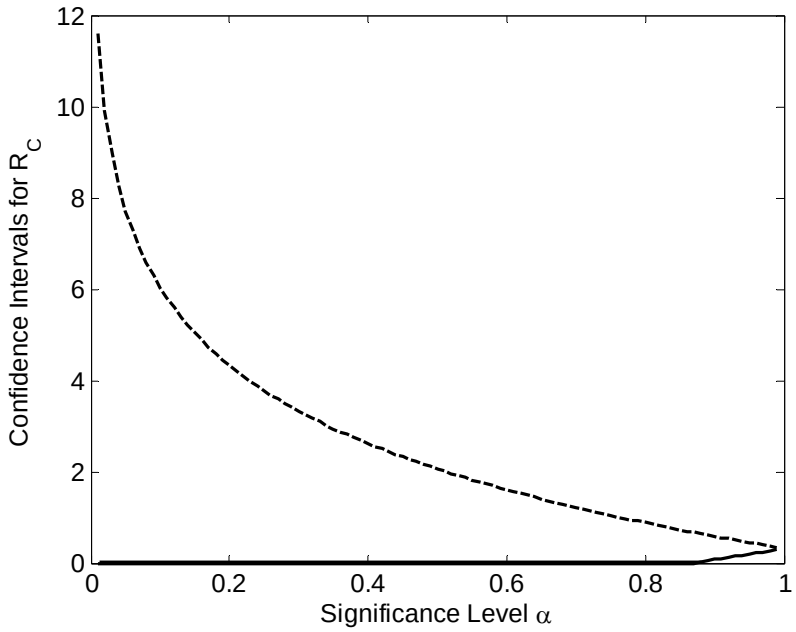

Figure 2.3: Two-sided confidence intervals for R_C for different values of the significance level α when $\hat{R}_C = 0.6585$. ---- upper confidence bound — lower confidence bound

Using the algorithm described in the section 2.5, the two-sided $100(1-\alpha)\%$ confidence interval for R_C is constructed for different values of α. When $\hat{R}_C = 0.6585$, the upper and lower limits are plotted in Figure 2.3. As α increases, the confidence limits narrow and converge to $\hat{R}_C = 0.6585$. The limits are very wide for typical values of α that are commonly-recommended for confidence intervals (i.e., $\alpha \leq 0.1$). Although $R_C = 0.3184 < 1$, which indicates that the SM is better than the EM, the results in Figure 2.3 show that, for a reasonable value of α (near 0.10), the two-sided confidence interval for R_C is $(0, 6)$. Since $k = 1$ is within this range, we are unable to distinguish whether the SM gives better predictions than the EM (this is a Type II error).

The confidence limits obtained from the mean value of $\hat{R}_C = 1.5821$ are shown in Figure 2.4. For α near 0.10, the two-sided confidence interval is $(0, 8.5)$, which is broader than that obtained in Figure 2.3.

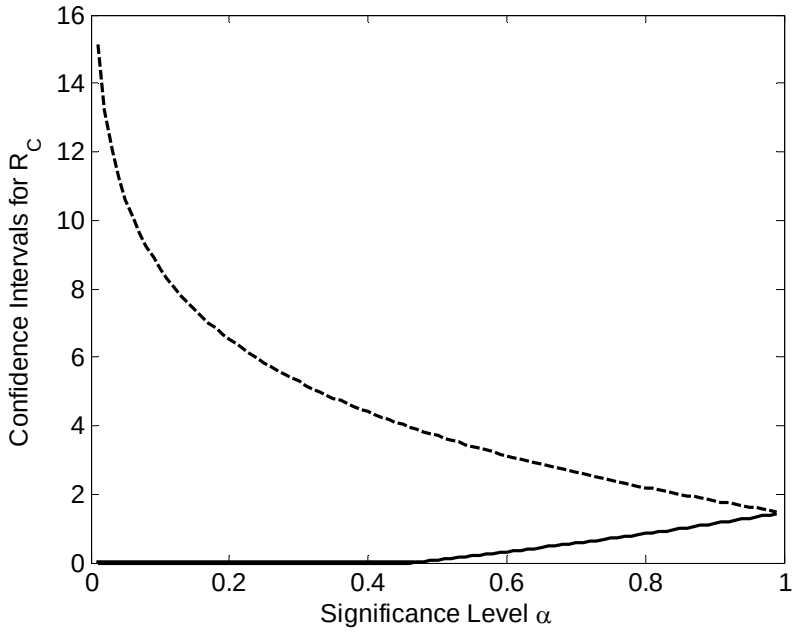

Figure 2.4: Two-sided confidence intervals for R_C for different values of the significance level α when $\hat{R}_C = 1.5821$. ---- upper confidence bound — lower confidence bound

Note that the confidence intervals in Figure 2.3 and Figure 2.4 are exact (Steiger, 2004). The high probability of mistakenly accepting the null hypothesis (unable to conclude that the SM is significantly better than the EM) arises from the limited data (few data points, limited input range and correlated experimental design). In situations where there are more data points or there are more parameters in the EM ($q > 1$), the confidence intervals for R_C become narrower (higher degrees of freedom in the noncentral *F* distribution). Narrower confidence intervals lead to better discrimination about which model is better.

Statistical Inference based on the SM

The following analysis compares various estimates for the variance of the additive noise ε and the variance of parameter estimates obtained using the SM. Two cases are considered: 1) $\lambda = 0.90$, which corresponds to $R_C = 3.04 > 1$, so that the EM is better; and 2) $\lambda = 0.99$, which corresponds to $R_C = 0.3184 < 1$, so that the SM is better (on average).

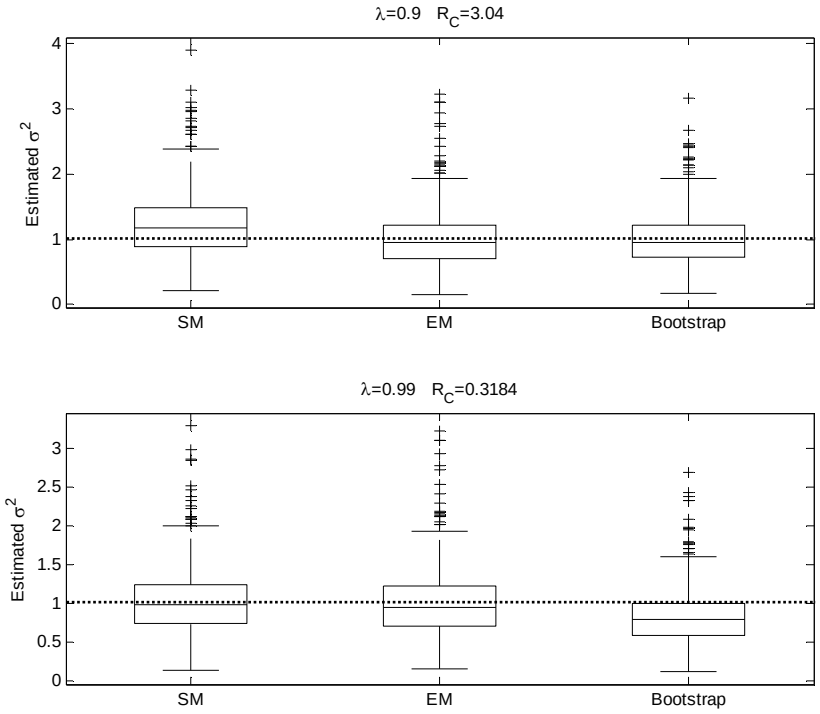

Figure 2.5: Comparison of estimated noise variance from the SM (s_S^2), the EM (s_E^2) and nonparametric bootstrapping (s_B^2)

As there are no replicate experiments available in this example, this analysis only consider using s_S^2 (from the SM residuals), s_E^2 (from the EM residuals) and s_B^2 (from nonparametric bootstrapping) to estimate the noise variance σ^2. To evaluate these three methods, a total of 1000

data sets were generated using different random noise sequences. For each simulated data set, $B = 200$ bootstraps were performed. In all simulations, it is assumed that the noise is independently and identically distributed following a standard Normal distribution. The estimated noise variances from each situation ($\lambda = 0.90$ and $\lambda = 0.99$) are compared using boxplots in Figure 2.5. The dotted line is the true value of the noise variance used in the simulation ($\sigma^2 = 1$). It is seen that, in the case when the EM is better, s_S^2 is biased upward, and s_E^2 and s_B^2 provide good estimate of σ^2. However, in the case when the SM is better, s_S^2 and s_E^2 are good estimates of σ^2. Based on these simulations, it seems that, when the correctly-structured EM is available, it should be used to estimate the noise variance because it provides an unbiased estimate. However, this issue requires further investigation because bias is not the entire problem. Perhaps other variance estimators will provide estimates with lower mean-squared error.

The estimated variance of $\hat{\beta}_{11S}$ from the SM could be obtained by three methods: 1) conventional methods (Eqn. (2.35) with s_S^2 (SM) or s_E^2 (EM)); 2) sandwich estimator (Eqn. (2.36)); and 3) nonparametric bootstrapping (Eqn. (2.39)). The results from each situation (for both $\lambda = 0.90$ and $\lambda = 0.99$) are compared in Figure 2.6. Since $\hat{\beta}_{12S}$ and $\hat{\beta}_{13S}$ are the same as those obtained using the EM (Figure 2.1), they are not considered in this analysis.

Figure 2.6 shows that, for this particular example, all methods considered give similar results. Several nonparametric bootstrapping algorithms can be directly used to construct the confidence intervals for parameter estimates. There is no noticeable difference between the results obtained using the algorithms described by Efron and Tibshirani (1994) and by Martinez and Martinez (2002).

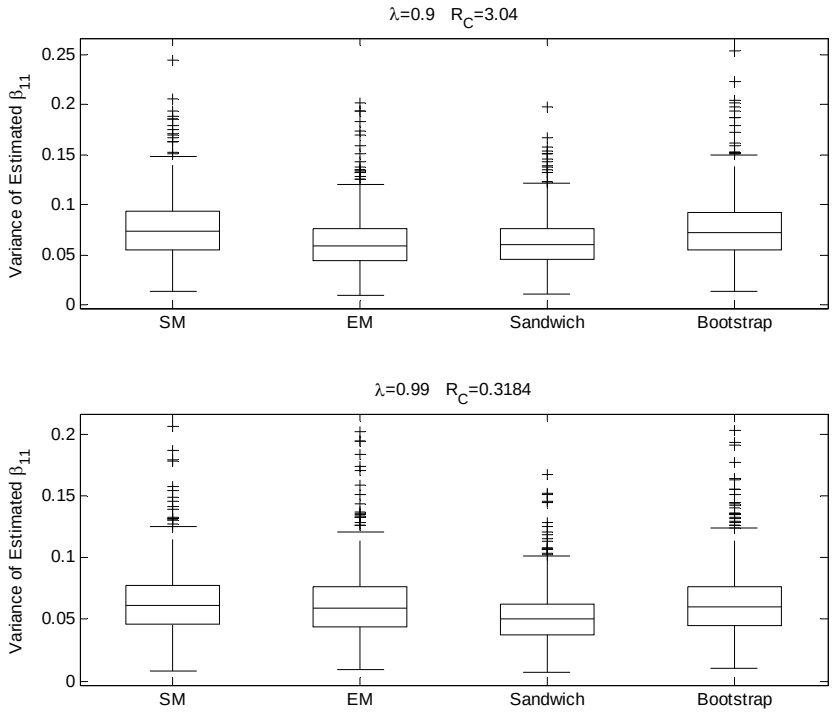

Figure 2.6: Comparison of estimated variance of $\widehat{\boldsymbol{\beta}}_{11S}$ from: 1) conventional methods (based on the SM (s_S^2) and EM (s_E^2)); 2) sandwich estimator; and 3) nonparametric bootstrapping

2.8 Conclusions

In this Chapter, a number of important issues related to the use of simplified or misspecified models have been reviewed. Much of the statistical literature has focused on validation and the use of models that are assumed to be statistically valid. There are, however, many instances when it is impossible to construct a model that is deemed statistically acceptable. There are other instances where the use of a correctly-structured extended model is undesirable due to the inherent complexity of the model form. Additionally, there are instances where simplified or misspecified models can give predictions that are superior, in sense of mean-squared error, to those obtained using the extended model.

For models that are linear in the parameters, it is possible to study the implications of using simplified or extended models. This study was undertaken using both theoretical analysis and Monte Carlo simulations. The simplified model gives better parameter estimates and model predictions (on average) than the extended model if the inequalities (involving the critical ratio R_C) described in (2.26) and (2.32) are satisfied. In these situations, the simplified model is superior, even though the extended model is correctly-structured and the simplified model is misspecified. It was demonstrated that these inequalities are satisfied when there are high noise levels, strong correlations among input variables, small number of experiments, a small range of independent variable settings, or small true values for parameters that are excluded from the simplified model. All of these situations are unfavourable for obtaining precise parameter estimates. When modellers are faced with uninformative and noisy data from poorly designed experiments, they should not try to estimate too many model parameters. Rather, they should confine themselves to fitting only a few key parameters that appear in the most important parts of their models. Unfortunately, there are considerable challenges in deciding which model should be preferred using the limited data that are available for parameter estimation. It is demonstrated how confidence intervals can be constructed for the critical ratio R_C. These intervals are often quite wide, especially when the number of parameters excluded is small. The result is that the statistical tests, while exact in their construction, can have poor discrimination properties for alternative hypotheses (either that the simplified model is better or that the extended model is better) when the data are uninformative. However, when the data are informative, and the terms left out of the simplified model are important, the lower confidence bound for R_C becomes greater than 1, and firm conclusions can be drawn that the extended model is better.

Several methods were investigated for estimating the noise variance and variance-covariance matrix of the parameter estimates obtained from the simplified or misspecified model. Interest in this area has been revived by the availability of inexpensive computing for computationally intensive methods such as the nonparametric bootstrapping.

The focus in this analysis was on models that are linear in the parameters, but the results are very important for phenomenological-based models that are nonlinear in the parameters. In these instances, there are often competing models that can be used. The difference in complexity between a simplified model and an extended model can be substantial. The application of the results in this Chapter can be used, in the first instance, on the linearized representation of the nonlinear model. In these instances, X is replaced by a parametric sensitivity matrix, whose i^{th} column is $\partial f(X,\beta)/\partial \beta_i|_{\beta=\hat{\beta}}$, where $\hat{\beta}$ is either a least-square estimate of β or an initial guess for the parameter values. In nonlinear models, the Fisher information matrix, which corresponds to $X^T X$ for a linear model, is often ill-conditioned (Bates and Watts, 1988; Seber & Wild, 2003; Kou et al., 2005a, b), so that conditions under which the simplified model gives superior predictions are often present.

2.9 Nomenclature

a	vector of coefficients
e	stochastic component
f, g	functions relating explanatory variables, parameters to response variable
k	number of explanatory variables in the true model
m	total number of parameters
n	number of observations
p	number of parameters in first part
p_C	cumulative probability of $\hat{R}_C$ based on central F distribution
q	number of parameters in second part
r	input range
s^2	sample variance

w	total number of predictions
x, z	single observation of explanatory variable
$\tilde{x}_i$	i^{th} row of X_1
y	single observation of response variable
A	auxiliary regression matrix
B	number of bootstraps
I	identity matrix
P	projection matrix
R_C	critical ratio
S	sum of squared residuals
W	vector of length n with entries of $\pm r$
X	matrix of regression variables
Y	response variables
Z	matrix of prediction variables

Greek Symbols

α	significance level
β, θ	unknown parameters
δ	noncentrality parameter
ε	stochastic component
λ	correlation factor
μ	expectation
σ^2	unknown noise variance
Γ	variance-covariance matrix of parameter estimates in the first part
Δ	bias
Σ	variance-covariance
Ψ	covariance matrix between the first part and the second
Ω	variance-covariance matrix of parameter estimates in the second part

Superscripts

$^{-1}$	matrix inverse
^{T}	matrix transpose
$\widehat{\ }$	estimated value
$\bar{\ }$	mean value

Subscripts

$_1$	first partitioned part
$_2$	second partitioned part
$_i$	index
$_{BOOT}$	results from nonparametric bootstrapping
$_{CONV}$	results from conventional methods
$_E$	extended model

L	lower confidence limit
S	simplified model
$SANW$	results from sandwich estimator
U	upper confidence limit
$*$	results from nonparametric bootstrapping

Abbreviations

min	minimization
Cov	covariance matrix
E	mathematical expectation
EM	correctly structured extended model
MSE	mean-squared error
MSEM	mean-squared error matrix
OLS	ordinary least-squares
RHS	right-hand side
SM	simplified/misspecified model
Tr	trace
Var	variance

2.10 References

Abdullaev, F. M. and E. K. Geidarov, "A Recursive 2-step Method of Least-Squares," Automat. Rem. Contr+ 46(1), 66-72 (1985).

Aerts, M. and G. Claeskens, "Bootstrap tests for misspecified models, with application to clustered binary data," Comput. Stat. and Data An. 36(3), 383-401 (2001).

Bagajewicz, M. J. and E. Cabrera, "Data Reconciliation in Gas Pipeline Systems," Ind. Eng. Chem. Res. 42(22), 5596-5606 (2003).

Bates, D. M. and D. G. Watts, "Nonlinear Regression Analysis and Its Applications," John Wiley & Sons, NY, p. 80 (1988).

Beck, J. V. and K. J. Arnold, "Parameter Estimation in Engineering and Science," John Wiley & Sons, NY, pp. 134, 218, 232-234 (1977).

Bera, A. K., "Hypothesis Testing in the 20th Century with a Special Reference to Testing with Misspecified Models," in "Statistics for the 21st Century: Methodologies for Applications of the Future," C. R. Rao G. J. Szekely, Marcel Dekkar, NY, pp. 33-92 (2000).

Box G. E. P. And N. R. Draper, “A Basis for the Selection of a Response Surface Design,” J. Am. Stat. Assoc. 54(287), 622-654 (1959).

Brendel, M., D. Bonvin and W. Marquardt, “Incremental identification of kinetic models for homogeneous reaction systems”, Chem. Eng. Sci. 61, 5404-5420 (2006).

Brooks, R. J. and A. M. Tobias, “Choosing the Best Model: Level of Detail, Complexity, and Model Performance,” Math. .Comput. Model. 24(4), 1-14 (1996).

Chang, S., T. D. Waite and A. G. Fane, “A Simplified Model for Trace Organics Removal by Continuous Flow PAC Adsorption/Submerged Membrane Processes,” J. Membrane Sci. 253(1-2), 81-87 (2005).

Chernick, M. R., “Bootstrap Methods: A Practitioner’s Guide,” John Wiley & Sons, NY, pp. 149-160 (1999).

Davison, A. C. and D. V. Hinkley, “Bootstrap Methods and Their Applications,” Cambridge University Press, US, pp. 256-384 (1997).

Donaldson, J. R. and R. B. Schnabel, “Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares,” Technometrics 29(1), 67-82 (1987).

Draper, N. R. and H. Smith, “Applied Regression Analysis,” 3rd Edn, John Wiley & Sons, NY, pp. 15-250 (1998).

Efron, B. and R. J. Tibshirani, “An Introduction to the Bootstrap,” Chapman and Hall, London, pp. 153-201 (1994).

Freund, R. J., C. W. Cluniesross and R. W. Vail, “Residual Analysis,” J. Am. Stat. Assoc. 56(293), 98-104 (1961).

Fushiki, T., “Bootstrap prediction and Bayesian prediction under misspecified models,” Bernoulli 11(4), 747-758 (2005).

Golbert, J. and D. R. Lewin, “Model-Based Control of Fuel Cells: (1) Regulatory Control,” J. Power Sources 135(1-2), 135-151 (2004).

Goldberg, A., “Stepwise Least-Squares – Residual Analysis and Specification Error,” J. Am. Stat. Assoc. 56(296), 998-1000 (1961).

Goldberg, A. and D. B. Jochems, “Note on Stepwise Least-Squares,” J. Am. Stat. Assoc. 56(293), 105-110 (1961).

Golden, R. M., “Making Correct Statistical Inferences Using a Wrong Probability Model,” J. Math. Psychol. 39, 3-20 (1995).

Good, P. I., "Resampling Methods: A Practical Guide to Data Analysis," 3rd Edn, Birkhäuser, US, pp. 17-27 (2006).

Gunst, R. F. and R. L. Mason, "Biased Estimation in Regression – Evaluation using Mean Squared Error," J. Am. Stat. Assoc. 72(359), 616-628 (1977).

Hocking, R. R., "Analysis and Selection of Variables in Linear Regression," Biometrics 32(1), 1-49 (1976).

Innis, G. and E. Rexstad, "Simulation Model Simplification Techniques," Simulation 41(1), 7-15 (1983).

Kabe, D. G., "Stepwise Multivariate Linear-Regression," J. Am. Stat. Assoc. 58(303), 770-773 (1963).

Kou, B., K. B. McAuley, C. C. Hsu and D. W. Bacon, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene/Hexene Copolymerization with Metallocene Catalyst. Macromol," Macromol. Mater. Eng. 290(6), 537-557 (2005a).

Kou, B., K. B. McAuley, C. C. Hsu, D. W. Bacon and K. Z. Yao, "Mathematical Model and Parameter Estimation for gas-phase ethylene homopolymerization with supported metallocene catalyst," Ind. Eng. Chem. Res. 44(8), 2428-2442 (2005b).

Lowerre, J. M., "Mean – Square Error of Parameter Estimates for Some Biased Estimators," Technometrics 16(3), 461-464 (1974).

Lv, P., J. Chang, T. Wang, C. Wu and N. Tsubaki, "A Kinetic Study on Biomass Fast Catalytic Pyrolysis," Energ. Fuel. 18(6), 1865-1869 (2004).

Maria, G., "A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems", Chem. Biochem. Eng. Q. 18(3), 195-222 (2004).

Martinez W. L. and A. R. Martinez, "Computational Statistics Handbook with MATLAB," Champman & Hall/CRC, US, pp. 214-227 (2002).

Mchaweh, A., A. Alsaygh, K. Nasrifar and M. Moshfeghian, "A Simplified Method for Calculating Saturated Liquid Densities," Fluid Phase Equilib. 224(2), 157-167 (2004).

Miller A. J., "Subset Selection in Regression," Chapman and Hall, London, 169-209 (1990).

Montgomery, D. C., E. A. Peck and G. G. Vining, "Introduction to Linear Regression Analysis," 3rd Edn, John Wiley & Sons, NY, pp. 87-108, 510-511, 582-588 (2001).

Montgomery, D. C. and G. C. Runger, "Applied Statistics and Probability for Engineers," 3rd Edn, John Wiley & Sons, NY, pp. 372-467 (2003).

O'Brien, S. M., L. L. Kupper and D. B. Dunson, "Performance of tests of association in misspecified generalized linear models," J. Stat. Plan. and Infer. 136, 3090-3100 (2006).

Perregaard, J., "Model Simplification and Reduction for Simulation and Optimization of Chemical Processes," Comput. Chem. Eng. 17(5-6), 465-483 (1993).

Price, J. M., "Comparisons among Regression – Estimators under the Generalized Mean – Square Error Criterion", Commun. Stat.-Theor. M. 11(17), 1965-1984 (1982).

Rao, P., "Some Notes on Misspecification in Multiple Regressions," Am. Stat. 25(5), 37-39 (1971).

Rao C. R. and Y. Wu, "On Model Selection," in "Model Selection", P. Lahiri, Institute of Mathematical Statistics, Beachwood, OH, pp. 1-64 (2001).

Rexstad, E. and G. S. Innis, "Model Simplification – 3 Applications," Ecol. Model. 27(1-2), 1-13 (1985).

Romdhane, M. and C. Tizaoui, "The Kinetic Modelling of a Steam Distillation Unit for the Extraction of Aniseed (Pimpinelia Anisum) Essential Oil," J. Chem. Technol. Biot. 80(7), 759-766 (2005).

Rosenberg, S. H. and P. S. Levy, "Characterization on Misspecification in General Linear Regression Model," Biometrics 28(4), 1129-1133 (1972).

Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," John Wiley & Sons, NJ, pp. 23-25, 68-89, 103-126, 572-574 (2003).

Steiger, J. H., "Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis," Psychol. Methods 9 (2), 164-182 (2004).

Sun, C. and J. Hahn, "Parameter Reduction for Stable Dynamical Systems based on Hankel Singular Values and Sensitivity Analysis," Chem. Eng. Sci. 61, 5393-5403 (2006)

Toro-Vizcarrondo, C. and T. D. Wallace, "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," J. Am. Stat. Assoc. 63(322), 558-572 (1968).

Toutenburg, H. and G. Trenkler, "Mean – Square Error Matrix Comparisons of Optimal and Classical Predictors and Estimators in Linear – Regression," Comput. Stat. Data An. 10(3), 297-305 (1990).

Velilla S., "On the Bootstrap in Misspecified Regression Models," Comput. Stat. Data An. 36(2), 227-242 (2001).

Waldorp, L. J., R. P. P. P. Grasman and H. M. Huizenga, "Goodness-of-fit and Confidence Intervals of Approximate Models," J. Math. Psychol. 50, 203-213 (2006).

Waldorp, L. J., H. M. Huizenga and R. P. P. P. Grasman, "The Wald Test and Cramér-Rao Bound for Misspecified Models in Electromagnetic Source Analysis," IEEE T. Signal Proces, 53(9), 3427-3435 (2005).

Wallace, T. D., "Efficiencies for Stepwise Regressions," J. Am. Stat. Assoc. 59(308), 1179-1182 (1964).

Wang, S. G. and S. C. Chow, "Advanced Linear Models: Theory and Applications," Marcel Dekker, NY, p. 66 (1994).

White, H., "Consequences and Detection of Misspecified Nonlinear Regression Models," J. Am. Stat. Assoc. 76(374), 419-433 (1981).

White, H., "Maximum Likelihood Estimation of Misspecified Models," Econometrica 50(1), 1-26 (1982).

Yoshida, H., Y. Takahashi, Y. and M. Terashima, "A Simplified Reaction Model for Production of Oil, Amino Acids, and Organic Acids from Fish Meat by Hydrolysis under Sub-Critical and Supercritical Conditions," J. Chem. Eng. JPN. 36(4), 441-448 (2003).

Zhang, F., "Matrix Theory: Basic Results and Techniques," Springer-Verlag, NY, pp. 159-207 (1999).

Chapter 3

Selection of Simplified Models: I. Analysis of Model-Selection Criteria Using Mean-Squared Error[2]

3.1 Summary

In this Chapter, Mean-Squared Error (MSE) is used to analyze nine commonly-used model-selection criteria (MSC) for their performance of selecting simplified models (SMs). Expressions are derived to enable exact calculations of the probability that a particular MSC will select a SM. For several common MSC, the relative propensities to select SMs are independent of the model structure and the available data for parameter estimation. It is shown that MSC that are effective in preventing overfitting are prone to underfitting when information content of the data is low. In Chapter 4, results are extended to develop a new MSE-based MSC for selecting nonlinear multi-response SMs.

3.2 Introduction

In many modelling situations in science and engineering, modellers have sufficient scientific knowledge to derive complex phenomenological models. However, it is often too difficult or costly to obtain enough good data to reliably estimate all of the unknown model parameters (e.g., Perregaard, 1993; Gray, 1997; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Lv et al.,

[2] The work summarized in this Chapter was submitted to *Canadian Journal of Chemical Engineering* in September 2009. Drs. Kim McAuley and Thomas Harris were co-authors of this research work. Note that this thesis has been prepared using a manuscript format, so some nomenclature used is not consistent throughout the entire thesis. Please refer to Section 3.7 for the nomenclature used in this Chapter.

2004; Maria, 2004, 2006; Mchaweh et al., 2004; Chang et al., 2005; Marquardt, 2005; Romdhane and Tizaoui, 2005; Wang et al., 2007). For complex models with many parameters and limited data, the resulting parameter estimates and model predictions may exhibit high variability. The decisions made using these models (or their parameter estimates) may be unreliable. As a result, good mathematical modelling involves deriving or selecting equations with an appropriate level of details for the intended model use.

Inevitably, modelling leads to the situation where some parameters and terms that would appear in the "true" model are missing, either by mistake or by design. There are many reasons to choose a simplified model (SM) with fewer parameters and terms than the correctly-structured complex model (Neto and Cotta, 1992; Talukdar and Basu, 1995; Gray, 1997; Zhang, 1997; Ismail, 2004; Jaree et al., 2004; Forney et al., 2005; Marquardt, 2005). The practical advantages of a parsimonious model often overshadow concerns over the correctness of the model structure. Sometimes SMs can be expected to give better predictions, in sense of mean-squared prediction error, than the correctly-structured model, even though the SM is misspecified (Rao, 1971; Hocking, 1976; Wu et al., 2007). Predictions and parameters estimates obtained using SMs tend to be better than those from corresponding correctly-structured models in situations where the number of data points available for parameter estimation is small, measurements are noisy, the range of input-variable settings is small, and/or experimental designs are highly correlated.

In these less-than-ideal situations, modellers need to make decisions regarding which terms and parameters to include in their models, and which to leave out, so that they can obtain the best possible predictions using their scientific and engineering knowledge and the experimental data that they possess. Model-selection criteria (MSC) are useful aids for deciding which parameters to include in or exclude from a model. Many different MSC have been proposed, such as the

Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Final Prediction Error (FPE) criterion and Mallow's C_p criterion, as well as variants on these criteria that correct for deficiencies. The statistics literature on this topic is extensive, e.g., Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Lanterman (2001), Rao and Wu (2001), Burnham and Anderson (2002), Stoica and Selen (2004), Konishi and Kitagawa (2008) and the many references contained therein.

In chemical engineering applications, MSC have been extensively studied and used by modellers. For example, Schaper et al. (1994) used the AIC to select an appropriate state-space model of a pilot-scale distillation column. Kendi and Doyle (1996) used AIC, BIC and FPE to select a simplified nonlinear model for control of a fluidized-bed reactor. Li et al. (2002) and Chetouani (2007) used the AIC and the BIC to determine a suitable NARMAX model for a reactor-exchanger. Toher et al. (2007) used AIC and BIC, respectively, for selecting multivariate statistical models for assessing product quality. Choi et al. (2008) applied the AIC to select an autoregressive model used for batch process performance monitoring. Scherr et al. (2008) used the corrected Akaike Information Criterion (AIC_C) and the adjusted coefficient of determination (R^2_{adj}) to select between two candidate kinetic models proposed for estrone-3-sulfate aerobic degradation.

Practical issues involving model selection using limited data are far from settled. Lanterman (2001) states: "Unfortunately, model order estimation remains a subject of tremendous controversy; there is little agreement on what the 'best' approach is, and indeed little agreement on if there is, in fact, such a thing as a 'best' approach." In a similar vein, Shibata (1989) notes that consistency in selecting the true system is only meaningful when the true system is simple

and when one of the candidate models can describe the system without error. He also notes that properly selecting the true model does not always lead to the best parameter estimates.

In this Chapter, the properties of nine commonly-used MSC are examined in situations where small data sets are available to fit or calibrate models. The use of the expected MSE of model predictions provides a convenient means for comparison and analysis of the various MSC. Although the main interest of this research is in multi-response fundamental and semi-empirical models that are nonlinear in the parameters, the analysis in this Chapter focuses on single-response linear statistical models, which will enable the derivation of exact theoretical results. While this choice might at first seem restrictive, note that: 1) the statistical analysis of nonlinear models often involves linearization of the model equations around the nominal parameter values (Seber and Wild, 2003); and 2) multi-response models can be expressed in a "rolled-out" form to behave like univariate models by stacking the various response variables in a single scaled vector (Seber and Wild, 2003). Thus, the results of the analysis in this Chapter will be helpful for selection of appropriate nonlinear multi-response models.

Using the expected mean-squared prediction error for selecting candidate models was outlined by Hocking (1976), and further developed by Linhart and Zucchini (1986) and Wu et al. (2007). In this Chapter, nine commonly-used MSC are summarized and a MSE-based interpretation is demonstrated to provide a convenient and insightful way to examine the performance of these various MSC. Theoretical expressions are derived and used to illustrate that the probability of selecting a particular SM, using a particular MSC, depends on a critical ration R_C (Wu et al., 2007). This critical ratio is defined as the total squared bias in the model predictions, divided by the expected total variance reduction. The bias results from the misspecified structure of the SM, and the variance reduction results from the smaller number of parameters being estimated in the

SM than in the EM. The critical ratio R_C is used to show that, for many of the MSC, their relative propensities for selecting SMs are independent of model structure and the available data set. Where possible, the various criteria are ranked according to this propensity. Finally, a simulation study is presented to confirm the results from the theoretical analysis. This simulation study uses a linear regression example that can be easily adapted to provide data with different levels of information content.

Note that, in the literature, regularization methods (Neumaier, 1998) are also used for model selection by implicitly or explicitly penalizing models based on the number of their parameters. These methods include penalized least squares and penalized maximum likelihood (Fan and Li, 2001) and the LASSO estimator (Tibshirani, 1996, 1997), which can be used to simultaneously estimate parameters and force less-important parameters to zero. Li and Lin (2002) have used entropy penalties to study the connection between penalized least squares and two MSC (i.e., AIC and BIC). Usually the application of these methods requires extensive computations, which might not be practical for large models (e.g. nonlinear dynamic models of chemical processes). Therefore, in this analysis, no studies on these regularization methods are carried out.

3.3 Model-Selection Criteria

Nine commonly-used MSC are summarized in Table 3.1. It is convenient to classify the various MSC into three categories based on similarities in their formulae (for linear statistical models) rather than similarities in the assumptions used in their development. The first category, log-based MSC, includes the AIC (Akaike, 1973) and two corrected versions AIC_C (Sugiura, 1978; Hurvich and Tsai, 1989) and AIC_U (McQuarrie et al., 1997), as well as the BIC (Akaike, 1978; Schwarz, 1978). The second category contains four prediction-based MSC: the FPE

(Akaike, 1973) and its corrected version FPE_U (McQuarrie and Tsai, 1998), the S_p criterion (Breiman and Freedman, 1983; Linhart and Zucchini, 1986), and the adjusted coefficient of determination R^2_{adj}. The third category contains Mallow's C_p (Mallow, 1973), which is structurally different from the other criteria, because it includes the sum of squared residuals from the correctly-structured EM. An assumed correctly-structured model is not required to compute values of the other MSC.

Table 3.1: List of model-selection criteria studied with formula based on models with k parameters. n is the number of observations and p is the number of parameters in the correctly-structured EM. "SSE" denotes the Sum of Squared Residuals.

MSC		Formula	Reference
Log-Based Criteria	AIC	$log\left(\frac{SSE_k}{n}\right)+\frac{2(k+1)}{n}$	Akaike (1973)
	AIC_C	$log\left(\frac{SSE_k}{n}\right)+\frac{n+k}{n-k-2}$	Sugiura (1978) Hurvich and Tsai (1989)
	AIC_U	$log\left(\frac{SSE_k}{n-k}\right)+\frac{n+k}{n-k-2}$	McQuarrie et al. (1997)
	BIC	$log\left(\frac{SSE_k}{n}\right)+\frac{log(n)k}{n}$	Akaike (1978) Schwarz (1978)
Prediction-Based Criteria	FPE	$SSE_k\frac{n+k}{n(n-k)}$	Akaike (1973)
	FPE_U	$SSE_k\frac{n+k}{(n-k)^2}$	McQuarrie and Tsai (1998)
	S_p	$SSE_k\frac{1}{(n-k)(n-k-1)}$	Breiman and Freedman (1983)
	$1-R^2_{adj}$	$SSE_k\frac{1}{(n-k)MST}$ [3]	McQuarrie and Tsai (1998)
C_p		$(n-p)\frac{SSE_k}{SSE_E}-n+2k$	Mallow (1973)

When modellers use any of the MSC in Table 3.1, they compute the criterion value from the available data for each candidate model, and they select the model that corresponds to the best

[3] MST is defined as $MST=\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)$, where $\bar{y}=\sum_{i=1}^{n}y_i/n$. MST is totally determined by the data, and is independent with the model from. In this work, we show results for $1-R^2_{adj}$ rather than R^2_{adj}, so that, as in the case of the other MSC, the modeller will select the MSC with the smallest criterion value.

value of the particular MSC. When using R^2_{adj}, the model with the largest value is selected, whereas for the other eight criteria in Table 3.1, the model with the smallest calculated criterion value is selected.

3.4 Mean-Squared Error Interpretations of Model-Selection Criteria

Assume that the true process can be described by

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{3.1}$$

where there are p unknwon parameters in β, p_1 unknown parameters in β_1, p_2 unknown parameters in β_2, and n data points available for parameter estimation. We refer to this correctly-structured or extended model as the EM. The usual ordinary least-squares (OLS) assumptions are made (Beck and Arnold, 1977): 1) X_1 and X_2 are deterministic and have full column rank; and 2) the stochastic component ε is independently and identically distributed with zero mean and constant variance σ^2.

A simplified (or under-parameterized) model (SM) is of the form

$$Y = X_1\beta_1 + e \tag{3.2}$$

where $e = X_2\beta_2 + \varepsilon$ is the stochastic component combined with any model mismatch. Using OLS, model predictions made at input settings X are

$$\begin{aligned} \hat{Y}_S &= X_1(X_1^T X_1)^{-1} X_1^T Y = P_1 Y \\ \hat{Y}_E &= X(X^T X)^{-1} X^T Y = PY \end{aligned} \tag{3.3}$$

where $P_1 = X_1(X_1^T X_1)^{-1} X_1^T$ and $P = X(X^T X)^{-1} X^T$. The subscripts "S" and "E" indicate the use of the SM and the EM, respectively. The corresponding sums of squared residuals from the SM and the EM are

$$SSE_S = Y^T(I_n - P_1)Y$$
$$SSE_E = Y^T(I_n - P)Y \tag{3.4}$$

Note that, $SSE_S \geq SSE_E$ is always true, and that the equality holds only when the true values of all parameters in β_2 are zero.

For the models described in Eqns. (3.1) and (3.2), the (total) mean-squared prediction errors at the design points are (Beck and Arnold, 1977)

$$MSE_S = \sigma^2 p_1 + \beta_2^T X_2^T (I_n - P_1) X_2 \beta_2$$
$$MSE_E = \sigma^2 p \tag{3.5}$$

The first term on the right-hand side of the expression for MSE_S is $\sigma^2 p_1$, which is the total variance in model predictions made at the design points. The second term, $\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2$, is the corresponding total squared bias. Wu et al. (2007) defined a critical ratio R_C

$$R_C = \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{(p - p_1)\sigma^2} \tag{3.6}$$

which is the total squared bias in the model prediction (introduced by removing parameters from the EM) divided by the variance reduction in the model predictions (due to fewer parameters being estimated in the SM). The inequality

$$R_C < 1 \tag{3.7}$$

provides a necessary and sufficient condition for the model predictions obtained using the SM having a smaller mean-squared error than those from the EM (Wu et al., 2007). A similar result was obtained by Linhart and Zucchini (1986) who considered the average mean-squared prediction error. Their analysis can be extended to give an alternative critical ratio $\tilde{R}_C = (p - p_1)R_C$ and the inequality $\tilde{R}_C < p - p_1$, which is equivalent to the inequality (3.7).

When the unknown parameter values β_2 and the unknown noise variance σ^2 are replaced by estimates obtained from the EM, an estimator for R_C is

$$\hat{R}_C = \frac{(SSE_S - SSE_E)/(p - p_1)}{SSE_E/(n - p)} \tag{3.8}$$

Calculation of $\hat{R}_C$ requires knowledge (or an assumption) about the form of the true model to enable the calculation of SSE_E. However, knowledge of the true model structure is not required for the theoretical comparison of the properties of the various MSC studied in this Chapter.

The right-hand side of Eqn. (3.8) is a likelihood ratio statistic (Wang and Chow, 1994). When the stochastic component ε in the EM is Normally distributed, $\hat{R}_C$ has a noncentral F distribution with $p - p_1$ and $n - p$ degrees of freedom and noncentrality parameter λ, where

$$\lambda = (p - p_1)R_C \tag{3.9}$$

To aid in the comparison of the MSC in Table 3.1, the difference in the numerical values of a particular criterion (i.e., $\Delta MSC = MSC_S - MSC_E$) is considered. If $\Delta MSC < 0$, the corresponding SM will be preferred to the EM by the particular criterion. If several candidate SMs are being compared, the modeller will select the SM corresponding to the smallest value of MSC_S, which is equivalent to selecting the model with the smallest ΔMSC.

For log-based MSC (i.e., the AIC, AIC_C, AIC_U and BIC), the expressions for ΔMSC are of the form

$$\Delta MSC = log\left(\frac{SSE_S}{SSE_E}\right) - a(n, p, p_1) \tag{3.10}$$

where a is a function of n, p and p_1 only. Expressions for a are summarized in Table 3.2. For prediction-based MSC (i.e., FPE, FPE_U, S_p and $1 - R^2_{adj}$ criteria), the expressions for ΔMSC are of the form

$$\Delta MSC = bSSE_S - cSSE_E \tag{3.11}$$

Expressions for b and c are also summarized in Table 3.2. For Mallow's C_p, the MSC difference is

$$\Delta MSC = (n-p)\frac{SSE_S}{SSE_E} - (n+p-2p_1) \tag{3.12}$$

The probability that a SM will be preferred to the EM, using a particular MSC, is the probability that $\Delta MSC \leq 0$, which is related to $\hat{R}_C$ by

$$Pr(\Delta MSC \leq 0) = Pr\left(\hat{R}_C \leq f(n,p,p_1)\right) \tag{3.13}$$

where $f(n,p,p_1)$ is a critical value determined by the number of data points and the number of parameters in the EM and SM. Eqn. (3.13) was derived using

$$\frac{SSE_S}{SSE_E} = \frac{p-p_1}{n-p}\hat{R}_C + 1 \tag{3.14}$$

from Eqn. (3.8). Expressions for f have been derived and are shown in the final column of Table 3.2.

Table 3.2: Constant terms a (Eqn. (3.10)), b (Eqn. (3.11)) and c (Eqn. (3.11)) values and critical values for various MSC when there are p_1 parameters in the SM, p parameters in the correctly-structured EM and n data points available for parameter estimation.

MSC		**a Value**		**Critical Value f**
Log-Based Criteria	AIC	$\frac{2(p-p_1)}{n}$		$\frac{n-p}{p-p_1}(e^a-1)$
	AIC_C	$\frac{n+p}{n-p-2} - \frac{n+p_1}{n-p_1-2} \equiv d$		
	AIC_U	$log\left(\frac{n-p_1}{n-p}\right) + d$		
	BIC	$\frac{log(n)}{n}(p-p_1)$		
		b Value	**c Value**	
Prediction-Based Criteria	FPE	$\frac{n+p_1}{n(n-p_1)}$	$\frac{n+p}{n(n-p)}$	$\frac{n-p}{p-p_1}\left(\frac{c}{b}-1\right)$
	FPE_U	$\frac{n+p_1}{(n-p_1)^2}$	$\frac{n+p}{(n-p)^2}$	
	S_p	$\frac{1}{(n-p_1)(n-p_1-1)}$	$\frac{1}{(n-p)(n-p-1)}$	
	$1-R^2_{adj}$	$\frac{1}{(n-p_1)MST}$	$\frac{1}{(n-p)MST}$	
C_p		———		2

Based on results summarized in Table 3.2, the following ordering of the critical values for various MSC can be established

$$\begin{gathered} f_{AIC} < f_{C_p} < f_{AIC_C} < f_{AIC_U} \\ f_{AIC} < f_{BIC} \ \ (n > 7) \\ f_{FPE} < f_{FPE_U} \\ f_{R^2_{adj}} < f_{FPE} < f_{C_p} < f_{S_p} \end{gathered} \tag{3.15}$$

This ordering of the critical values corresponds to the ordering of the probabilities of the various MSC for selecting a SM rather than the corresponding EM. For example, $f_{AIC} < f_{C_p}$ indicates that the Mallow's C_p criterion has a greater tendency to select SMs than the AIC does.

The differences among the critical f values for various MSC can be dramatic, as seen in Figure 3.1, which shows results for an example with $n = 16$ data points and $p = 5$ parameters in the EM.

Figure 3.1 indicates that, for this example, AIC_U has the largest critical values, and hence the greatest tendency to prefer a SM, and that R^2_{adj} has the lowest critical values for all cases considered in Figure 3.1, and hence the least tendency to prefer a SM rather than the EM. From the inequalities in Eqn. (3.15), it is apparent that many of the rankings among the MSC always hold, and are neither model- nor data-dependent.

Figure 3.2 shows a comparison of the probabilities (computed from the noncentral *F* distribution) that various MSC will choose a particular SM, rather than the EM, for different values of R_C when $n = 16$ and $p = 5$. Figure 3.2 shows that all of the MSC are more likely to select the SM as R_C becomes smaller. The reduction in R_C is associated with relative improvement in predictions of the SM, compared to the EM. The vertical line at $log(R_C) = 0$

corresponds to $R_C = 1$. When $R_C < 1$, the SM will be expected to provide lower MSE predictions than the EM (at the design points used for parameter estimation).

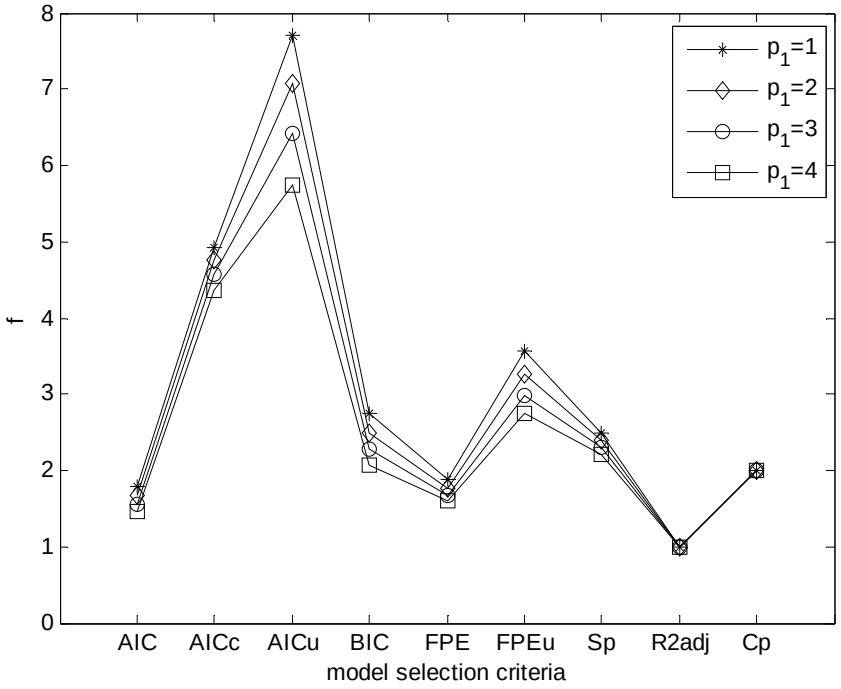

Figure 3.1: Critical values relating $\widehat{R}_C$ to the various MSC

Figure 3.2 also confirms the ordering of tendencies established by the inequalities in Eqn. (3.15). For example, the AIC_U criterion has a larger tendency to prefer simplified models than does AIC_C, C_p or AIC.

For the particular example studied, the AIC_U criterion has a large propensity for selecting over-simplified models. Figure 3.2 shows that, when $p_1 = 2$ and $R_C = 5$, so that $log(R_C) \approx 1.6$, the probability of the AIC_U criterion selecting the SM is approximately 60%, even though the EM can be expected to give superior predictions (because $R_C > 1$). The R^2_{adj} criterion has the lowest

probability of preferring a particular SM to the EM, compared to the other MSC. For example, consider the situation where $p_2 = 2$ and $R_C = 0.5$, so that $log(R_C) \approx -0.7$. Since $R_C < 1$, the SM can be expected to give lower MSE predictions than the EM. However, using R^2_{adj}, the modeller will select the EM more than 60% of the time.

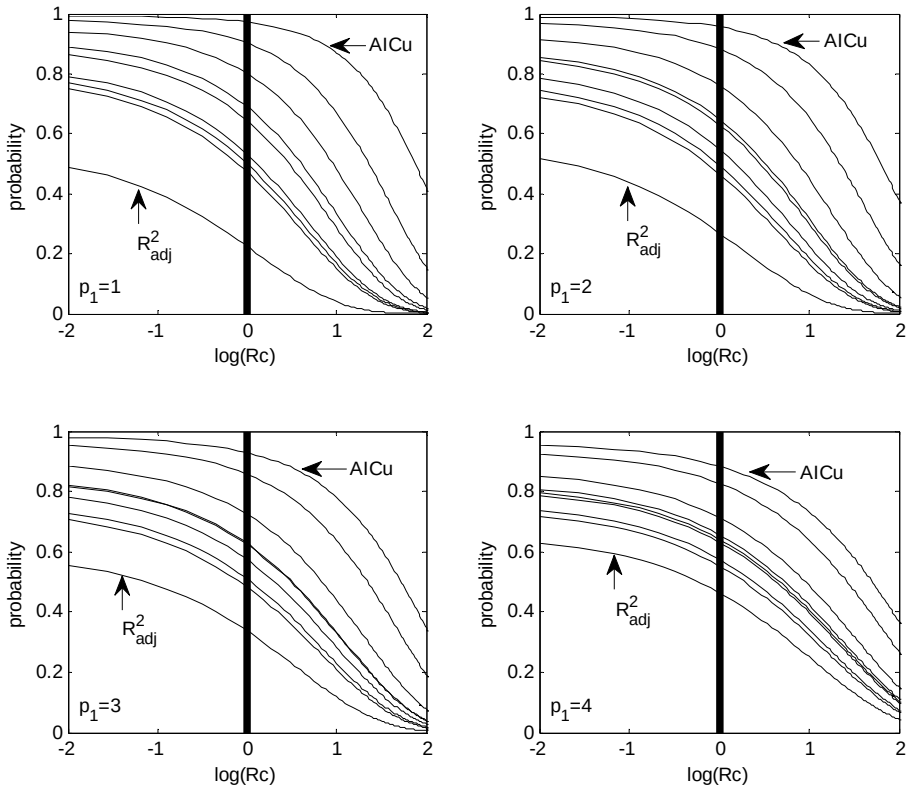

Figure 3.2: Probability of the SM being preferred to the EM using various MSC when 16 data points are used to fit the model and the EM has 5 parameters. From top to bottom, the curves on subplots $p_1 = 1$ and $p_1 = 2$ correspond to AIC_U, AIC_C, FPE_U, BIC, S_p, C_p, FPE, AIC and R^2_{adj}, and the curves on subplots $p_1 = 3$ and $p_1 = 4$ correspond to AIC_U, AIC_C, FPE_U, S_p, BIC, C_p, FPE, AIC and R^2_{adj}. Note that the curve for BIC is above the curve for S_p in subplots $p_1 = 1$ and $p_1 = 2$, whereas the curves for S_p is above the curve for BIC when $p_1 = 3$ and $p_1 = 4$. The order for all other curves is the same in all four subplots.

Considerable research has focused on the ability of MSC to guard against overfitting and to select the true model. In their comprehensive analysis, McQuarrie and Tsai (1998) used a signal-to-noise-ratio approach to analyze and compare the tendencies of various MSC to select over-parameterized models. The probabilities of overfitting were derived both asymptotically and for small-sample cases. They showed that the corrected versions of the Akaike Information Criterion (AIC_C and AIC_U) and the Final prediction Error criterion (FPE_U) have stronger penalty functions than the original versions, which result in smaller probabilities of selecting over-parameterized models. McQuarrie and Tsai (1998) acknowledged that "criteria with penalty functions that excessively resist overfitting may be prone to underfitting, particularly when the true model is weak", which corresponds to limited data for fitting the true model. This is indeed what is proven in this study using the critical ratio R_C when the values of R_C is small.

The analysis in this section has focused on the relative propensities for the various MSC to select SMs, rather the correctly-structured EM. However, this analysis does not provide information about whether a particular MSC is proficient in selecting the best model from a set of candidate SMs and the EM. If $R_C < 1$, an SM is expected to give lower total mean-squared error than the EM for predictions made at the design points. When comparing several SMs, however, it is not always true that the SM with the lowest value of R_C has the smallest expected mean-squared prediction error. When a particular SM is used, it can be verified using Eqn. (3.5) that

$$\Delta MSE = MSE_S - MSE_E = \sigma^2(p - p_1)(R_C - 1) \tag{3.16}$$

When two SMs contains different numbers of parameters, p_1, the SM with the lowest value of R_C may not correspond to the lower value of ΔMSE, due to the factor $p - p_1$ on the right-hand side of Eqn. (3.16). In the next Chapter, a new MSC is proposed based on Eqn. (3.16) to assist the

modeller to select the SM, with lowest mean-squared prediction error, from a group of candidate models.

3.5 Monte Carlo Simulations

Consider a linear example with five parameters in the correctly-structured EM. The true parameter values are

$$\beta = (\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4 \quad \beta_5)^T = (1 \quad 1/2 \quad 1/3 \quad 1/4 \quad 1/5)^T \tag{3.17}$$

Input-variable settings for the experiments are

$$X = \left(W_1 \quad W_2 \quad W_3 \quad rW_1 + \sqrt{1-r^2}W_4 \quad rW_2 + \sqrt{1-r^2}W_5\right) \tag{3.18}$$

where W_i $(i = 1,2,3,4,5)$ are orthogonal column vectors defined as

$$(W_1 \quad W_2 \quad W_3 \quad W_4 \quad W_5) = \begin{pmatrix} 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 \end{pmatrix} \tag{3.19}$$

The design factor r in Eqn. (3.18) affects the correlation between the columns of X. When $r = 0$, all columns of X are orthogonal. As $r \to 1$, columns 4 and 5 become highly correlated with columns 1 and 2, respectively. The additive noise ε in the EM is independently and identically distributed, following a Normal distribution with zero mean and constant variance.

To illustrate the theoretical analysis in the previous section, and to compare the performance of the various MSC for selecting simplified models (or over-parameterized models), nine candidate models summarized in Table 3.3 are considered. Models M_1 to M_7 are SMs, and model M_8 is the EM used to generate the simulated data. Model M_9 is an over-parameterized model with a nuisance parameter $\beta_6 = 0$. The input-variable settings corresponding to β_6 are orthogonal to all of the W_i $(i = 1,2,3,4,5)$

$$W_6^T = (1 \quad -1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 1 \quad -1 \quad -1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 1 \quad -1 \quad 1) \qquad (3.20)$$

Note that, although $W_6 = -W_2 W_5$, the input settings corresponds to β_5 is $rW_2 + \sqrt{1-r^2}W_5$, and no interaction term is considered in this analysis.

Table 3.3: Candidate models used for evaluating model-selection criteria. Model M_8 is the true model used in Monte Carlo simulations. Models M_1 to M_7 are simplified models. Column with a "√" indicates that the corresponding term is included in the particular model. Model M_9 is an over-parameterized model with an extra parameter β_6.

Model Index	Parameters Included in Candidate Models					
	β_1	β_2	β_3	β_4	β_5	β_6
M_1	√					
M_2				√		
M_3				√	√	
M_4	√	√	√			
M_5	√		√		√	
M_6		√	√	√		
M_7	√	√	√	√		
M_8	√	√	√	√	√	
M_9	√	√	√	√	√	√

Table 3.4 summarizes the true values of R_C and the total mean-squared prediction error (at the design points) obtained using each candidate model in different situations. In the following analysis, four situations are considered corresponding to different measurement noise variances ($\sigma^2 = 0.15$ is the low noise-level setting and $\sigma^2 = 10$ is the high noise-level setting) and

different levels of correlation between columns in X ($r = 0.20$ is the weak correlation setting, and $r = 0.99$ is the strong correlation setting).

Table 3.4: Theoretical values of R_C and mean-squared prediction error obtained using the candidate models in Table 3.3 for different values of σ^2 and correlation factor r.

		M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
Number of Parameters		1	1	2	3	3	3	4	5	6
$\sigma^2 = 0.15$ $r = 0.20$	R_C	13.36	37.36	46.62	5.25	16	53.25	4.10	——	
	MSE	8.17	22.57	21.28	2.02	5.25	16.42	1.21	**0.75**	0.90
$\sigma^2 = 10$ $r = 0.20$	R_C	0.20	0.56	0.70	0.079	0.24	0.80	0.06	——	
	MSE	**18.02**	32.42	40.98	31.57	34.80	45.97	40.61	50	60
$\sigma^2 = 0.15$ $r = 0.99$	R_C	16.01	16.50	4.84	0.11	0.33	1.10	0.085	——	
	MSE	9.76	10.05	2.48	**0.48**	0.55	0.78	0.61	0.75	0.90
$\sigma^2 = 10$ $r = 0.99$	R_C	0.24	0.25	0.073	0.002	0.005	0.017	0.001	——	
	MSE	**19.61**	19.90	22.18	30.03	30.10	30.33	40.01	50	60

The total mean-squared prediction errors for the SMs (M_1 to M_7) and EM (M_8) are calculated using Eqn. (3.5). For the over-parameterized model M_9, there is no bias in the model predictions, and

$$MSE\left(\hat{Y}_{M_9}\right) = \sigma^2(p+1) \tag{3.21}$$

where $p = 5$ is the number of parameters in the correctly-structured model M_8. Note that inequality $MSE\left(\hat{Y}_{M_9}\right) > MSE\left(\hat{Y}_{M_8}\right)$ is always true, which means that, in sense of expected MSE, M_9 is always worse than M_8.

Monte Carlo simulations were run 10000 times in each set of simulations, each time with a different random noise sequence. Mallow's C_p requires an assumed full model to calculate a sum of squared residuals, which was computed using M_9 (the over-parameterized model), because a modeller who would consider M_9 as a candidate model would not know that $\beta_6 = 0$.

Table 3.5: Percentage of each candidate model being selected in each situation by using the various MSC. Results are obtained from 10000 simulations.

		M_{1-7}	M_8	M_9	$M_{R_C<1}$	Ranking of Propensity to Select a SM
$\sigma^2 = 0.15$ $r = 0.20$	AIC	17.92	**57.39**	24.69	-	8
	AIC_C	55.04	**41.64**	3.32	-	2
	AIC_U	**69.17**	29.34	1.49	-	1
	BIC	26.6	**56.84**	16.57	-	5
	FPE	19.83	**58.36**	21.81	-	7
	FPE_U	36.14	**53.40**	10.45	-	3
	S_p	28.27	**57.84**	13.89	-	4
	R^2_{adj}	11.62	**56.34**	32.04	-	9
	C_p	23.54	**58.36**	18.10	-	6
$\sigma^2 = 10$ $r = 0.20$	AIC	87.98	3.88	8.14	-	8
	AIC_C	99.44	0.33	0.23	-	2
	AIC_U	**99.87**	0.08	0.05	-	1
	BIC	94.95	1.72	3.33	-	5
	FPE	89.71	3.47	6.82	-	7
	FPE_U	97.61	1.00	1.39	-	3
	S_p	94.99	1.85	3.61	-	4
	R^2_{adj}	74.41	8.27	17.32	-	9
	C_p	90.77	2.87	6.36	-	6
$\sigma^2 = 0.15$ $r = 0.99$	AIC	83.36	5.92	10.72	63.39	8
	AIC_C	98.82	0.72	0.46	73.97	2
	AIC_U	**99.55**	0.27	0.18	73.33	1
	BIC	91.66	3.08	5.26	69.18	5
	FPE	85.58	5.28	9.14	66.95	7
	FPE_U	95.69	1.92	2.39	71.98	3
	S_p	92.45	3.16	4.39	69.90	4
	R^2_{adj}	71.27	9.88	18.85	55.33	9
	C_p	87.89	4.42	7.69	66.59	6
$\sigma^2 = 10$ $r = 0.99$	AIC	83.67	6.97	9.36	-	8
	AIC_C	99.01	0.69	0.30	-	2
	AIC_U	**99.77**	0.17	0.06	-	1
	BIC	93.13	3.05	3.82	-	4
	FPE	85.85	6.34	7.81	-	7
	FPE_U	96.44	1.85	1.71	-	3
	S_p	92.82	3.58	3.60	-	5
	R^2_{adj}	67.52	13.42	**19.06**	-	9
	C_p	86.94	5.68	7.38	-	6

The column in Table 3.5 under the header "M_{1-7}" displays the percentage of the time that each criterion resulted in the selection any simplified model (probabilities for M_1 to M_7 added together). The next column shows the percentage of simulated experiments where the true model was selected, and the column after shows the percentage of simulations where the over-parameterized model was selected. The column under the header "$M_{R_C<1}$" shows the percentage of the time that each criterion selected any one of the SMs with $R_C < 1$. Note that no values appear in this column in the situations where $R_C > 1$ for all SMs (the first case) and in situations where $R_C < 1$ for all SMs (the second and the fourth cases).

Low Noise Variance and Low Correlation ($\sigma^2 = 0.15,\ r = 0.20$)

In this situation, the available data are not very noisy and the input settings in X are weakly correlated. The critical ratio R_C is greater than 1 for all of the SMs (M_1 to M_7), so that the true model M_8 is preferable for making predictions. Table 3.5 shows that the true model is selected most often by all of the criteria, except for AIC_U. Note that AIC_U selects both M_4 and M_7 more often than M_8 (data not shown). From the MSE entries in Table 3.4, it is apparent that M_4 and M_7 are the two good SMs for making predictions at the design points, but that these models are expected to give worse predictions than M_8. Among all of the criteria, AIC_U selects the over-parameterized model M_9 least often, and R^2_{adj} has the greatest tendency to select M_9. BIC and FPE_U do a good job of selecting the best model M_8, while avoiding both over- and under-fitted models.

High Noise Variance and Low Correlation ($\sigma^2 = 10,\ r = 0.20$)

The correlation factor r is the same as in the first situation, but the noise variance has increased considerably, making it more difficult to obtain reliable estimates of the parameters.

The true values of R_C decrease to values less than 1 for all candidate SMs (see Table 3.4), and the mean-squared prediction errors increase due to the lower-quality data. Table 3.4 shows that model M_1, which contains only the intercept, is the best model for making predictions, due to the poor data quality. Note that the relative rankings for the various MSC criteria in the final column of Table 3.5 are the same as for the first set of simulations.

Low Noise Variance and High Correlation ($\sigma^2 = 0.15,\ r = 0.99$)

In this situation, there is a small noise variance, but the input settings are highly correlated, and a SM is expected to provide the best predictions. Table 3.4 shows that model M_4 (containing parameters β_1, β_2 and β_3) is the best SM with the lowest value of the expected mean-squared prediction error. This is not surprising, since, even though the noise variance is small, the experimental design is highly correlated, and the true values of β_4 and β_5 are small. In this situation, predictions from simplified model M_4 are better, on average, than predictions from the true model M_8. The critical ratio R_C exceeds 1 for models M_1, M_2, M_3 and M_6, and indicates that the EM, M_8, is preferred to these SMs. Inequality $R_C < 1$ is satisfied for models M_4, M_5 and M_7, indicating that these SMs are preferred to M_8. Note that R_C and MSE for the various SMs are quadratic functions of r for the particular example studied, so that R_C and MSE do not decrease monotonically as the correlation factor r increases from 0.20 to 0.99. The fourth column in Table 3.5 shows the percentage of time that models M_4, M_5 and M_7 (which are better than M_8) are selected. The ranking results in Table 3.5 for the various MSC support the theoretical analysis.

High Noise Variance and High Correlation ($\sigma^2 = 10,\ r = 0.99$)

In this situation, the data are noisy and the input settings are highly correlated, so a model with the simplest structure is preferred. Table 3.4 shows that model M_1 is the best SM with lowest

value of MSE for model predictions. The pattern of results for this case is similar to that discovered in previous cases.

In all of the simulated cases, the relative rankings in the final column of Table 3.5 are consistent with the theoretical rankings in Eqn. (2.15). Note that criterion AIC_U is ranked first for all four sets of simulations, with the largest tendency to select SMs. R^2_{adj} has the lowest tendency, in all four sets of simulations, to select SMs. Also note that the rankings of the criteria in the four different situations are quite similar and that these simulation results are consistent with the ranking of the criteria in Figure 3.2.

3.6 Conclusions

The performance of nine commonly-used model-selection criteria was examined for selection of simplified linear regression models. It was shown that a critical ratio R_C and its estimator $\hat{R}_C$ provide a convenient theoretical connection among the various model-selection criteria and their sampling properties. This connection enables the ranking of the relative probabilities of the criteria for preferring simplified models rather than the correctly-structured model. It was established that there exist preferential orderings for many of the model-selection criteria that are independent of the model structure and the particular data set. For example, the following order of probabilities for selecting simplified models was proven: S_p, C_p, FPE, R^2_{adj}, wherein the S_p criterion has the largest probability of the four for selecting simplified models rather than the properly-structure model, and the other three criteria have progressively lower probabilities. It is also observed that model-selection criteria with strong tendencies to guard against overfitting have a high tendency to select simplified models when underfitting is considered.

Results from the theoretical analysis was confirmed in Monte Carlo simulations based on a carefully designed linear regression example, which can be easily adapted to generate data with different levels of information content. For the particular example examined, the Bayesian Information Criterion (BIC) and a corrected version of the Final Prediction Error criterion (FPE_U) did a good job of selecting the best model, with the lowest total mean-squared prediction error at the design points.

In Chapter 4, a new model-selection criterion is proposed to help modellers in selecting models with lowest mean-squared prediction error. A linear statistical model is used in the development of this new criterion, and applicability to the selection of nonlinear multivariate models of chemical processes is demonstrated.

3.7 Nomenclature

a, b, c	coefficients
e	stochastic component with any model mismatch
f	critical value
k	number of parameters in the candidate model
log	natural logarithm
n	number of data points for a single response variable
p	total number of unknown parameters
r	design factor
I_n	$(n \times n)$ identity matrix
M	candidate model
P	projection matrix
P_r	cumulative probability
R_C	critical ratio
R^2_{adj}	coefficients of determination
W	cector of length n with entries of ± 1
X	matrix of regression variables
Y	response variable

Greek Symbols

β	unknown parameters

ϵ	stochastic component
λ	noncentrality parameter
σ^2	noise variance
Δ	difference

Superscripts

$\widehat{\ }$	estimated value
$\sim$	corrected or alternative
T	matrix transcript
-1	matrix inverse

Subscripts

1	first partitioned part
2	second partitioned part
i	index
E	extended model
S	simplified model

Abbreviations

AIC	Akaike Information Criterion
AIC_C, AIC_U	corrected Akaike Information Criterion
BIC	Bayesian Information Criterion
EM	Extended Model
FPE	Final Prediction Error
FPE_U	corrected Final Prediction Error
MSC	Model-Selection Criteria
MSE	Mean-Squared-Error
SM	Simplified Model
SSE	Sum of Squared Residuals

3.8 Acknowledgements

The authors would like to thank Cybernetica, DuPont, Hatch, Matrikon, SAS, MITACS (Mathematics of Information Technology and Complex Systems) and NSERC (TJH) for financial support of this research.

3.9 References

Akaike, H. “Information Theory and an Extension of the Maximum Likelihood Principle,” 2. Int. S. Inf. Theor., Ed. Petrov, B. N. and Csaki F., pp. 267-281. Budapest: Akademia Kiado (1973).

Akaike, H. “A Bayesian Analysis of the Minimum AIC Procedure,” Ann. I. Stat. Math. 30, Part A, 9-14 (1978).

Bagajewicz, M. J. and E. Cabrera, “Data Reconciliation in Gas Pipeline Systems,” Ind. Eng. Chem. Res. 42(22), 5596-5606 (2003).

Beck, J. V. and K. J. Arnold, “Parameter Estimation in Engineering and Science,” John Wiley and Sons, NY, pp. 134-154 (1977).

Breiman, L. and D. Freedman, “How Many Variables Should Be Entered in a Regression Equation?” J. Am. Stat. Assoc. 78 (381), 131-136 (1983).

Burnham, K. P. and D. R. Anderson, “Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,” 2nd Edn. Springer, NY, pp. 49-148 (2002).

Chang, S., T. D. Waite and A. G. Fane, “A Simplified Model for Trace Organics Removal by Continuous Flow PAC Adsorption/Submerged Membrane Processes,” J. Membrane Sci. 253(1-2), 81-87 (2005).

Chetouani, Y., “Modeling and Prediction of the Dynamic Behavior in a Reactor-Exchanger Uisng NARMAX Neural Structure,” Chem. Eng. Commun. 194(5), 691-705 (2007).

Choi S. W., J. Morris and I. B. Lee, “Dynamic Model-based Batch Process Monitoring,” Chem. Eng. Sci. 63(3), 622-636 (2008).

Fan, J. and R. Li, “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties,” J. Am. Stat. Assoc. 96, 1348-1360 (2001).

Forney, L. J., J. M. Brown, B. M. Kadlubowski and J. T. Sommerfeld, “Simplified Model for Oxygen Transport with Reaction in a Polymer-Electrolyte Fuel Cell,” Can. J. Chem. Eng. 83(3), 500-507 (2005).

Gray, M. R. “Through a Glass, Darkly: Kinetics and Reactors for Complex Mixtures,” Can. J. Chem. Eng. 75(3), 481-493 (1997).

Hocking, R. R. “Analysis and Selection of Variables in Linear Regression,” Biometrics 32(1), 1-49 (1976).

Hurvich, C. M. and C. L. Tsai, “Regression and Time Series Model Selection in Small Samples,” Biometrika 78, 297-307 (1989).

Ismail, A. S., “Holdup Profile in Multistage Stirred-Cell Liquid-Liquid Extraction Column Using Population Balance Model,” Can. J. Chem. Eng. 82(5), 1037-1043 (2004).

Jaree, A., R. R. Hudgins, H. Budman, P. L. Silveston, M. Menzinger, “Numerical Investigation of Resonance Behaviour of a Tubular Packed-Bed Reactor with Non-uniform Activity,” Can. J. Chem. Eng. 82(2), 387-391 (2004).

Kendi, T. A. and F. J. Doyle, “Nonlinear Control of a Fluidized Bed Reactor using Approximate Feedback Linearization,” Ind. Eng. Chem. Res. 35(3), 746-757 (1996).

Konishi, S. and G. Kitagawa, “Information Criteria and Statistical Modeling,” Springer, NY, pp. 29-254 (2008).

Lanterman, A. D. “Schwarz, Wallace, and Rissanen: Intertwining Themes in Theories of Model Selection,” Int. Stat. Rev. 69(2), 185-212 (2001).

Li B. B., J. Morris and E. B. Marin, “Model Selection for Partial Least Sqaures Regression,” Chemom. Intell. Lab. Syst., 64(1), 79-89 (2002).

Li, R. and D. K. J. Lin, “Data Analysis in Supersaturated Designs,” Stat. Probabil. Lett. 59, 135-144 (2002).

Linhart, H. and W. Zucchini, “Model Selection,” John Wiley and Sons, NY, pp. 1-38, 112-118 (1986).

Lv, P., J. Chang, T. Wang, C. Wu and N. Tsubaki, “A Kinetic Study on Biomass Fast Catalytic Pyrolysis,” Energ. Fuel. 18(6), 1865-1869 (2004).

Mallow, C. L. “Some Comments on Cp,” Technometrics, 15, 661-675 (1973).

Maria, G., “A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems,” Chem. Biochem. Eng. Q. 18(3), 195-222 (2004).

Maria, G., “Application of Lumping Analysis in Modeling the Living Systems - a Trade-off between Simplicity and Model Quality,” Chem. Biochem. Eng. Q. 20(4), 353-373 (2006).

Marquardt, W., “Model-Based Experimental Analysis of Kinetic Phenomena in Multi-Phase Reactive Systems,” Chem. Eng. Res. Des. 83(A6), 561-573 (2005).

Mchaweh, A., A. Alsaygh, K. Nasrifar and M. Moshfeghian, “A Simplified Method for Calculating Saturated Liquid Densities,” Fluid Phase Equilib. 224(2), 157-167 (2004).

McQuarrie, A. D. R., R. H. Shumway and C. L. Tsai, "The Model Selection Criterion AICu," Stat. Probabil. Lett. 34, 285-292 (1997).

McQuarrie, A. D. R. and C. L. Tsai, "Regression and Time Series Model Selection," World Scientific, Singapore, pp. 15-87 (1998).

Neto, F. S. and R. M. Cotta, "Lumped-Differential Analysis of Concurrent Flow Double-Pipe Heat-Exchanger," Can. J. Chem. Eng. 70(3), 592-595 (1992).

Neumaier, A., "Solving Ill-conditioned and Singular Linear Systems: A Tutorial on Regularization," SIAM Rev. 40, 636-666 (1998).

Perregaard, J., "Model Simplification and Reduction for Simulation and Optimization of Chemical Processes," Comput. Chem. Eng. 17(5-6), 465-483 (1993).

Rao, P. "Some Notes on Misspecification in Multiple Regressions," Am. Stat. 25(5), 37-39 (1971).

Rao, C. R. and Y. Wu, "On Model Selection (with Discussion)," in "Model Selection," (Ed. by P. Lahiri), IMS Lecture Notes – Monograph Series 38, 1-64 (2001).

Romdhane, M. and C. Tizaoui, "The Kinetic Modeling of a Steam Distillation Unit for the Extraction of Aniseed (Pimpinelia Anisum) Essential Oil," J. Chem. Technol. Biot. 80(7), 759-766 (2005).

Schaper C. D., W. E. Larimore, D. E. Seborg and D. A. Mellichamp, "Identification of Chemical Processes Using Canonical Variate Analysis," Comput. Chem. Eng., 18(1), 55-69 (1994).

Scherr, F. F., A. K. Sarmah, H. J. Di and K. C. Cameron, "Modeling Degradation and Metabolite Formation Kinetics of Estrone-3-sulfate in Agricultural Soils," Environ. Sci. Technol. 42(22), 8388-8394 (2008).

Schwarz, G. "Estimating the Dimension of a Model," Ann. Stat. 6, 461-464 (1978).

Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," John Wiley and Sons, NY, pp. 23-25, 529-531 (2003).

Shibata, R., "Statistical Aspects of Model Selection," In "From Data to Model," J. C. Willems, Springer-Verlag, NY, pp. 215-240 (1989).

Stoica, P. and Y. Selen, "Model-Order Selection: A Review of Information Criterion Rules," IEEE Signal Proc. Mag. 21(4), 36-47 (2004).

Sugiura, N. "Further Analysis of the Data by Akaike's Information Criterion and the Finite Corrections," Commun. Stat. A-Theor. 7, 13-26 (1978).

Talukdar, J. and P. Basu, "A Simplified Model of Nitric-Oxide Emission from a Circulating Fluidized-Bed Combustor," Can. J. Chem. Eng. 73(5), 635-643 (1995).

Tibshirani, R. J., "Regression Shrinkage and Selection via the Lasso," J. R. Stat. Soc. B. 58, 267-288 (1996).

Tibshirani, R. J., "The LASSO Method for Variable Selection in the Cox Model," Stat. Med. 16, 385-395 (1997).

Toher D., G. Downey and T. B. Murphy, "A Comparison of Model-based and Regression Classification Techniques applied to Near Infrared Spectroscopic Data in Food Authentication Studies," Chemom. Intell. Lab. Syst. 89(2), 102-115 (2007).

Wang, F. Y., Z. H. Zhu, P. Massarotto and V. Rudolph, "A Simplified Dynamic Model for Accelerated Methane Residual Recovery from Coals," Chem. Eng. Sci. 62(12), 3268-3275 (2007).

Wang, S. G. and S. C. Chow, "Advanced Linear Models: Theory and Applications," Marcel Dekker, NY, pp. 228-243 (1994).

Wu, S., T. H. Harris and K. B. McAuley, "The Use of Simplified or Misspecified Models: Linear Case," Can. J. Chem. Eng. 85, 386-398 (2007).

Yoshida, H., Y. Takahashi and M. Terashima, "A Simplified Reaction Model for Production of Oil, Amino Acids, and Organic Acids from Fish Meat by Hydrolysis under Sub-critical and Supercritical Conditions," J. Chem. Eng. JPN. 36(4), 441-448 (2003).

Zhang, P. "Comment on 'An Asymptotic Theory for Linear Model Selection'," Stat. Sinica, 7, 254-258 (1997).

Chapter 4

Selection of Simplified Models: II. Development of a Model-Selection Criterion Based on Mean-Squared Error[4]

4.1 Summary

Simplified models (SMs) with a reduced set of parameters are used in many practical situations, especially when the available data for parameter estimation are limited. A variety of candidate models are often considered during the model formulation, simplification and parameter estimation processes. In this Chapter, a new criterion is proposed to help modellers select the best SM, so that predictions with lowest expected mean-squared error can be obtained. The effectiveness of the proposed criterion for selecting simplified nonlinear univariate and multivariate models is demonstrated using Monte Carlo simulations and is compared with the effectiveness of the Bayesian Information Criterion (BIC).

4.2 Introduction

A mathematical model is a representation, in mathematical terms, of certain aspects of a nonmathematical system (Aris, 1999). In science and engineering, mathematical modelling plays an important role, and models are used for simulating, designing, controlling and optimizing industrial production processes. In many modelling situations in chemical engineering, modellers

[4] The work summarized in this Chapter was submitted to *Canadian Journal of Chemical Engineering* in September 2009. Drs. Kim McAuley and Thomas Harris were co-authors of this research work. Note that this thesis has been prepared using a manuscript format, so some nomenclature used is not consistent throughout the entire thesis. Please refer to Section 4.8 for the nomenclature used in this Chapter.

have sufficient scientific knowledge to derive complex phenomenological models, which can be expected to match the underlying behaviour of the process very well. Unfortunately, it is often too difficult or costly to obtain enough good data to reliably estimate all of the unknown model parameters (e.g., Perregaard, 1993; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Lv et al., 2004; Maria, 2004, 2006; Mchaweh et al., 2004; Chang et al., 2005; Romdhane and Tizaoui, 2005; Wang et al., 2007). For complex models with many parameters, the resulting parameter estimates and model predictions may exhibit high variability, especially when the data available are limited (e.g., the number of data points is small, measurements are noisy, the range of input-variable settings is small, and/or experimental designs are highly correlated) (Wu et al., 2007). The decisions made using these models (or their parameter estimates) may be unreliable. As a result, it is important to avoid estimating too many model parameters when the information content of the data is low.

Because of the difficulties associated with formulating complex models and with obtaining good estimates for all of the unknown parameters, engineers often use simplified models (SMs) that are known to be structurally imperfect. A variety of candidate SMs can be obtained by making different assumptions during the model formulation and simplification process. For example, temperature dependencies of heat capacities or kinetic rate constants can be ignored, small terms in material or energy balances can be neglected, and unknown parameters can be fixed at reasonable values (obtained from similar systems) to reduce the number of parameters that need to be estimated. There are many reasons to choose a SM with fewer parameters and terms than the correctly-structured or extended model (EM) (Zhang, 1997). The practical advantages of a parsimonious model often overshadow concerns over the correctness of the model structure. When the available data are not informative, SMs can be expected to give better

predictions with lower mean-squared error (MSE) than the EM (Rao, 1971; Hocking 1976; Wu et al., 2007).

When experimental data are insufficient to support the use of complex models, modellers must make decisions about model simplification. They need to know which terms and parameters to include, which parameters to fix at nominal values, and which terms to leave out, so that they can obtain the best possible predictions using the data that they possess along with their scientific and engineering knowledge. Many different strategies have been developed for selecting appropriate SMs.

Model-Selection Criteria (MSC) have been studied and used for model selection since the Akaike Information Criterion (AIC) was proposed in 1973 (e.g., Akaike (1973), Linhart and Zucchini (1986), McQuarrie and Tsai (1998), Rao and Wu (2001), Burnham and Anderson (2002), Konishi and Kitagawa (2008)). When using a MSC, such as the AIC, the Final Prediction Error Criterion (FPE), or Mallow's C_p, the criterion value for different candidate models is calculated directly from the model equations and the residuals that are obtained during parameter estimation. The candidate model with the lowest criterion value is selected as the best model. MSC are simple to use because no numerical optimization is required beyond the parameter estimation step. In Chapter 3 in this thesis, nine commonly-used MSC were compared for their performance and tendencies for selecting SMs when the number of data points is small and experimental designs are correlated. It is shown that the expected MSE provides a convenient theoretical means for analyzing the relative tendencies of these MSC.

In this Chapter, a new MSC is developed to assist the modeller in selection of the SM with the lowest expected MSE for model predictions made at the design points. This new MSC explicitly

accounts for bias due to imperfect model structure and for variance in model parameters and predictions arising from noisy data.

It is well known that removing parameters from a correctly-structured EM will introduce bias, but may decrease variance in model predictions (Rao, 1971; Hocking 1976). Use of the MSE for selecting appropriate SMs has been studied by Linhart and Zucchini (1986) and Wu et al. (2007). Linhart and Zucchini (1986) proposed a hypothesis-test approach to compare two nested models and to select the one with lower mean-squared prediction error. Wu et al. (2007) summarized the many quantitative and qualitative results in the literature concerned with using and selecting SMs. A confidence-interval approach was developed to assess the uncertainty associated with whether a SM or the EM will provide lower-MSE model predictions. It was shown that, when SMs are preferable due to limited data, decisions concerned with whether the EM or SM will give better predictions are very uncertain.

One short-coming of the approaches proposed by Linhart and Zucchini (1986) and Wu et al. (2007) is that they can only be used for comparing two nested models, where the SM is a simplified version of the more complex EM. However, in many practical situations, modellers often consider a set of candidate SMs, which may or may not be nested with each other. In this Chapter, a new model-selection criterion is proposed for selecting the best model (with the lowest expected MSE for predictions) from a group of candidate models that includes the EM and several SMs. This criterion is developed using univariate linear models, and is extended for selecting univariate nonlinear models. The performance of the proposed model-selection criterion is demonstrated using Monte Carlo simulations and is compared with the performance of the Bayesian Information Criterion (BIC). It is also shown that the proposed criterion is effective for selecting multivariate nonlinear models when the noise variance-covariance matrix is known.

Difficulties associated with selecting simplified multivariate models when the noise variance-covariance matrix is unknown are discussed.

4.3 Development of MSE-based Model-Selection Criterion

Consider a correctly-structured EM that can be described by the following univariate linear model

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{4.1}$$

where there are p unknwon parameters in β, p_1 unknown parameters in β_1, and n data points available for parameter estimation.

In this model, the noise ε is additive, and the noise-free response of the process is $Y_{true} = X\beta$. Also, assume that (Beck and Arnold, 1977): 1) input settings X_1 and X_2 are deterministic with full column rank; and 2) the stochastic component ε is independently and identically distributed with zero mean and constant variance σ^2. A particular simplified version of the EM is of the form

$$Y = X_1\beta_1 + e \tag{4.2}$$

where $e = X_2\beta_2 + \varepsilon$ incorporates the stochastic component combined with any model mismatch. The SM is nested within the EM.

The current analysis focuses on selecting the best SM from a set of candidate models to obtain the lowest total MSE for model predictions. The MSE is defined as the expected squared difference between the model prediction, $\hat{Y}$, and the noise-free response of the process, Y_{true} (Rice, 1995).

For a column vector of predictions $\hat{Y}$ obtained using a candidate model, the total MSE[5] is

$$\begin{aligned} MSE(\hat{Y}) &= E\left(\left(\hat{Y} - Y_{true}\right)^T\left(\hat{Y} - Y_{true}\right)\right) \\ &= \left(E(\hat{Y}) - Y_{true}\right)^T\left(E(\hat{Y}) - Y_{true}\right) + tr\left(Cov(\hat{Y})\right) \end{aligned} \tag{4.3}$$

where $E(\cdot)$, $Cov(\cdot)$, and $tr(\cdot)$ denote the expected value, variance-covariance matrix and trace, respectively. The second line in Eqn. (4.3) shows that MSE is equal to the squared bias plus the total variance of the model predictions (Rice, 1995). As a result, MSE, which accounts for both bias and variance, is an appropriate criterion for analyzing simplified or misspecified models.

When unknown parameters in the EM (Eqn. (4.1)) and SM (Eqn. (4.2)) are estimated using ordinary least-squares (OLS), the expected total MSE for predictions is (Beck and Arnold, 1977)

$$\begin{aligned} MSE_E &= \sigma^2 p \\ MSE_S &= \sigma^2 p_1 + \beta_2^T X_2^T (I_n - P_1) X_2 \beta_2 \end{aligned} \tag{4.4}$$

where the subscripts "E" and "S" indicate the use of the EM and the SM, respectively.

Wu et al. (2007) developed a strategy to determine whether the SM or the EM is expected to give predictions with lower MSE (at the design points used for parameter estimation). This strategy relies on a critical ratio R_C, which is defined as

$$R_C = \frac{\beta_2^T X_2^T (I_n - P_1) X_2 \beta_2}{(p - p_1)\sigma^2} \tag{4.5}$$

where $P_1 = X_1(X_1^T X_1)^{-1} X_1^T$. The numerator of R_C is the squared bias introduced by removing parameters associated with X_2 from the model, and the denominator is the variance reduction

[5] Note that, "MSE" has different meanings when used in different situations by different researchers. Throughout this thesis, "MSE" associated with an estimator is defined as the corresponding quadratic loss function, and the reader is referred to pp.127-128 of Rice (1995) for the definition.

(due to fewer parameters being estimated) when a particular SM is used rather than the EM. As a result,

$$R_C < 1 \tag{4.6}$$

is a necessary and sufficient condition for $MSE_S < MSE_E$, which implies that the SM is preferable to the EM for making predictions. This critical ratio has also been used to compare the tendencies of various commonly-used MSC for selecting SMs in Chapter 3.

The true value of R_C depends on unknown β_2 and σ^2. Fitting the EM using OLS provides $\hat{\beta}_2$ and s_E^2, which are unbiased estimators of β_2 and σ^2. Therefore, an estimator of R_C can be obtained as

$$\hat{R}_C = \frac{\hat{\beta}_2^T X_2^T (I_n - P_1) X_2 \hat{\beta}_2}{(p - p_1) s_E^2} = \frac{(SSE_S - SSE_E)/(p - p_1)}{SSE_E/(n - p)} \tag{4.7}$$

where "SSE" denotes the sum of squared residuals. Note that the calculation of $\hat{R}_C$ requires knowledge (or an assumption) about the form of the correctly-structured EM. Given the additional assumption that ε is Normally distributed, $\hat{R}_C$ in Eqn. (4.7) is a likelihood ratio statistic (Wang and Chow, 1994), which follows a noncentral F distribution (Montgomery et al., 2001) with $p - p_1$ and $n - p$ degrees of freedom, and noncentrality parameter

$$\lambda = (p - p_1) R_C \tag{4.8}$$

$\hat{R}_C$ given in Eqn. (4.7) is also the test statistic of a partial F test for testing the hypothesis $H_0: \beta_2 = 0$ (Montgomery et al., 2001).

In situations where σ^2 is known from prior information about the variability of the response variable, an estimator for R_C can be obtained from

$$\hat{R}_C = \frac{\hat{\beta}_2^T X_2^T (I_n - P_1) X_2 \hat{\beta}_2}{(p - p_1) \sigma^2} = \frac{(SSE_S - SSE_E)/(p - p_1)}{\sigma^2} \tag{4.9}$$

where $(p - p_1)\hat{R}_C$ follows a noncentral χ^2 distribution (Montgomery et al., 2001) with $p - p_1$ degrees of freedom and the noncentrality parameter λ from Eqn. (4.8).

From Eqns. (4.4) and (4.5), the reduction in total MSE at the design points, when a particular SM is used, is

$$\Delta MSE = MSE_S - MSE_E = \sigma^2(p - p_1)(R_C - 1) \tag{4.10}$$

The model with the smallest value of ΔMSE will provide the best predictions, on average, at the design points. Due to the $(p - p_1)$ term in Eqn. (4.10), when two SMs contain different numbers of parameters (i.e. different values of p_1), the SM with the lower value of R_C may not correspond to the lower MSE_S. As a result, we propose to use the following corrected critical ratio R_{CC} for comparing several models with different number of parameters

$$R_{CC} = \frac{(MSE_S - MSE_E)/n}{\sigma^2} = \frac{p - p_1}{n}(R_C - 1) \tag{4.11}$$

R_{CC} is the increase in MSE (per data point) arising from the selection of a candidate SM (rather than the EM) normalized by the noise variance. The true value of R_{CC} corresponding to the EM is zero. When the available data are informative enough to support the use of the EM, R_C for all SMs will tend to be larger than 1, and the corresponding R_{CC} will be positive. In situations when the available data are limited, R_{CC} for some SMs will tend to be negative, indicating that these SMs will give better predictions than the EM. The SM with lowest value of R_{CC} will give the best predictions in terms of MSE.

Based on the relationship between R_C and R_{CC} in Eqn. (4.11), estimates for R_{CC} can be obtained using $\hat{R}_C$. Unfortunately, $\hat{R}_C$ given in Eqn. (4.7) or (4.9) is biased and has a large variance (Kubokawa et al., 1993). Improved point estimates for R_C can be obtained using various

estimators for the noncentrality parameter λ (Pandey and Rahman, 1971; Kubokawa et al., 1993). A brief summary is provided in the Appendix in this Chapter.

When σ^2 is unknown, we propose that the following truncated estimator for R_C should be used

$$\hat{R}_{CK} = max\left(\frac{n-p-2}{n-p}\hat{R}_C - 1, \frac{2(n-p-2)}{(p-p_1+2)(n-p)}\hat{R}_C\right) \tag{4.12}$$

where the subscript "K" indicates that this estimator was derived using the improved estimator for λ developed by Kubokawa et al. (1993). Note that $\hat{R}_C$, from Eqn. (4.7), follows a noncentral F distribution. In situations when σ^2 is known, the appropriate truncated estimator is

$$\hat{R}_{CK} = max\left(\hat{R}_C - 1, \frac{2}{p-p_1+2}\hat{R}_C\right) \tag{4.13}$$

where $\hat{R}_C$ is obtained from Eqn. (4.9) and $(p-p_1)\hat{R}_C$ follows a noncentral χ^2 distribution.

The truncated estimators in Eqns. (4.12) and (4.13) have lower MSE than the original estimators in Eqns. (4.7) and (4.9), and are less computationally demanding than a corresponding maximum likelihood estimator based on the method of Pandey and Rahman (1971). As a result, we propose that modellers should select the best model using

$$\hat{R}_{CC} = \frac{p-p_1}{n}\left(\hat{R}_{CK} - 1\right) \tag{4.14}$$

The candidate model (either an SM or the EM) with the lowest value of $\hat{R}_{CC}$ is expected to give the lowest total mean-squared prediction error at the design points. This new model-selection criterion will be extended to the selection of univariate and multivariate nonlinear models, which are of greater interest to chemical engineers than are univariate linear models.

4.4 Extension to Selection of Univariate Nonlinear Models

In the nonlinear case, the EM has the form

$$y = f(X, \theta) + \varepsilon \tag{4.15}$$

where $f(X, \theta)$ is nonlinear in some or all of the parameters θ, X contains all the input variable settings, and ε is independently and identically distributed with zero mean and constant variance σ^2. Unlike the linear case, numerical optimization is required to obtain the parameter estimates $\hat{\theta}$ (Seber and Wild, 2003). When there are too many unknown parameters and the available data are limited (e.g., data may be noisy, or may be obtained from poorly designed experiments), SMs, which contain only a subset of the unknown parameters, are often preferred, so that difficulties associated with poor numerical conditioning can be avoided. These candidate SMs can be formulated, either by making different assumptions in the model formulation process, or by leaving some parameters fixed at their initial guesses. Good initial guesses can be obtained based on the available data (Bates and Watts, 1988), the modeller's engineering knowledge and experience, or from similar studies in the literature.

Using an appropriated formulated or selected SM with a small set of parameters can significantly reduce the complexity and nonlinearity of the model, as well as the variability of model predictions. However, use of the SM will introduce bias in model predictions. For the nonlinear model described in Eqn. (4.15), the MSE, which accounts for bias and variance, can be defined as (Rice, 1995)

$$\begin{aligned} MSE(\hat{y}) &= E\left(\left(\hat{y} - f(X, \theta)\right)^T \left(\hat{y} - f(X, \theta)\right)\right) \\ &= \left(E(\hat{y}) - f(X, \theta)\right)^T \left(E(\hat{y}) - f(X, \theta)\right) + tr\left(Cov(\hat{y})\right) \end{aligned} \tag{4.16}$$

where $\hat{y} = f\left(X, \hat{\theta}\right)$. For nonlinear models, R_{CC} is defined as

$$R_{CC} = \frac{(MSE_S - MSE_E)/n}{\sigma^2} \tag{4.17}$$

This expression can be compared to Eqn. (4.11) for linear models. For a given set of candidate models, the one with lowest value of R_{CC} corresponds to the lowest mean-squared prediction error, and therefore, should be selected as the best model. For nonlinear univariate models, R_{CC} can also be written as

$$R_{CC} = \frac{p - p_1}{n}(R_C - 1) \tag{4.18}$$

where R_C, based on linearization of the nonlinear model, is approximately the squared bias introduced by estimating only a subset of parameters, divided by the associated variance reduction. Unfortunately, for nonlinear models, no explicit expression can be written for R_C. When σ^2 is unknown, R_C can be estimated from the data using the likelihood ratio statistic

$$\hat{R}_C = \frac{(SSE_S - SSE_E)/(p - p_1)}{SSE_E/(n - p)} \tag{4.19}$$

which is the same as Eqn. (4.7) for univariate linear models. When σ^2 is known, R_C can be estimated using the right-hand side of Eqn. (4.9). As a result, the associated truncated estimator $\hat{R}_{CK}$, which is given in Eqn. (4.12) or (4.13), can also be used in nonlinear case, so that R_{CC} can be estimated as

$$\hat{R}_{CC} = \frac{p - p_1}{n}\left(\hat{R}_{CK} - 1\right) \tag{4.20}$$

The candidate nonlinear model with the lowest value of $\hat{R}_{CC}$ is expected to give predictions with the lowest total MSE.

The proposed $\hat{R}_{CC}$ criterion for selection of nonlinear univariate models relies on the assumption that $\hat{R}_C$ from Eqn. (4.19) follows a noncentral F distribution. Gallant (1987) showed that likelihood ratio statistics for nonlinear models, like the one on the right-hand side of Eqn.

(4.19), can be adequately described by a noncentral F distributions with noncentrality parameter $\lambda = (p - p_1)R_C$. Calculation of $\hat{R}_C$ using Eqn. (4.19) requires SSE_E, the sum of squared residuals from the EM. In situations where it is impossible to estimate all of the unknown parameters in the EM due to problems of ill conditioning, the value of SSE_E can be approximated using a SM with a sufficiently large number of parameters, so that estimation of additional parameters does not produce a noticeable improvement in the objective function for parameter estimation.

In the next section, Monte Carlo simulations are performed using the Lubricant model of Witt (1974) described by Bates and Watts (1988) to demonstrate: 1) the validity of the approximate noncentral F distribution for $\hat{R}_C$ in Eqn. (4.19); 2) the effectiveness of the proposed MSE-based criterion for selecting the best nonlinear univariate model; and 3) the effects of various factors (e.g., noise variance, number of data points, initial parameter guesses) on the selection of the best model.

4.4.1 Example: Lubricant Model

The Lubricant model predicts the logarithm of the kinematic viscosity of a lubricant as a function of temperature (°C) and pressure (atm/1000). This relationship is described by the following nonlinear empirical model

$$y = f(X, \theta) + \varepsilon \tag{4.21}$$

with

$$f(X, \theta) = \frac{\theta_1}{\theta_2 + x_1} + \theta_3 x_2 + \theta_4 x_2^2 + \theta_5 x_2^3 + (\theta_6 + \theta_7 x_2^2) x_2 exp\left(\frac{-x_1}{\theta_8 + \theta_9 x_2^2}\right) \tag{4.22}$$

where x_1 is temperature, x_2 is pressure, and $X = (x_1 \quad x_2)$. The additive noise ε is assumed to be independently and identically distributed following a Normal distribution with zero mean and constant variance σ^2 (Linssen, 1975). There are $p = 9$ unknown parameters in the EM (Eqn. (4.22)). The original data set consists of $n = 53$ data points obtained at four temperature settings (0℃, 25℃, 37.8℃ and 98.9℃) and a variety of pressure settings ranging from 1 atm to 7469.35 atm (Bates and Watts, 1988).

In the Monte Carlo simulations used for testing the performance of $\hat{R}_{CC}$, the true noise variance is set as

$$\sigma^2 = 0.002 \tag{4.23}$$

which was estimated based on the original data set. Note that, this true value of σ^2 is only used for generating the data in Monte Carlo simulations. We assume that σ^2 is unknown when selecting the best model.

Table 4.1: True parameter values and initial parameter guesses used in Monte Carlo simulations.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
$\theta_{i,true}$	1000	200	1.36	-0.2	-0.02	0.4	0.1	50	-0.45
θ_i^0	960	210	1.42	-0.3	-0.022	0.35	0.05	48	-0.49
deviation	1.66	1.88	1.71	8.00	1.49	1.53	34.12	2.41	1.80

The true parameter values θ_{true} and the initial parameter guesses θ^0 were set at the values given in Table 4.1. These values of θ_{true} were obtained by rounding off parameter estimates from the EM using the original data. The initial guesses in Table 4.1 were arbitrarily chosen to illustrate the proposed methodology, but in practice, these values should be selected carefully using prior knowledge and engineering judgment. The last row of Table 4.1 shows the difference between θ_{true} and θ^0 as a multiple of the standard error for the parameter estimates. For

example, the initial guess of -0.3 for θ_4^0 is 8.00 times of standard errors away from the true value of -0.2, which is used to generate the simulated data.

Set of Candidate Models

To demonstrate the usefulness of the proposed MSE-based criterion, a set of candidate models is arbitrarily chosen, which is shown in Table 4.2.

Table 4.2: Candidate models.

Candidate Model	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	Number of Parameters p_1
SM_1		√		√			√	√		4
SM_2			√		√	√	√	√		5
SM_3	√	√		√		√	√	√		6
SM_4	√	√	√	√		√	√		√	7
EM	√	√	√	√	√	√	√	√	√	$p = 9$

The check mark "√" indicates that the corresponding parameter is estimated in the candidate model. Parameters without a check mark were held at their initial guesses (given in Table 4.1) and were not estimated in the corresponding SMs. In practical applications, these candidate models should be carefully formulated based on the modeller's engineering knowledge and experience.

Assessing the Validity of the Noncentral *F* Distribution Approximation for $\widehat{R}_C$

The assumption that $\hat{R}_C$ (Eqn. (4.19)) adequately follows a noncentral *F* distribution is a key requirement in the development of the proposed $\hat{R}_{CC}$ criterion for selecting nonlinear models. In this section, the validity of this approximation is tested for the nonlinear SMs in Table 4.2. 10000 Monte Carlo simulations with different additive random noise sequences were used to generate

simulated data based on the model in Eqns. (4.21) and (4.22), using the noise variance in Eqn. (4.23), the true parameter values in Table 4.1 and the input settings from the original data set.

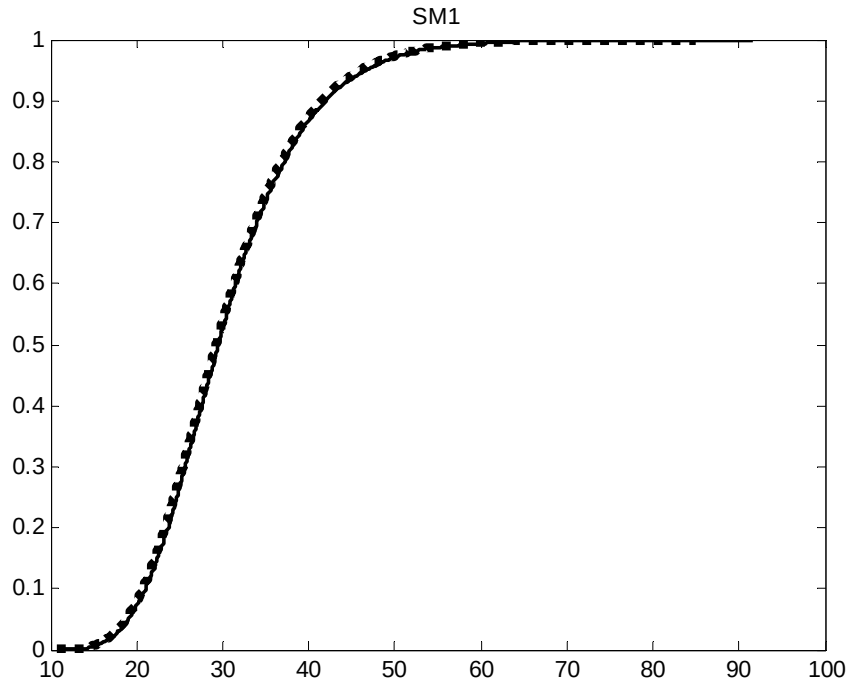

Figure 4.1: Comparison of the theoretical cumulative distribution (----) for $\widehat{R}_C$ and the empirical distribution (—) obtained from 10000 Monte Carlo simulations for SM_1. Note that the two curves are nearly coincident.

For each generated data set, $\hat{R}_C$ was calculated for each candidate model. Empirical and theoretical cumulative distributions of $\hat{R}_C$ for SM_1 are compared in Figure 4.1. The empirical distribution was obtained by sorting the 10000 values for $\hat{R}_C$ from lowest to highest and plotting the fraction of the 10000 values that is below each possible value of $\hat{R}_C$. The theoretical distribution function was calculated using the noncentral *F* distribution function in MATLAB™ with $\lambda = (p - p_1)R_C$. The true value of R_C was calculated using Eqns. (4.17) and (4.18), where

the MSE was approximated using sample biases and sample variances from the complete set of 10000 predictions.

Figure 4.1 shows a close match between the empirical and the theoretical curves. Note that similar results were observed for all of the other SMs in Table 4.2, confirming that the noncentral F distribution is a good approximation for the distribution of $\hat{R}_C$ obtained using these nonlinear models. This result demonstrates that it is appropriate to use the truncated Kubokawa estimator when computing $\hat{R}_{CC}$, which is similar to the more general theoretical conclusion of Gallant (1987).

Model Selection based on Original Data and $\widehat{R}_{CC}$

In this section, $\hat{R}_{CC}$ is used to select the best model based on the original data set, starting from the initial parameter guesses given in Table 4.2. In the first step, each candidate model was fitted using nonlinear least-squares, and the corresponding sum of squared residuals was calculated, and $\hat{R}_{CC}$ values for each model were obtained using Eqns. (4.12), (4.19) and (4.20). Based on the original data set, the EM, with all the nine parameters estimated has the smallest $\hat{R}_{CC}$ value and is therefore selected as the best model. This is not surprising, due to the fact that the modeller used considerable effort and expertise to select the model form (Witt, 1974; Bates and Watts, 1988), which is considered as the EM in this analysis.

In the following section, the performance of the proposed $\hat{R}_{CC}$ model-selection criterion is illustrated using Monte-Carlo simulations involving situations with different noise variances, numbers of data points, and initial parameter guesses.

Performance of the Proposed Model-Selection Criterion

To demonstrate the effectiveness of the proposed $\hat{R}_{CC}$ criterion for selecting the model with the lowest total mean-squared prediction error, four sets of Monte Carlo simulations were conducted. In the first set (Case 1), the noise setting is the same as in Eqn. (4.23) and data obtained at all four temperature settings are used for parameter estimation. In the second set (Case 2), the noise variance was increased by a factor of 5, making it more difficult to obtain good parameter estimates. In the third set (Case 3), the simulation settings were the same as in Case 1, except that simulated data obtained at $T = 98.9℃$ were not available for parameter estimation. In the fourth set (Case 4), the simulation settings were the same as in Case 3, but the initial guess for θ_5 was changed from -0.022 to 0, which is farther away from the true value of -0.02. Since θ_5 is held constant in SM_1, SM_3 and SM_4, by setting $\theta_5 = 0$, the corresponding cubic term ($\theta_5 x_2^3$) in the model is deleted. As a result, these SMs have a simpler model structure than the EM.

In each case, Monte Carlo simulations were performed 10000 times using different random noise sequences. Sample means and sample variances from the 10000 sets of predictions were used to compute theoretical values of MSE and R_{CC} (Eqn. (4.17)), which are shown in Table 4.3 for each candidate model in all of the four cases considered. The smallest MSE and R_{CC} values, which correspond to the best model, are highlighted in bold. The value of R_{CC} for the EM is zero by definition. The results in Table 4.3 indicate that the EM will give the best model predictions, on average, using the settings from Cases 1 and 4, and that SM_4, which has two fewer parameters, is preferred in Cases 2 and 3, when the data are less informative. Simplified models SM_1, SM_2 and SM_3 will give worse predictions, on average, than the EM in all four Cases, as indicated by the larger values of MSE and positive values of R_{CC} in Table 4.3.

Table 4.3: MSE for model predictions and corresponding true values of R_{CC} for each candidate model in all four cases studied.

		SM_1	SM_2	SM_3	SM_4	EM
Case 1	MSE	0.2934	1.1293	0.2147	0.0261	**0.0184**
	R_{CC}	2.5942	10.4799	1.8517	0.0721	**0**
Case 2	MSE	0.3257	1.1690	0.2627	**0.0821**	0.0905
	R_{CC}	0.4439	2.0350	0.3249	**-0.0157**	0
Case 3	MSE	0.0310	0.9557	0.0308	**0.0175**	0.0182
	R_{CC}	0.1684	12.3355	0.1657	**-0.0088**	0
Case 4	MSE	0.1238	0.9557	0.0943	0.0404	**0.0182**
	R_{CC}	1.3892	12.3355	1.0010	0.2928	**0**

Table 4.4 shows the fraction of the time that each model was selected using the 10000 simulated data sets. Results for the models that were selected most often are highlighted in bold. In Case 1, the EM, which is theoretically the best model according to Table 4.3, was selected as the best model 71.14% of the time and SM_4 was selected 28.86% of the time. In Cases 2 and 3, the best model SM_4 was selected most often using $\hat{R}_{CC}$ criterion, and in Case 4, the EM was selected most often due to the poor initial guess for θ_5. These results demonstrate the effectiveness of the proposed model-selection criterion, even in situations when the difference in MSE between the best model and the second-best model is small (as shown in Case 3 in Table 4.3). These simulation results also confirm that when data are noisy (Case 2) or few data points are available (Case 3), a simpler model tends to give better predictions. When the initial guess for a particular parameter is poor, as in Case 4, the proposed selection criterion tends to automatically select a model wherein that parameter is estimated.

For comparison, Table 4.4 also shows the frequencies for each candidate model being selected using the Bayesian Information Criterion (BIC). In Chapter 3, the tendencies of nine different model-selection criteria for selecting SMs were compared and it is determined that BIC did a

reliable job for selecting the best model, with the lowest mean-squared prediction error, for the particular example studied. BIC was computed for each SM from (McQuarrie and Tsai, 1998):

$$BIC = log\left(\frac{SSE_S}{n}\right) + \frac{log(n)}{n} p_1 \tag{4.24}$$

Use of BIC does not require assumptions about the true model structure, which is an advantage over $\hat{R}_{CC}$ in situations where an EM is not available.

Table 4.4: Fraction of each candidate model being selected using $\hat{R}_{CC}$ and BIC.

		Fraction of Each Model Being Selected				
		SM_1	SM_2	SM_3	SM_4	EM
Case 1	$\hat{R}_{CC}$	0	0	0	0.2886	**0.7114**
	BIC	0	0	0	**0.6563**	0.3437
Case 2	$\hat{R}_{CC}$	0.0037	0	0.0055	**0.7088**	0.2820
	BIC	0.0939	0	0.0106	**0.8515**	0.0440
Case 3	$\hat{R}_{CC}$	0.2007	0	0.0479	**0.5274**	0.2240
	BIC	**0.5777**	0	0.0235	0.3788	0.0200
Case 4	$\hat{R}_{CC}$	0	0	0	0.0435	**0.9565**
	BIC	0.0034	0	0.0005	0.2012	**0.7949**

Comparison of the $\hat{R}_{CC}$ and BIC results indicate that, for the particular example studied, the BIC tends to prefer simpler models. The BIC selected SM_4 65.63% of the time in Case 1 and selected the more complex EM preferred by $\hat{R}_{CC}$ only 34.37% of the time. In Case 3, the BIC selected SM_1, with only 4 parameters, most often, whereas $\hat{R}_{CC}$ selected the more complex SM_4 most often. In all four Cases shown in Table 4.4, the $\hat{R}_{CC}$ criterion selected the best model (with the lowest expected total MSE for model predictions) most often, whereas the BIC only selected the best model in Cases 2 and 4.

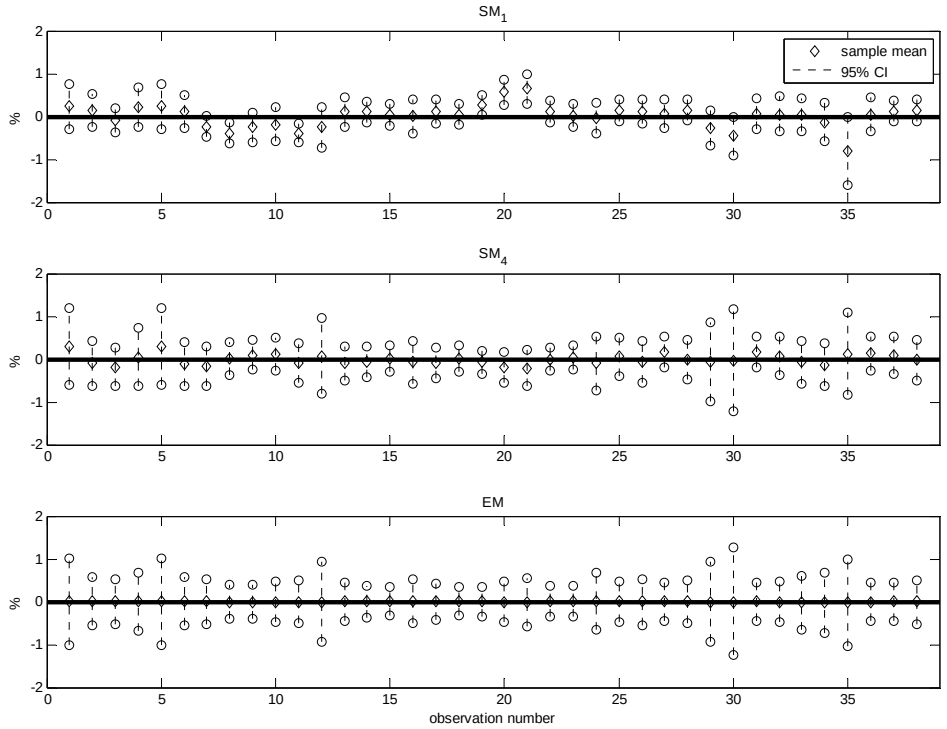

Figure 4.2: Sample mean and 95% empirical confidence intervals (CI) of $\left(f(X,\hat{\theta}) - f(X,\theta)\right)/f(X,\theta)$ for SM_1, SM_4 and the EM at each prediction points (Case 3)

Figure 4.2 shows the sample means and 95% empirical confidence intervals for the percent error in the model predictions obtained at each experimental setting. These results were obtained using the 10000 Monte Carlo simulations for Case 3 and compares the quality of the predictions obtained from SM_1, SM_4 and the EM. Note that SM_1 was selected most often as the best model using the BIC, whereas SM_4 was selected most often using the proposed criterion $\hat{R}_{CC}$. As expected, the EM, which was used to generate the data, gives unbiased model predictions (i.e., the sample mean is approximately zero at all design points). Also, as expected, the predictions obtained from the EM have wider 95% confidence intervals, corresponding to larger prediction

variances. Predictions obtained using SM_4, which is the best model in terms of total MSE for the model predictions has narrower confidence intervals than the EM and has relatively small bias. Predictions from SM_1, which was preferred by BIC, have even smaller variance than predictions from SM_4, but larger bias, on average. Data settings that result in particularly biased predictions from SM_1 are the settings for experiments 20, 21, 30 and 35. The bias introduced by using SM_4 is considerably smaller at all experimental settings, confirming the effectiveness of the proposed $\hat{R}_{CC}$ criterion for this particular example.

4.5 Extension to Selection of Multivariate Models

Many models that appear in the engineering literature have multiple types of response variables (e.g., temperatures, pressures, concentrations, yields). If the model form is complicated, or the data available are not sufficiently informative, estimating all of the unknown parameters may be very difficult or even impossible (e.g., Kou et al., 2005a, b; Ben Zvi et al., 2004). In these complicated nonlinear situations, there are often many competing simplified models that could be used, depending on the simplifying assumptions that are made and the subset of parameters that is estimated (Chu and Hahn, 2008; Lund and Foss, 2008; Thompson et al., 2007, 2009). The difference in complexity and nonlinearity between candidate SMs and the corresponding EM can be substantial. For example, Kou et al. (2005a) developed an EM for ethylene-hexene copolymerization that had 55 parameters, and chose a SM with only 37 parameters because of the limited data available for parameter estimation. Note that Kou et al. (2005a) encountered many difficulties deciding how many parameters to estimate and how many to hold constant at their initial values.

In this section, the proposed MSE-based model-selection criterion is extended to the selection of simplified multivariate nonlinear models.

Assume that a model of the form

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix} = \begin{pmatrix} f_1(X,\theta) \\ f_2(X,\theta) \\ \vdots \\ f_d(X,\theta) \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_d \end{pmatrix} \tag{4.25}$$

can describe the behaviour of the process, where there are d different response variables. Eqn. (4.25) may contain a set of algebraic equations or may be numerical solution of a set of differential and algebraic equations.

Responses obtained using n different sets of experimental conditions can be stacked vertically in a "rolled-out" format (Seber and Wild, 2003), so that n responses for the first variable are at the top, followed by n responses for the second variable, and so on, to give

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \\ \vdots \\ y_{dn} \end{pmatrix} = \begin{pmatrix} f_1(X_1,\theta) \\ f_1(X_2,\theta) \\ \vdots \\ f_1(X_n,\theta) \\ f_2(X_1,\theta) \\ \vdots \\ f_2(X_n,\theta) \\ \vdots \\ f_d(X_n,\theta) \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{dn} \end{pmatrix} \tag{4.26}$$

It is convenient to express these equations in the form

$$y = f(X,\theta) + \varepsilon \tag{4.27}$$

where there are N elements in y. N is the total number of data values available for parameter estimation. If all measurements are available for each set of independent variables, $N = nd$.

Due to the limited information content in the available data, the extension in this analysis focuses on situations where the noise in different response variables is independent, and ε_i $(i = 1,2,\dots,d)$ is independently and identically distributed following a Normal distribution with

zero mean and *known* variance σ_i^2. Prior information about σ_i^2 may have been obtained from repeated experiments on a similar system. Situations when σ_i^2 is unknown are discussed at the end of this section.

Based on the above assumptions,

$$\varepsilon \sim N(0, V) \tag{4.28}$$

with variance-covariance matrix

$$V = \begin{pmatrix} \sigma_1^2 I_n & 0 & \cdots & 0 \\ 0 & \sigma_2^2 I_n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 I_n \end{pmatrix} \tag{4.29}$$

where I_n is an $n \times n$ identity matrix. The use of V with diagonal structure in Eqn. (4.29) is for illustration purposes, and the following analysis is also applicable in situations where the noise in different response variables is correlated with *known* variance-covariance matrix.

To apply the proposed criterion for selecting multivariate simplified models, the model in Eqn. (4.29) is scaled as:

$$\tilde{y} = \tilde{f}(X, \theta) + \tilde{\varepsilon} \tag{4.30}$$

where

$$\tilde{y} = Ly \quad \tilde{f}(X, \theta) = Lf(X, \theta) \quad \tilde{\varepsilon} = L\varepsilon \tag{4.31}$$

The scaling matrix is $L = (U^T)^{-1}$, where U is an upper triangular matrix obtained from Cholesky decomposition of the variance-covariance matrix V (i.e., $U^T U = V$).

In this analysis, the selection of multivariate models focuses on selecting the best model with lowest MSE for the scaled predictions, $\hat{\tilde{y}} = \tilde{f}\left(X, \hat{\theta}\right)$, where $\hat{\theta}$ is obtained by fitting the scaled model in Eqn. (4.30) using nonlinear least-squares. Note that, after putting the multivariate model in "rolled-out" format and scaling using known variances, a univariate nonlinear model is

formulated, where the noise, $\tilde{\varepsilon}$, is independently and identically distributed following a standard Normal distribution. Therefore, the results derived in the previous section for selecting nonlinear univariate models can be used directly for selecting multivariate models. Note that the parameter estimates obtained using ordinary least-squares and the scaled model are identical to those that would be obtained using generalized least-squares based on the original un-scaled multivariate model (Seber and Wild, 2003).

Because the variance of the noise is known (i.e., unit variance after scaling), R_C in Eqn. (4.18) can be estimated using

$$\hat{R}_C = (SSE_S - SSE_E)/(p - p_1) \tag{4.32}$$

where "SSE" is the sum of squared residuals for the scaled model. Note that Eqn. (4.32) is the same as the expression in Eqn. (4.9) for univariate linear models with $\sigma^2 = 1$. Similar to univariate models, the distribution of $(p - p_1)\hat{R}_C$ can be approximated by a noncentral χ^2 distribution, with $p - p_1$ degrees of freedom and noncentrality parameter $\lambda = (p - p_1)R_C$. As a result, the truncated estimator given in Eqn. (4.13) can be used, and R_{CC} in Eqn. (4.18) can be estimated as

$$\hat{R}_{CC} = \frac{p - p_1}{N}\left(\hat{R}_{CK} - 1\right) \tag{4.33}$$

where N is the total number of data values available for parameter estimation. The candidate model with the lowest value of $\hat{R}_{CC}$ is expected to give the lowest total MSE for the scaled predictions.

In the next section, a dynamic α-pinene model (Fuguitt and Hawkins, 1947; Box et al., 1973) is used in Monte Carlo simulations to demonstrate: 1) the quality of the approximate noncentral χ^2 distribution for $(p - p_1)\hat{R}_C$; and 2) the performance of the proposed MSE-based criterion for selecting simplified multivariate nonlinear models.

4.5.1 Example: α-Pinene Model

The α-pinene thermal isomerization process studied by Fuguitt and Hawkins (1947) has three measured responses and the EM consists of five ordinary differential equations (ODEs) (Bates and Watts, 1988)

$$\begin{aligned}
\frac{df_1}{dt} &= -(\theta_1 + \theta_2)f_1 && y_1 = f_1 + \varepsilon_1 \\
\frac{df_2}{dt} &= \theta_1 f_1 \\
\frac{df_3}{dt} &= \theta_2 f_1 - (\theta_3 + \theta_4)f_3 + \theta_5 f_5 && y_3 = f_3 + \varepsilon_3 \\
\frac{df_4}{dt} &= \theta_3 f_3 \\
\frac{df_5}{dt} &= \theta_4 f_3 - \theta_5 f_5 && y_5 = f_5 + \varepsilon_5
\end{aligned} \tag{4.34}$$

where f_i $(i = 1, 2, \ldots, 5)$ correspond to the concentrations of α-pinene, dipentene, alloocimene, pyronene, and dimer, respectively, (in mole %) taken at various times. Only three independent measurements are available (y_1, y_3 and y_5). Although the right-hand sides of the ODEs are linear in the parameters, the predicted responses are nonlinear in the parameters due to exponentials that appear in the analytical solution for the model equations (Box et al., 1973).

The initial values used in the experimental runs are

$$\begin{aligned}
f_1^0 &= 100\% \\
f_2^0 = f_3^0 = f_4^0 = f_5^0 &= 0\%
\end{aligned} \tag{4.35}$$

In the simulated experiments used to test the proposed MSE-based MSC, the times (in minutes) at which simulated measurements were generated match the times used by Fuguitt and Hawkins:

$$t = (1230 \quad 3060 \quad 4920 \quad 7800 \quad 10680 \quad 15030 \quad 22620 \quad 36420)^T \tag{4.36}$$

When generating the simulated experiments, it was assumed that the additive noise in y_1, y_3 and y_5 is Normally distributed with zero mean and variance-covariance matrix

$$V = Cov\left(\begin{pmatrix}\varepsilon_1\\ \varepsilon_3\\ \varepsilon_5\end{pmatrix}\right) = \begin{pmatrix}0.6I_n & 0 & 0\\ 0 & 0.3I_n & 0\\ 0 & 0 & 0.8I_n\end{pmatrix} \tag{4.37}$$

These variances are consistent with the original data. In the analysis below, V is assumed to be known both for model scaling and for model selection.

In this example, there are $d = 3$ response variables and $n = 8$ observations for each response, so there are $N = nd = 24$ data points available for parameter estimation. The true parameter values θ_{true} used to conduct the Monte Carlo simulations were obtained by fitting the model to the original data and rounding the resulting parameter estimates. These true values are listed in Table 4.5, along with initial parameter guesses used for parameter estimation from the 10000 simulated data sets. The last row of Table 4.5 shows the deviation between θ_{true} and θ^0 as a multiple of the standard error for the parameter estimates, which was computed from the 10000 simulations.

Table 4.5: True parameter values and initial parameter guesses used in Monte Carlo simulations.

	θ_1	θ_2	θ_3	θ_4	θ_5
$\theta_{i,true}(\times 10^{-5})$	6	3	2	28	4
$\theta_i^0(\times 10^{-5})$	5.84	2.65	1.63	24.5	5.5
deviation	2.91	7.16	1.87	1.61	2.02

To demonstrate the usefulness of the proposed MSE-based criterion, a set of candidate models was arbitrarily chosen by fixing some parameters and estimating the others, as shown in Table 4.6.

Table 4.6: List of candidate models. Parameters indicated by "√" are included for estimation in the corresponding SM and the remaining parameters are fixed at their initial guesses given in Table 4.5.

Candidate Model	θ_1	θ_2	θ_3	θ_4	θ_5	Number of Parameters p_1
SM_1	√					1
SM_2		√		√		2
SM_3	√		√		√	3
SM_4	√	√	√		√	4
EM	√	√	√	√	√	$p = 5$

Assessing the Validity of the Noncentral χ^2 Distribution Approximation for $(p - p_1)\hat{R}_C$

The validity of the noncentral χ^2 approximation for $(p - p_1)\hat{R}_C$ determines the effectiveness of the Kubokawa estimator for R_C in Eqn. (4.13). 10000 Monte Carlo simulations with different additive random noise sequences were used to generate simulated data based on the above settings. For each generated data set, $\hat{R}_C$ was calculated for each SM using Eqn. (4.32). The noncentrality parameter $\lambda = (p - p_1)R_C$ for the theoretical noncentral χ^2 distributions was calculated from the set of 10000 simulations using the sample mean and sample variance for the scaled predictions.

Figure 4.3 shows the empirical cumulative distributions of $(p - p_1)\hat{R}_C$ for SM_1 and the theoretical χ^2 distribution with $p - p_1$ degrees of freedom and noncentrality parameter λ. The empirical curve matches the theoretical curve closely, confirming that the noncentral χ^2 distribution is a good approximation for the distribution of $(p - p_1)\hat{R}_C$ obtained using this nonlinear model. Similar results were observed for the other SMs in Table 4.6, as would be expected based on the theoretical results of Gallant (1987).

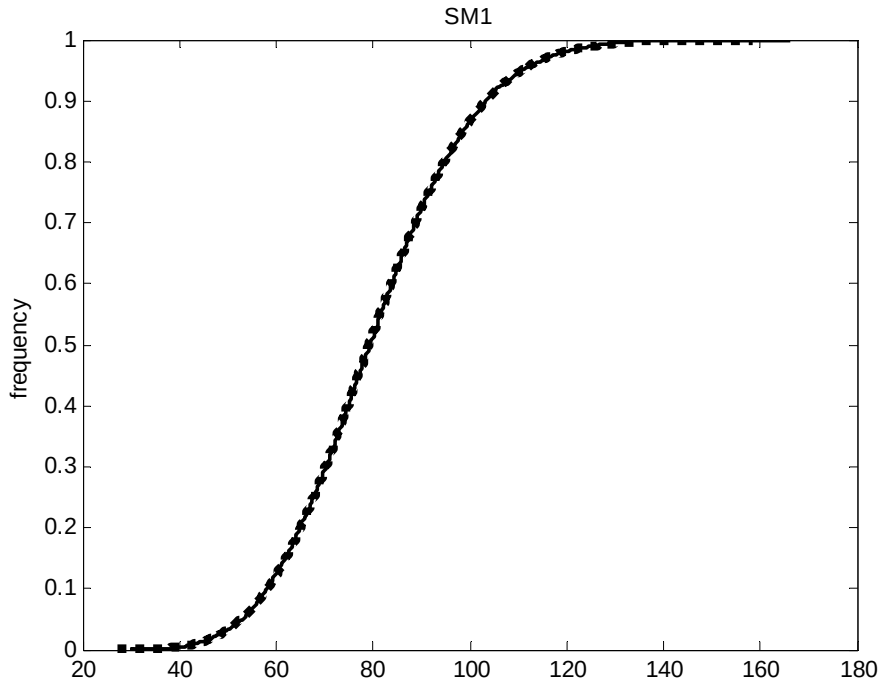

Figure 4.3: Comparison of the theoretical cumulative distribution (----) for $\widehat{R}_C$ and the empirical distribution (—) obtained from 10000 Monte Carlo simulations for SM_1 in Table 4.6.

Performance of the Proposed Model-Selection Criterion

We now examine the performance of the proposed $\hat{R}_{CC}$ criterion for selecting the best model from the set of candidate models in Table 4.6. Table 4.7 shows the theoretical value of the total MSE and R_{CC} for each candidate model, computed using Eqns. (4.16) and (4.17). Note that the EM, which has the smallest value of MSE and R_{CC}, is the best model in terms of mean-squared prediction error. Based on 10000 Monte Carlo simulations, frequencies of each model being selected using $\hat{R}_{CC}$ are also summarized in Table 4.7. The proposed criterion selects the EM (42.35% of the time), more often than it selects the other models, which demonstrates the effectiveness of this criterion for determining the best model.

Table 4.7: True values of MSE and R_{CC} and the frequencies of each model being selected using $\widehat{R}_{CC}$ and BIC.

		SM_1	SM_2	SM_3	SM_4	EM
Theoretical Value	MSE	77.2407	9.4627	10.1075	7.3132	**5.0643**
	R_{CC}	3.0073	0.1833	0.2101	0.0937	**0**
Frequencies	$\widehat{R}_{CC}$	0	0.2473	0.1397	0.1895	**0.4235**
	BIC	0	**0.5425**	0.2160	0.1599	0.0816

Table 4.7 also shows the frequencies of each model being selected using BIC. These results were obtained using Eqn. (4.24) based on the scaled model in Eqn. (4.30). Note that, BIC selected SM_2 most often (54.25% of the time), and selected the best model (the EM) only 8.16% of the time.

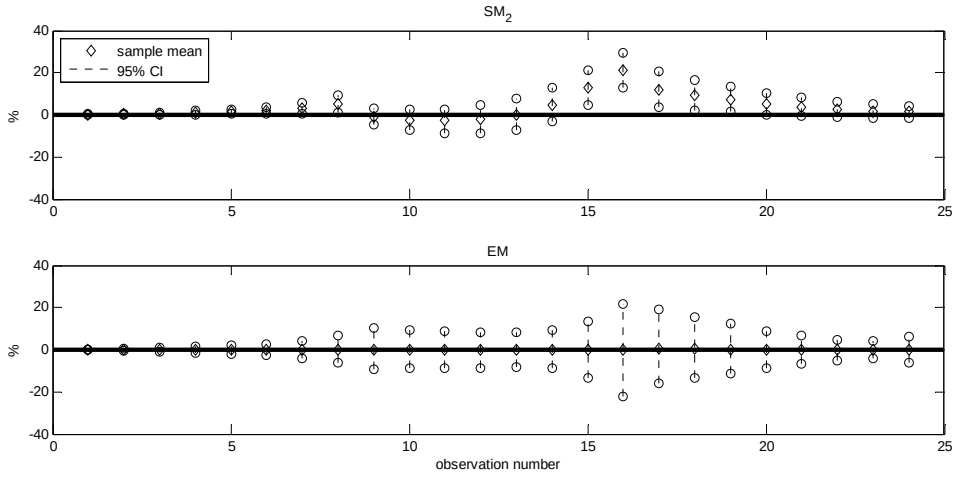

Figure 4.4: Sample mean and 95% empirical confidence intervals (CI) of $\left(\tilde{f}(X,\widehat{\theta}) - \tilde{f}(X,\theta)\right)/\tilde{f}(X,\theta)$ for SM_2 and the EM at each prediction point. Observation number 1 to 8 correspond to α-pinene (f_1) predictions, numbers 9 to 16 correspond to alloocimene (f_3) and numbers 17 to 24 correspond to the dimer (f_5).

Figure 4.4 shows the sample means and 95% empirical confidence intervals for the per cent error in the scaled model predictions obtained at each experimental setting. These results were

generated to compare the quality of the predictions obtained from SM_2 and the EM. Note that SM_2 was selected most often as the best model by the BIC, whereas the EM was preferred using the proposed criterion $\hat{R}_{CC}$. As expected, the EM gives unbiased model predictions. Predictions obtained using SM_2 have narrower confidence intervals than the EM, indicating smaller prediction variances, and have considerable bias for some data settings, especially predictions for data settings from 15 to 19. These results confirm the tendency of the BIC to select simpler models, with higher mean-squared error, than the model selected by the proposed $\hat{R}_{CC}$ criterion.

4.5.2 Selecting Multivariate Models when Noise Variances are Unknown

The analysis in the previous section was based on the assumption that the variance-covariance matrix for the model errors was known *a priori* by the modeller. Although this assumption is valid in situations where there has been considerable prior experimentation on similar systems, there are many situations where the modeller will have limited information about the variances of measured responses. There are both practical and theoretical challenges to extending the proposed $\hat{R}_{CC}$ criterion to selection of models when the variance-covariance matrix is unknown, especially in the situations where the information content in the available data is too weak to support reliable estimation of all unknown model parameters. In these less-than-ideal situations, attempting to estimate the parameters and the noise variances using, e.g., iteratively reweighted least-squares or maximum likelihood methods (Seber and Wild, 2003), is not feasible.

A Bayesian solution for dealing with uncertainties in assumed known variances for measured response variables would be to specify prior distributions for uncertain elements in the variance-covariance matrix, incorporating any knowledge about uncertainties that might be available to the modeller. Random sampling from the prior distributions could be used to scale the model,

estimate the parameters and then calculate many different values of $\hat{R}_{CC}$ for the candidate models under consideration. After a large number of re-sampling, parameter estimation and model selection calculations, the final best model could be determined as the one that was selected most often. This brute-force approach may not be computationally feasible for complex models of chemical processes, where computation times would be prohibitive if the time required to solve the model equations and to estimate the parameters is considerable.

4.6 Conclusions

Parameter estimation in complex models of chemical processes can be difficult, especially when there are many unknown parameters and the available data for parameter estimation are limited (e.g., when noisy data are obtained from poorly designed experiments). In these situations, simplified models with a reduced set of unknown parameters are often preferred to complex models because they can give more reliable predictions. Candidate simplified models can be obtained by making different assumptions during model formulation or by fixing some parameters at nominal values. Modellers want to determine which simplified model will result in the best predictions, given the available data for parameter estimation.

In this Chapter, a reliable and easy-to-use model-selection criterion is developed to assist modellers in the selection of simplified linear or nonlinear models. The new criterion $\hat{R}_{CC}$ is derived using total mean-squared error (MSE) to account for bias and variance in the model predictions. Calculation of $\hat{R}_{CC}$ requires the modeller to have knowledge of (or to make an assumption about) the structure of a full model that is expected to be able to adequately describe the underlying behaviour of the process. The effectiveness of this new criterion is demonstrated

theoretically and using Monte Carlo simulations involving nonlinear single-response and multi-response models.

The performance of the $\hat{R}_{CC}$ criterion is compared with that of the Bayesian Information Criterion (BIC). Both criteria are effective, in that they tend to select simplified models, rather than complex models when data are limited. For the examples studied, the $\hat{R}_{CC}$ criterion consistently selects simplified models that give the lowest total mean-squared prediction errors. In some situations, the BIC tends to select over-simplified models rather than the best model.

The proposed $\hat{R}_{CC}$ criterion can be applied to the selection of multi-response models when variances for the different response variables are assumed to be known. Difficulties associated with multivariate model selection with unknown noise variances are discussed.

4.7 Appendix: Summary of Various Estimators for the Noncentrality Parameter

Improved estimators for R_C can be derived based on various estimators for the noncentrality parameter λ, which appears in noncentral F and χ^2 distributions. Results from Kubokawa et al. (1993) are summarized below.

Given a random variable S, such that $S \sim \chi^2_v(\lambda)$, the unbiased estimators for λ is

$$\hat{\lambda}_0 = S - v \tag{4.38}$$

which can take negative values. The following improved estimator, which results in smaller mean-squared error and no negative values, was proposed by Kubokawa et al.

$$\hat{\lambda}_K = max\left(\hat{\lambda}_0, \frac{2}{v+2}S\right) \tag{4.39}$$

Similarly, in the case when random variable S follows a noncentral F distribution, $S \sim F_{v_1,v_2}(\lambda)$, the uniformly minimum-variance unbiased estimators for λ is

$$\hat{\lambda}_0 = \frac{v_1(v_2 - 2)}{v_2} S - v_1 \tag{4.40}$$

Kubokawa et al. proposed the following truncated estimator, which cannot take negative values and which results in lower mean-squared error

$$\hat{\lambda}_K = max\left(\hat{\lambda}_0, \frac{2v_1(v_2 - 2)}{v_2(v_1 + 2)}\right) \tag{4.41}$$

The truncated estimator for R_C in Eqns. (4.13) and (4.12) was derived from Eqns. (4.39) and (4.41), respectively.

4.8 Nomenclature

d	number of response variable
e	stochastic component with any model mismatch
f	nonlinear model, deterministic response
log	natural logarithm
n	number of data points for a single response variable
p	total number of unknown parameters
s^2	noise variance estimates
t	time
v	degree of freedom
x	input variable
y	vector of response variable
E	expected value
I_n	$(n \times n)$ identity matrix
L	scaling matrix
N	total number of data points
P	projection matrix
R_C	critical ratio
R_{CC}	corrected critical ratio
S	random variable
U	upper triangular matrix
V	variance-covariance matrix
X	matrix of regression variables
Y	response variable

Greek Symbols

β, θ	unknown parameters
ϵ	stochastic component
λ	noncentrality parameter
σ^2	noise variance

Superscripts

$\widehat{\ }$	estimated value
0	initial values
T	matrix transcript
-1	matrix inverse

Subscripts

1	first partitioned part
2	second partitioned part
i	index
$true$	noise-free response or true values
E	extended model
K	Kubokawa estimate
S	simplified model

Abbreviations

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
Cov	Variance-Covariance Matrix
EM	Extended Model
FPE	Final Prediction Error
max	maximum
MSC	Model-Selection Criteria
MSE	Mean-Squared-Error
SM	Simplified Model
tr	trace

4.9 Acknowledgements

The authors would like to thank Cybernetica, DuPont, Hatch, Matrikon, SAS, MITACS (Mathematics of Information Technology and Complex Systems) and NSERC (TJH) for financial support of this research.

4.10 References

Akaike, H. “Information Theory and an Extension of the Maximum Likelihood Principle,” 2. Int. S. Inf. Theor., Ed. PETROV, B. N. and CSAKI F., pp. 267-281. Budapest: Akademia Kiado (1973).

Aris, R. “Mathematical Modeling: A Chemical Engineer’s Perspective,” Academic Press, NY, p.3 (1999).

Bagajewicz, M. J. and E. Cabrera, “Data Reconciliation in Gas Pipeline Systems,” Ind. Eng. Chem. Res. 42(22), 5596-5606 (2003).

Bates, D. M. and D. G. Watts, “Nonlinear Regression Analysis and Its Applications,” John Wiley & Sons, NY, pp. 72-76, 87-89, 147-149, 272-275 (1988).

Beck, J. V. and K. J. Arnold, “Parameter Estimation in Engineering and Science,” John Wiley & Sons, NY, pp. 134-154 (1977).

Ben-Zvi, A., K. McAuley and J. McLellan, “Identifiability Study of a Liquid-Liquid Phase-Transfer Catalyzed Reaction System,” AIChE J. 50(10), 2493-2501 (2004).

Box, G. E. P., W. G. Hunter, J. F. MacGregor and J. Erjavec, “Some Problems Associated with the Analysis of Multiresponse data,” Technometrics 15, 33-5 (1973).

Burnham, K. P. and D. R. Anderson, “Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,” 2nd Edn., Springer, NY, pp. 49-148 (2002).

Chang, S., T. D. Waite and A. G. Fane, “A Simplified Model for Trace Organics Removal by Continuous Flow PAC Adsorption/Submerged Membrane Processes,” J. Membrane Sci. 253(1-2), 81-87 (2005).

Chu Y. and J. Hahn, “Integrating Parameter Selection with Experimental Design Under Uncertainty for Nonlinear Dynamic Systems,” AIChE J. 54(9), 2310-2320 (2008).

Fuguitt, R. E. and J. E. Hawkins, “Rate of Thermal Isomerization of α-pinene in the Liquid Phase,” J. Am. Chem. Soc. 69, 319-322 (1947).

Gallant A. R., “Nonlinear Statistical Models,” John Wiley & Sons, NY, pp. 47-90 (1987).

Hocking, R. R. “Analysis and Selection of Variables in Linear Regression,” Biometrics 32(1), 1-49 (1976).

Konishi, S. and G. Kitagawa, “Information Criteria and Statistical Modeling,” Springer, NY, pp. 29-254 (2008).

Kou B., K. B. McAuley, J. C. C. Hsu and D. W. Bacon, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene/Hexene Copolymerization with Metallocene Catalyst," Macromol. Mater. Eng. 290, 537-557 (2005a).

Kou B., K. B. McAuley, C. C. Hsu, D. W. Bacon and K. Z. Yao, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene Homopolymerization with Supported Metallocene Catalyst," Ind. Eng. Chem. Res. 44, 2428-2442 (2005b).

Kubokawa, T., C. P. Robert and A. K. Md. E. Saleh, "Estimation of Noncentrality Parameters," Can. J. Stat. 21(1), 45-57 (1993).

Linhart, H. and W. Zuccini, "Model Selection," John Wiley & Sons, NY, pp. 1-38, 115-118 (1986).

Linssen, H. N., "Nonlinearity Measures: A Case Study," Stat. Neerl. 29, 93-99 (1975).

Lund B. F. and B. A. Foss, "Parameter Ranking by Orthogonalization – Applied to Nonlinear Mechanistic Models," Automatica 44, 278-281 (2008).

Lv, P., J. Chang, T. Wang, C. Wu and N. Tsubaki, "A Kinetic Study on Biomass Fast Catalytic Pyrolysis," Energ. Fuel. 18(6), 1865-1869 (2004).

Maria, G. "A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems," Chem. Biochem. Eng. Q. 18(3), 195-222 (2004).

Maria, G. "Application of Lumping Analysis in Modeling the Living Systems - a Trade-off between Simplicity and Model Quality," Chem. Biochem. Eng. Q. 20(4), 353-373 (2006).

Mchaweh, A., A. Alsaygh, K. Nasrifar and M. Moshfeghian, "A Simplified Method for Calculating Saturated Liquid Densities," Fluid Phase Equilib. 224(2), 157-167 (2004).

McQuarrie, A. D. R. and C. L. Tsai, "Regression and Time Series Model Selection," World Scientific, Singapore, pp. 15-87 (1998).

Montgomery, D. C., E. A. Peck and G. G. Vining, "Introduction to Linear Regression Analysis," 3rd ed, John Wiley & Sons, NY, pp. 93-94, 582-588 (2001).

Pandey, J. N. and M. Rahman, "The Maximum Likelihood Estimate of the Noncentrality Parameter of a Noncentral F Variate," Siam J. Math. Anal. 2(2), 269-276 (1971).

Perregaard, J., "Model Simplification and Reduction for Simulation and Optimization of Chemical Processes,". Comput. Chem. Eng. 17(5-6), 465-483 (1993).

Rao, C. R. and Y. Wu, "On Model Selection (with Discussion)," in "Model Selection," P. Lahiri, IMS Lecture Notes – Monograph Series 38, 1-64 (2001).

Rao, P. "Some Notes on Misspecification in Multiple Regressions," Am. Stat. 25(5), 37-39 (1971).

Rice, J. A., "Mathematical Statistics and Data Analysis," 2nd Ed., Duxbury Press, Belmont, CA, pp. 127-128 (1995).

Romdhane, M. and C. Tizaoui, "The Kinetic Modeling of a Steam Distillation Unit for the Extraction of Aniseed (Pimpinelia Anisum) Essential Oil," J. Chem. Technol. Biot. 80(7), 759-766 (2005).

Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," John Wiley & Sons, NY, 32-42, 86-89, 529-533, 587-618 (2003).

Thompson D. E., K. B. McAuley and P. J. McLellan, "A Simplified Model for Prediction of Molecular Weight Distributions in Ethylene-Hexene Copolymerization Using Ziegler-Natta Catalysts," Macromol. React. Eng. 1, 523-536 (2007).

Thompson D. E., K. B. McAuley and P. J. McLellan, "Parameter Estimation in a Simplified MWD Model for HDPE Produced by a Ziegler-Natta Catalyst," Macromol. React. Eng. 3, 160-177 (2009).

Vitt, L. Die Berechnung physikalischer und thermodynamischer Kennworte von Druckflussigkeiten, sowie die Bestimmung des Gesamtwirkungsgrades an Rumpen under Berucksichtigung der Thermodynamik fur die Druckflussigkei. Ph.D. Thesis, Technological University Eindhoven. (1974).

Wang, F. Y., Z. H. Zhu, P. Massarotto and V. Rudolph, "A Simplified Dynamic Model for Accelerated Methane Residual Recovery from Coals," Chem. Eng. Sci. 62(12), 3268-3275 (2007).

Wang, S. G. and S. C. Chow, "Advanced Linear Models: Theory and Applications," Marcel Dekker, NY, pp. 228-243 (1994).

Wu, S., T. J. Harris and K. B. McAuley, "The Use of Simplified or Misspecified Models: Linear Case," Can. J. Chem. Eng. 85, 386-398 (2007).

Yoshida, H., Y. Takahashi and M. Terashima, "A Simplified Reaction Model for Production of Oil, Amino Acids, and Organic Acids from Fish Meat by Hydrolysis under Sub-Critical and Supercritical Conditions," J. Chem. Eng. JPN. 36(4), 441-448 (2003).

Zhang, P. "Comment on 'An Asymptotic Theory for Linear Model Selection'". Stat. Sinica 7, 254-258 (1997).

Chapter 5

Selection of an Optimal Parameter Set Using Estimability Analysis and Mean-Squared Error[6]

5.1 Summary

Parameter estimation in complex models of chemical processes is difficult, especially when there are too many unknown parameters to estimate, and the available data for parameter estimation are limited. Estimability analysis ranks parameters from most estimable to least estimable based on the model structure, the available data, and uncertainties in initial parameter guesses. Difficulties associated with poor numerical conditioning are avoided by estimating the most estimable parameters. The remaining parameters are left at their initial values or can be removed from the model via simplification. In this Chapter, the mean-squared error based criterion developed in Chapter 4 is used to determine the optimal number of parameters to estimate from a ranked parameter list, so that the most reliable model predictions can be obtained.

5.2 Introduction

In many modelling situations described in the engineering and scientific literature, modellers have sufficient knowledge to formulate complex fundamental models that can be expected to adequately represent the underlying behaviour of the process studied. The appropriate use of

[6] The research summarized in this Chapter is in preparation for a journal article and will be submitted in the future. Drs. Kim McAuley and Thomas Harris are co-authors of this research work. Note that this thesis has been prepared using a manuscript format, so some nomenclature used is not consistent throughout the entire thesis. Please refer to Section 5.7 for the nomenclature used in this Chapter.

these models for simulating, designing, controlling and optimizing industrial production processes relies on reliable estimation of the many unknown parameters contained in the models. Unfortunately, the information content in the data available for parameter estimation are often limited (e.g., the number of data points is small, measurements are noisy, the range of input-variable settings is narrow, and/or experimental designs are highly correlated). Consequently, the resulting parameter estimates and model predictions exhibit high variability (Perregaard, 1993; Bagajewicz and Cabrera, 2003; Yoshida et al., 2003; Lv et al., 2004; Gadkar et al., 2005; Maria, 2004, 2006; Mchaweh et al., 2004; Chang et al., 2005; Romdhane and Tizaoui, 2005; Wang et al., 2007). Some parameters may have little effects on the model predictions, and the effects of some parameters on model predictions can be highly correlated with the effects of others. In some situations, it is even impossible to estimate all of the unknown parameters (Grewal and Glover, 1976; Eisenfeld, 1986; Vajda et al., 1989; Ljung and Glad, 1994; Dotsch and Van Den Hof, 1996; Walter and Pronzato, 1996; Ben-Zvi et al., 2003, 2004a, 2004b, 2006; Raue et al., 2009). As a result, modellers often use simplified models (SMs), which are known to be structurally imperfect and contain only a reduced set of parameters (Hiskens, 2001; Kou et al., 2005a, b; Anh et al., 2006; Bastogne et al., 2007; Sin and Vanrolleghem, 2007; Thompson et al., 2009). In these less-than-ideal situations, modellers must make decisions about which set of parameters to estimate, and which to hold constant, so that they can obtain the best possible predictions, using the available data and their scientific and engineering knowledge.

In the literature, several different strategies have been developed to select the appropriate set of parameters for estimation. Velez-Reyes and Verghese (1995) proposed a method for parameter subset selection that uses singular value decomposition of the parametric sensitivity matrix. They illustrated their technique using a simple six-parameter model. In their algorithm,

the subset of parameters to be estimated depends on an arbitrary user-supplied threshold value, which is used to decide which singular values are large and which are small. Weijers and Vanrolleghem (1997) proposed an improved two-step approach for the selection of the best parameter subset to estimate. First, they used a sensitivity analysis to screen for parameters that have little influence on model predictions, so that they can be removed from the list of candidate parameters for estimation. In this screening step, an arbitrary cut-off value of 0.2 for the mean scaled sensitivity with respect to each parameter was used to determine which parameters should be removed. In the second step, they computed Fisher information matrices for all possible subsets of the remaining influential parameters, and selected subsets with specified numbers of parameters that gave the largest value of a D-optimality criterion value. Finally, they used subjective arguments to decide how many parameters to estimate in order to avoid numerical problems. Weijers and Vanrolleghem pointed out that this method is not suitable for complex models due to the "combinatorial explosion" that occurs when the number of parameters becomes large. Note that, similar problems occur using the method of Velez-Reyes and Verghese (1995).

Sandink et al. (2001) used an extension of the relative gain array, the relative disturbance gain, and the disturbance condition number, in the selection of parameters for online updating in models used for process monitoring and control. The application of these techniques was demonstrated using a simple model of a gas-phase polyethylene reactor system, without any parameter ranking. Brun et al. (2002) used a two-step approach, similar to that of Weijers and Vanrolleghem to select parameters to estimate in an activated sludge model with more than 30 parameters. In an initial screening step, they removed parameters with small influence on model predictions. Next, they used a combinatorial approach to compute the value of a D-optimality criterion and a collinearity index for all possible subsets of the remaining parameters. They

selected the final number of parameters to estimate using an arbitrary cut-off value for the collinearity index.

Yao et al. (2003) proposed the orthogonalization algorithm shown in Table 5.1 to rank model parameters from most estimable to least estimable. This deflation algorithm accounts for both the magnitude of parameter influences and for correlation among the effect of various parameters during the ranking procedure. The linear least-squares calculations required are simple, even for complex models with large number of unknown parameters, so that no screening step is needed to remove unimportant parameters. Yao et al. (2003) applied their algorithm to a complex dynamic reactor model for ethylene-butene copolymerization that has more than 50 kinetic parameters. An arbitrary cut-off value was used to determine the number of parameters to estimate. This algorithm is equivalent to an algorithm later proposed by Lund and Foss (2008), which uses QR decomposition rather than matrix inversion.

Degenring et al. (2004) used a parameter selection approach based on Principle Component Analysis to decide which parameters to estimate in a complex metabolism model with 122 parameters. First, a sensitivity-based screening analysis was used to determine that 49 of the 122 parameters have negligible effects on the predicted responses. Next, a Fisher information matrix was computed using the remaining influential parameters and this matrix was decomposed using eigenvalue/eigenvector methods. Parameters that had large weights in eigenvectors that corresponded to small eigenvalues were removed from the set of parameters to be estimated. The authors selected the number of parameters to estimate from their ranked list arbitrarily, after examining the influence of the number of parameters on the objective function for parameter estimation. Li et al. (2004) developed a method to rank parameters for estimation using one metric that accounts for the sensitivity of model predictions to individual parameters and a second

metric that accounts for correlation between the effect of parameters. Computation of the second metric is expensive in problems with large numbers of parameters due to its combinatorial nature. The two metrics are combined in arbitrary fashion to achieve an overall ranking. The authors advocate estimating "as many parameters as possible, subject to the restriction that the optimal solution is not too strongly dependent on the initial parameter values". This approach could lead to estimation of a large number of parameters from limited data, resulting large variance in the model predictions.

Kou et al. (2005a, b) used the algorithm in Table 5.1 to rank parameters in dynamic models for gas-phase ethylene homopolymerization and ethylene/hexane copolymerization. The authors updated the parameter values used to compute the sensitivity coefficients between parameter estimation iterations to account for the influence of poor parameter guesses on the ranking and selection of parameters. The appropriate number of parameters to estimate was determined using subjective arguments related to uncertainties in response variables. Sun and Hahn (2006) developed a complicated screening step to remove parameters that have little influence on model predictions. This screening step, which uses an observability Grammian to account for uncertain initial parameter guesses and their influences on sensitivity coefficients, employs an arbitrary cut-off value to determine which parameters are important. The second step of their algorithm uses singular value decomposition of a sensitivity covariance matrix to identify important directions in the parameter space. The number of important directions to consider is selected arbitrarily, based on the relative sizes of eigenvalues. The authors advocate estimation of linear combinations the parameters, rather than a particular subset of the influential parameters. Chu and Hahn (2007) used a genetic algorithm to find optimal parameter subsets with different numbers of parameters to maximize a D-optimality criterion similar to that used by Brun et al. (2002). The number of

parameters to include in the estimation was then determined using subjective decisions. The authors noted that use of the genetic algorithm is computationally intensive, and convergence is not guaranteed. In a later article, Chu and Hahn (2009) developed a pair-wise clustering method for parameter subset selection, in an effort to reduce the required computation time. Unfortunately, this method is only able to account for pairwise correlations between the effects of parameters, and does not detect linear combinations of three or more parameters that lead to numerical conditioning problems during parameter estimation.

Thompson et al. (2009) used the algorithm in Table 5.1 to rank parameters in a model that predicts molecular weight distributions of ethylene/hexene copolymers produced by a Ziegler-Natta catalyst. When scaling the sensitivity coefficients, they accounted for uncertainties in different types of measurements and uncertainties in initial parameter guesses. Brute-force cross-validation (Stone, 1974) was used to determine the optimal number of parameters to estimate to obtain reliable model predictions. Other successful applications of the algorithm in Table 5.1 can be found in Gadkar et al. (2005), Puskas et al. (2005), Yue et al. (2006), Jayasankar et al. (2009), Koeva et al. (2009), and Littlejohns et al. (2009a,b). All of these researchers struggled to determine the appropriate number of parameters to estimate in their models.

In this Chapter, the effective and easy-to-use criterion developed in Chapter 4 is used for determining the optimal number of parameters to estimate. In section 5.3, the parameter ranking algorithm of Yao et al. (2003) is described in detail. Based on the ranked parameter list, a set of nested simplified models (SMs) is formulated by including more and more parameters to estimate, starting from the top of the list. In section 5.4, the mean-squared error (MSE) based model-selection criterion is used to select the appropriate number of parameters to estimate in nonlinear multivariate dynamic models using limited data. In section 5.5, a DAE model

formulated by the Dow Chemical Company is used to illustrate the effectiveness of the proposed criterion. Comparison with brute-force cross-validation confirms that the proposed criterion provides reliable results with modest computational effort.

5.3 Estimability Analysis

The orthogonalization algorithm proposed by Yao et al. (2003) (see Table 5.1) can be used to rank parameters in the following nonlinear dynamic model

$$\begin{aligned} \frac{dx}{dt} &= f(x, u, \theta) \\ y &= g(x, u, \theta) + \varepsilon \end{aligned} \tag{5.1}$$

which is assumed to be sufficiently complex so that it can adequately describe the behaviour of the process if appropriate parameter values were available.

Assume that: 1) there are m state variables in x, d response variables in y and p unknown parameters in θ; 2) data are available from r dynamic experiments which are conducted using input trajectories u that are perfectly known; 3) n measurements are collected at various times for each response variable in each experimental run; and 4) the random noise ε in the measured responses are Normally distributed.

In Eqn. (5.1), measured responses of different types obtained at different times in different runs are stacked in "rolled-out" format (Seber and Wild, 2003) as

$$\begin{pmatrix} y_{111} \\ \vdots \\ y_{1n1} \\ y_{211} \\ \vdots \\ y_{dn1} \\ y_{112} \\ \vdots \\ y_{dn2} \\ \vdots \\ y_{dnr} \end{pmatrix} = \begin{pmatrix} g_{11}(x,u_1,\theta) \\ \vdots \\ g_{1n}(x,u_1,\theta) \\ g_{21}(x,u_1,\theta) \\ \vdots \\ g_{dn}(x,u_1,\theta) \\ g_{11}(x,u_2,\theta) \\ \vdots \\ g_{dn}(x,u_2,\theta) \\ \vdots \\ g_{dn}(x,u_r,\theta) \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \vdots \\ \varepsilon_{1n1} \\ \varepsilon_{211} \\ \vdots \\ \varepsilon_{dn1} \\ \varepsilon_{112} \\ \vdots \\ \varepsilon_{dn2} \\ \vdots \\ \varepsilon_{dnr} \end{pmatrix} \tag{5.2}$$

where responses for the first variable obtained at different times in the first run are at the top, followed by responses for the second variable in the first run, and so on. Based on the model form in Eqn. (5.2), an unscaled parametric sensitivity matrix, Z, can be formulated as

$$Z = \begin{pmatrix} \left.\frac{\partial g_{11}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial g_{11}}{\partial \theta_p}\right|_{u_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial g_{1n}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial g_{1n}}{\partial \theta_p}\right|_{u_1} \\ \left.\frac{\partial g_{21}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial g_{21}}{\partial \theta_p}\right|_{u_1} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial g_{dn}}{\partial \theta_1}\right|_{u_1} & \cdots & \left.\frac{\partial g_{dn}}{\partial \theta_p}\right|_{u_1} \\ \left.\frac{\partial g_{11}}{\partial \theta_1}\right|_{u_2} & \cdots & \left.\frac{\partial g_{11}}{\partial \theta_p}\right|_{u_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial g_{dn}}{\partial \theta_1}\right|_{u_2} & \cdots & \left.\frac{\partial g_{dn}}{\partial \theta_p}\right|_{u_2} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial g_{dn}}{\partial \theta_1}\right|_{u_r} & \cdots & \left.\frac{\partial g_{dn}}{\partial \theta_p}\right|_{u_r} \end{pmatrix} \tag{5.3}$$

where

$$\left.\frac{\partial g_{il}}{\partial \theta_j}\right|_{u_k}, \quad \begin{pmatrix} i = 1,2,\dots,d; & l = 1,2,\dots,n \\ j = 1,2,\dots p; & k = 1,2,\dots,r \end{pmatrix} \tag{5.4}$$

is the first-order partial derivative of response variable g_i with respect to parameter θ_j, evaluated using the input trajectory u_k at time t_l. Note that these partial derivatives are computed using the initial parameter guesses θ^0. Each column in Z corresponds to one parameter, and each row corresponds to one response variable at a particular sampling time from a particular experimental run. If there are missing values in some response variables, the corresponding row in the Z matrix is deleted, without causing any problems for the ranking algorithm. In most practical situations, due to the high complexity of the model form, the elements of Z must be computed numerically by solving sensitivity equations or using difference approximations (Saltelli et al., 2000).

To effectively compare the effect of different parameters on the various model predictions, appropriate scaling is required. Thompson et al. (2009) scaled the sensitivity coefficients in the Z matrix using knowledge about uncertainties in initial parameter guesses and in the various types of measured responses

$$\left.\frac{s_{\theta_j^0}}{\sigma_i}\frac{\partial g_{il}}{\partial \theta_j}\right|_{u_k} \tag{5.5}$$

$s_{\theta_j^0}$ reflects the modeller's level of uncertainty in the initial guess for the j^{th} parameter, and σ_i is an estimate of the uncertainty in measured responses of type y_i, which may be available from replicate experiments. The scaling in Eqn. (5.5) is recommended for sensitivity analysis by the Intergovernmental Panel for Climate Changes (IPCC, 1999, 2000).

Note that $Z^T Z$ obtained from the properly scaled Z matrix is a Fisher information matrix, and that each row in Z corresponds to a measured value that will be used to estimate the parameters.

One benefit of the estimability algorithm is that data are not required to construct rows in Z, and additional rows can be computed corresponding to proposed experiments. Optimal designs can be obtained by selecting experimental settings that lead to new rows in Z to optimize the values of D-, A- or V-optimality criteria (Thompson et al., 2009).

Table 5.1: Orthogonalization algorithm for parameter ranking (Yao et al., 2003)

1.	Calculate the magnitude (i.e., the Euclidean norm) of each column in Z, using appropriate scaling for the sensitivity coefficients. The most estimable parameter corresponds to the column in Z with the largest magnitude.
2.	Put the corresponding column into a vector X_1.
3.	Use X_1 to predict columns in Z using ordinary least-squares, $\hat{Z} = X_1(X_1^T X_1)^{-1} X_1^T Z$, and calculate the residual matrix $R_1 = Z - \hat{Z}$.
4.	Calculate the magnitude of each column in R_1. The second most estimable parameter corresponds to the column in R_1 with the largest magnitude.
5.	Augment the column vector X_1 by the column in Z corresponding to the second most estimable parameter to produce a two-column matrix X_2.
6.	Advance the iteration counter (subscripts in X and R) and repeat Steps 3 to 5, until all parameters are ranked or when it is impossible to perform the least-square prediction of Z in Step 3 due to matrix singularity.

In this algorithm, the influence of each parameter on the response variable is determined by the magnitude of each column of the scaled Z matrix. Columns with large magnitudes correspond to parameters with large influences on the predicted responses, relative to the corresponding uncertainties. The most estimable parameter, which is ranked first by the algorithm, is the one with the largest overall influence. Step 3 and Step 4 account for correlations among the effect of parameters. Using this algorithm with the scaling in Eqn. (5.5), influential parameters that are poorly known appear near the top of the ranked list; unimportant parameters and parameters whose values are already well-known rank near the bottom. The calculations required are simple, even for complex nonlinear models of industrial processes containing many unknown parameters. The main computational effort is in the calculation of the sensitivity

coefficients, and the main challenge for the user is to specify appropriate initial parameter guesses and scaling factors.

Although ranking of the parameters is relatively easy, it has been difficult for users of this algorithm to decide how many parameters to estimate from the ranked list. Yao et al. (2003) and Kou et al. (2005a, b) used arbitrary cut-off values for the magnitude of columns in the residual matrix R. They estimated the parameters that were ranked before this cut-off value, and left the remaining parameters at their initial guesses or removed them from the model via simplification. Littlejohns et al. (2009a, b) and Koeva et al. (2009) ranked as many parameters as possible, and used the objective function for parameter estimation to decide how many parameters to estimate. They used their judgement to decide when the improvement in the fit became negligible as additional parameters were included in the estimation. Thompson et al. (2009) used computationally expensive brute-force cross-validation to determine the optimal number of parameters to estimate, so that the best possible predictions could be obtained.

When a subset of the model parameters is estimated and other parameters are fixed at nominal (and probably incorrect) values, the resulting parameter estimates and model predictions are biased. Estimating more parameters from the ranked list reduces bias in parameter estimates and model predictions, but increases the variance (Rao, 1971; Hocking, 1976; Wu et al., 2007). MSE, which is the sum of squared bias and total variance, is a convenient measure for the quality of model predictions. Chapter 2 summarized the extensive literature concerned with using and selecting SMs, and Chapter 4 developed a model-selection criterion to select models with lowest MSE for model predictions from a set of candidate models. In this Chapter, this MSE-based criterion is used to determine the optimal number of parameters to estimate from the ranked list,

so that the best possible model predictions (i.e., with the lowest expected MSE) can be obtained from the limited data using fast and simple calculations.

5.4 MSE-based Model-Selection Criterion

The model-selection criterion proposed in Chapter 4 used a critical ratio $\hat{R}_{CC}$ to compare the quality of predictions from different models in sense of mean-squared prediction error. This ratio, which was developed using linear statistical analysis, has been shown to be effective for selecting the best nonlinear multivariate model from a set of candidate models. When comparing several models, the model corresponding to the lowest value of $\hat{R}_{CC}$ is expected to give the best predictions made at the experimental settings used for parameter estimation.

In the context of estimability analysis, the ranked parameter list is used to formulate a set of nested SMs. Starting from the top of the list, each successive SM is slightly more complex in that it has one additional unknown parameter. Parameters in the j^{th} SM can be estimated by minimizing a weighted sum of squared residuals

$$J_j = \sum_{i=1}^{d} \sum_{k=1}^{r} \frac{\left(y_{ik} - \hat{y}_{ikj}\right)^T \left(y_{ik} - \hat{y}_{ikj}\right)}{\sigma_i^2} \tag{5.6}$$

where j is the number of parameters in the SM, and σ_i is the uncertainty in the i^{th} measured response variable, which was used to scale the sensitivity coefficients in Eqn. (5.5). For the j^{th} SM, the critical ratio $\hat{R}_{CC,j}$ can be calculated from

$$\hat{R}_{CC,j} = \frac{p - j}{N}\left(\hat{R}_{CK,j} - 1\right) \tag{5.7}$$

where p is the total number of parameters in the ranked list, and N is the total number of data points available for parameter estimation. If all responses are measured at every sampling time, then $N = ndr$. $\hat{R}_{CK,j}$ is

$$\hat{R}_{CK,j} = max\left(\hat{R}_{C,j} - 1, \frac{2}{p-j+2}\hat{R}_{C,j}\right) \tag{5.8}$$

where

$$\hat{R}_{C,j} = \left(J_j - J_p\right)/(p-j) \tag{5.9}$$

$\hat{R}_{C,j}$ is a likelihood ratio statistic, and $(p-j)\hat{R}_{C,j}$ follows a noncentral χ^2 distribution. Details concerned with the derivation and mean-square error interpretation of $\hat{R}_{C,j}$, $\hat{R}_{CK,j}$ and $\hat{R}_{CC,j}$ are given in Chapter 4. Computation of $\hat{R}_{CC,j}$ is straightforward. J_j in Eqn. (5.9) is the objective function value (Eqn. (5.6)) when j parameters are estimated from the top of the ranked list, and J_p is the objective function value when all p parameters are estimated. The model with the lowest value of $\hat{R}_{CC,j}$ contains the optimal set of parameters to estimate, and is expected to give predictions with lowest expected MSE.

The use of the above criterion requires that J_j can be obtained for each candidate model. In situations where it is impossible to estimate all p unknown parameters due to ill-conditioning, the value of J_p can be approximated using a SM with a sufficiently large number of parameters, so that estimation of additional parameters does not produce a noticeable improvement in the model fit.

In the next section, the critical ratio $\hat{R}_{CC,j}$ is used to determine the optimal number of parameters to estimate in a dynamic nonlinear model (i.e., the Dow Chemical model described by Biegler et al. (1986)). It is confirmed that the optimal number of parameters obtained using $\hat{R}_{CC,j}$ is the same as that obtained using brute-force cross-validation, for this particular example.

5.5 Example: Dow Chemical Model

The following model, formulated by the Dow Chemical Company, describes reaction kinetics in a batch reactor (Biegler et al., 1986). The model consists of six ordinary differential equations and four algebraic equations:

$$\begin{aligned}
\frac{dx_1}{dt} &= -k_2 x_2 x_8 \\
\frac{dx_2}{dt} &= -k_1 x_2 x_6 + k_{-1} x_{10} - k_2 x_2 x_8 \\
\frac{dx_3}{dt} &= k_2 x_2 x_8 + k_1 x_4 x_6 - 0.5 k_{-1} x_9 \\
\frac{dx_4}{dt} &= -k_1 x_4 x_6 + 0.5 k_{-1} x_9 \\
\frac{dx_5}{dt} &= k_1 x_2 x_6 - k_{-1} x_{10} \\
\frac{dx_6}{dt} &= -k_1 x_2 x_6 + k_{-1} x_{10} - k_1 x_4 x_6 + 0.5 k_{-1} x_9 \\
x_7 &= -[Q^+] + x_6 + x_8 + x_9 + x_{10} \\
x_8 &= \frac{K_2 x_1}{K_2 + x_7} \\
x_9 &= \frac{K_3 x_3}{K_3 + x_7} \\
x_{10} &= \frac{K_1 x_5}{K_1 + x_7}
\end{aligned} \tag{5.10}$$

where

$$\begin{aligned}
k_1 &= k_{10} exp\left(\frac{-E_1}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \\
k_2 &= k_{20} exp\left(\frac{-E_2}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \\
k_{-1} &= k_{-10} exp\left(\frac{-E_{-1}}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right)
\end{aligned} \tag{5.11}$$

In this model, T is the reactor temperature, $[Q^+]$ is the initial catalyst concentration (0.0131 $gmol \cdot kg^{-1}$), R is the ideal gas constant (1.986 $cal \cdot K^{-1} \cdot mol^{-1}$), and T_0 is the reference temperature (340.15 K).

There are four response variables:

$$\begin{aligned} y_1 &= x_1 + \varepsilon_1 \\ y_2 &= x_2 + \varepsilon_2 \\ y_3 &= x_3 + \varepsilon_3 \\ y_4 &= x_4 + \varepsilon_4 \end{aligned} \tag{5.12}$$

For each response variable, data are reported from three experimental runs conducted using different reactor temperatures and initial reactant concentrations (Biegler et al., 1986). Some measurements are not available at some sampling times. Due to numerical difficulties encountered when solving this DAE model at high operating temperatures, the estimability analysis and parameter estimation in this analysis use data obtained at 313.15 K (but not at 340.15 K and 373. 15 K). Note that other researchers have had difficulties estimating parameters in this stiff DAE model (Biegler et al., 1986; Stortelder; 1997) and that some authors have opted to use a simplified model with only three differential equations and a reduced set of parameters (Tjoa and Biegler, 1991; Guay and McLean, 1995; Sulieman et al., 2009). Also note that, in the original study (Biegler et al., 1986), only three of the response variables (y_1, y_2 and y_3) are measured independently, and that the values of the fourth variable are computed as

$$y_4 = y_1^0 - y_1 - y_3 \tag{5.13}$$

where y_1^0 is the initial concentration of y_1. A total of $N = 103$ data values are available for parameter estimation, including 27 measurements for y_1 (with 11 missing values), 38 measurements for y_2, and 38 measurements for y_3.

It is assumed that the uncertainties associated with the three measured response variables are

$$\sigma = [0.02 \quad 0.02 \quad 0.02]^T \tag{5.14}$$

These values were used by Becerra et al. (2001) in their simulation study. The uncertainty values in Eqn. (5.14) reflect the modeller's prior knowledge about reproducibility of measurements for different responses. Since all three measurements are concentrations, measured using the same device, these uncertainty levels are identical.

The model described in Eqns. (5.10) and (5.11) has 9 unknown parameters in total, as summarized in Table 5.2.

Table 5.2: Unknown parameters in the Dow Chemical model with optimal parameter estimates. The initial guesses and associated uncertainties are used in estimability analysis and in parameter estimation based on the various SMs.

Parameter	Optimal estimates $\hat{\theta}$	Initial guesses θ^0	Uncertainties s_{θ^0}	Relative uncertainties s_{θ^0}/θ^0
$k_{10}\ (kg \cdot gmol^{-1} \cdot hr^{-1})$	2.1560	1	2×10^{-1}	20%
$E_1\ (cal \cdot mol^{-1})$	1.8476×10^4	1×10^4	2×10^3	20%
$k_{20}\ (kg \cdot gmol^{-1} \cdot hr^{-1})$	3.3590	2	3×10^{-1}	15%
$E_2\ (cal \cdot mol^{-1})$	1.9075×10^4	1×10^4	1×10^4	100%
$k_{-10}\ (hr^{-1})$	3.7197×10^3	2×10^3	5×10^2	25%
$E_{-1}\ (cal \cdot mol^{-1})$	2.6046×10^4	1×10^4	4×10^3	40%
$K_1\ (gmol \cdot kg^{-1})$	2.5750×10^{-16}	3×10^{-16}	2×10^{-16}	67%
$K_2\ (gmol \cdot kg^{-1})$	4.8760×10^{-14}	5×10^{-14}	5×10^{-15}	10%
$K_3\ (gmol \cdot kg^{-1})$	1.7884×10^{-16}	2×10^{-16}	1×10^{-17}	5%

The optimal parameter estimates, which were obtained using data at all three temperature settings, were reported by Biegler et al. (1986). The researchers who obtained these parameter values converted three of the six ODEs to algebraic equations, used logarithm transformations of the state variables and unknown parameters to alleviate ill-conditioning problems, and enforced tight tolerances for the DAE solver to avoid numerical problems that occurred in the integration scheme. "Careful monitoring and appropriate intervention by the investigator over the course of the convergence history" was required to obtain these estimates (Biegler et al., 1986). Note that a

reference temperature T_0 is used to re-parameterize in the Arrhenius equations in this analysis to reduce correlation between estimated pre-exponential factors and activation energies (Bates & Watts, 1988; Stortelder, 1997). For example

$$k_i(T) = \alpha_i exp\left(\frac{-E_i}{RT}\right) = k_{i0} exp\left(\frac{-E_i}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right) \tag{5.15}$$

where $i = 1,2,-1$ so that

$$k_{i0} = \alpha_i exp\left(\frac{-E_i}{RT_0}\right) \tag{5.16}$$

The initial parameter guesses θ^0 and associated uncertainties s_{θ^0} in 3rd and 4th columns of Table 5.2 were set at plausible values to represent the modeller's prior knowledge. Initial parameter values are required for nonlinear least-squares parameter estimation and for parameter ranking using estimability analysis (Yao et al. 2003). The associated uncertainties s_{θ^0}, which reflect how much the modeller believes the initial parameter guesses (or how far the modeller is willing to allow the particular parameter to move away from the nominal value), are also required for the estimability analysis. Relative uncertainties s_{θ^0}/θ^0 are shown in the last column of Table 5.2. Note that large relative uncertainties are assumed for E_2^0 and K_1^0 to indicate that the modeller is not very certain about the validity of these initial parameter guesses.

Based on the above settings and assumptions, estimability analysis was performed and the resulting ranked parameter list is

$$[E_2 \quad K_1 \quad E_{-1} \quad E_1 \quad k_{-10} \quad k_{20} \quad k_{10} \quad K_2 \quad K_3]^T \tag{5.17}$$

Due to large sensitivities of model predictions to parameters E_2 and K_1, and due to large uncertainties in their initial values, E_2 and K_1 are ranked at the top of the list, indicating that they are the two most estimable parameters using the available data. Note that it is possible to estimate the activation energy E_2 using data obtained at only the low temperature setting, because

the modeller has some prior knowledge about k_{20} at the reference temperature of $T_0 = 340.15\ K$, which is indicated by the relatively small value for $s_{k_{20}^0}$. In future work, different scaling factors used in Eqn. (5.5) will be studied for the Dow Chemical model to examine their effect on the final parameter ranking.

Based on the ranked list in Eqn. (5.17), a set of nested SMs can be formulated (i.e., the first SM contains only E_2 as an unknown parameter, and all other parameters are fixed at the initial values in Table 5.2; the second SM contains E_2 and K_1 as unknown parameter, and the eight SM contains all parameters except for K_3). In the following analysis, the best SM (or alternatively, the optimal number for parameters to estimate) is selected using two methods: 1) the proposed MSE-based model selection criterion described in Chapter 4, and 2) the brute-force cross-validation method used by Thompson et al. (2009). Results from both methods are analyzed and compared.

5.5.1 Application of MSE-based Model-Selection Criterion to Dow Chemical Problem

For each candidate SM, the parameters were estimated using the program "NLSCON" (Nowak and Weimann, 2001) by minimizing the weighted sum of squared residuals (Eqn. (5.6)). The Matlab™ DAE solver "ode15s" for stiff DAE models was used to solve the model equation, using parameter values determined by NLSCON. Figure 5.1 shows the minimum objective function values for the candidate SMs with different numbers of parameters. The objective function value at $j = 0$ was evaluated using the initial parameter guesses, with no parameters being estimated. J_j refers to the optimum value of the weighted least-squares objective function when j parameters were included in the SM. As expected, J_j decreases as more and more parameters are estimated. However, when $j \geq 4$, there is little improvement in the model fit.

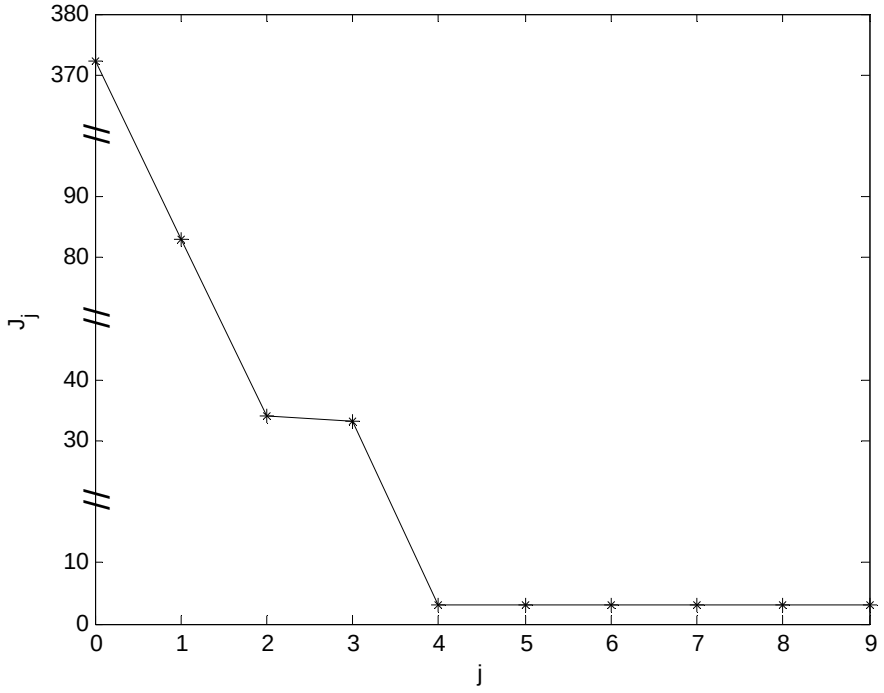

Figure 5.1: J_j values versus the number of parameter estimated from the top of the ranked list

The $\hat{R}_{CC,j}$ values for all of the candidate SMs are plotted in Figure 5.2. When $j \leq 3$, $\hat{R}_{CC,j} > 0$, indicating that the complex model gives better predictions than the corresponding SM. Note that the $\hat{R}_{CC,j}$ value decrease from $j = 1$ to $j = 2$ because the model fit (and predictive ability) improve considerably when 2 parameters are estimated rather than only 1 (see Figure 5.1). The expected MSE of the model predictions decreases because estimation of the additional parameter results in a large reduction in prediction bias compared to the corresponding increase in prediction variance. When the 3 most estimable parameters are estimated, there is only a small improvement in the model fit, therefore $\hat{R}_{CC,j}$ increases, indicating an increase in the expected MSE for model predictions. When $j = 4$, there is a substantial improvement in the model fit and

$\hat{R}_{CC,j}$ reaches its minimum value, indicating that the optimal number of parameters to estimate is 4 (i.e., parameters E_2, K_1, E_{-1} and E_1should be estimated). The small improvement in model fit when $j > 4$ results in an increase in $\hat{R}_{CC,j}$ for larger values of j, corresponding to a larger expected MSE for model predictions when too many parameters are estimated.

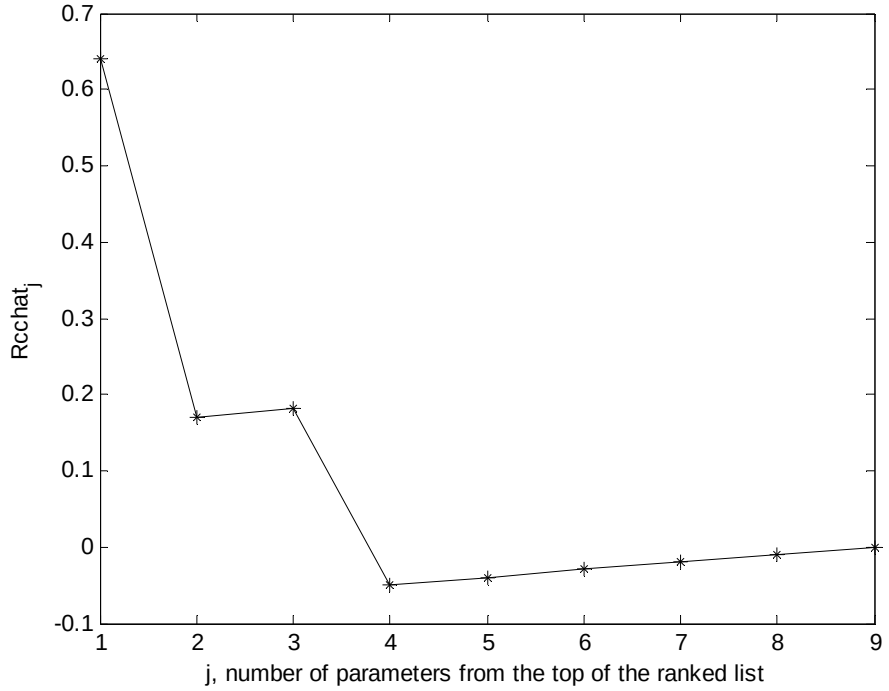

Figure 5.2: $\hat{R}_{CC,j}$ values versus number of parameters estimated from the top of the ranked list

Figure 5.3 shows that model predictions obtained from the SM with E_2, K_1, E_{-1} and E_1 estimated match the data very well. The other 5 parameters were fixed at their initial guesses provided in Table 5.2.

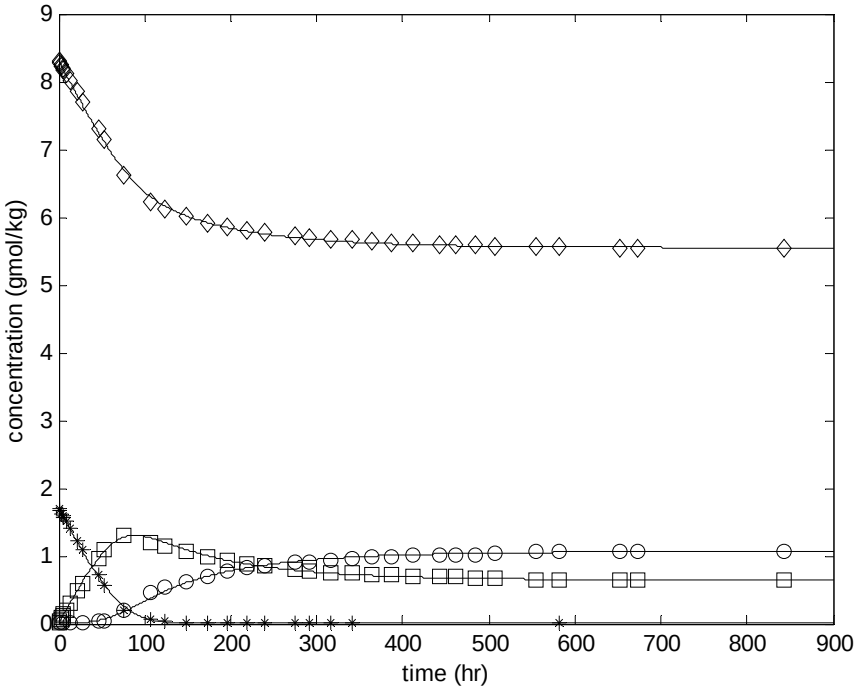

Figure 5.3: model predictions when top four parameters are estimated. ◊: [BM] measurements (y_2); □: [HABM] measurements (y_3); *: [HA] measurements (y_1); ○: [AB] measurements (y_4); —: model predictions

5.5.2 Application of Cross-Validation to Dow Chemical Problem

Cross-validation is a common technique for model selection and validation (Stone, 1974). Using this method, each of the $N = 103$ data points was left out, one at a time, and the remaining $N - 1 = 102$ data values were used for parameter estimation. Each of the resulting 103 sets of estimated parameters was then used to predict the corresponding withheld data value. The resulting weighted squared residuals were then added together to compute the objective function for cross-validation

$$CV_j = \sum_{q=1}^{N} CV_{j,q} \tag{5.18}$$

Subscript j refers to the number of parameters included in the nested SM. This procedure was repeated, for $j = 1$ to $j = 9$. A low value of CV_j indicates good predictive ability of the model. This procedure is known as LOO (leave one out) method (Stone, 1974). It is computationally extensive, especially when the number of data points available for parameter estimation is large.

Due to the heavy computational load, Thompson et al. (2009) only selected 4 key runs from a total of 31 runs for cross-validation when determining the number of parameters to estimate in a polyethylene molecular weight distribution model. They plotted CV_j versus j and concluded that, the predictive ability of their model tended to improve as more parameters were estimated, up to 14 parameters, after which CV_j increased with increasing j. Therefore, the top 14 parameters from their ranked list were selected for parameter estimation using all of the 31 runs.

For the Dow Chemical model studied in this analysis, all data values are used for cross-validation. The resulting CV_j values are plotted in Figure 5.4. As additional parameters were estimated, CV_j decreased until $j = 4$, indicating an improvement in the predictive ability of the model as additional parameters were included. When $j > 4$, estimating additional parameters resulted in worse model predictions than for $j = 4$. Therefore, the cross-validation method indicates that the optimal number of parameters to estimate is 4. This result is the same as that obtained using the proposed MSE-based $\hat{R}_{CC}$ criterion, confirming its effectiveness for this particular example. However, the cross-validation computations required more than 100 times as much computational effort.

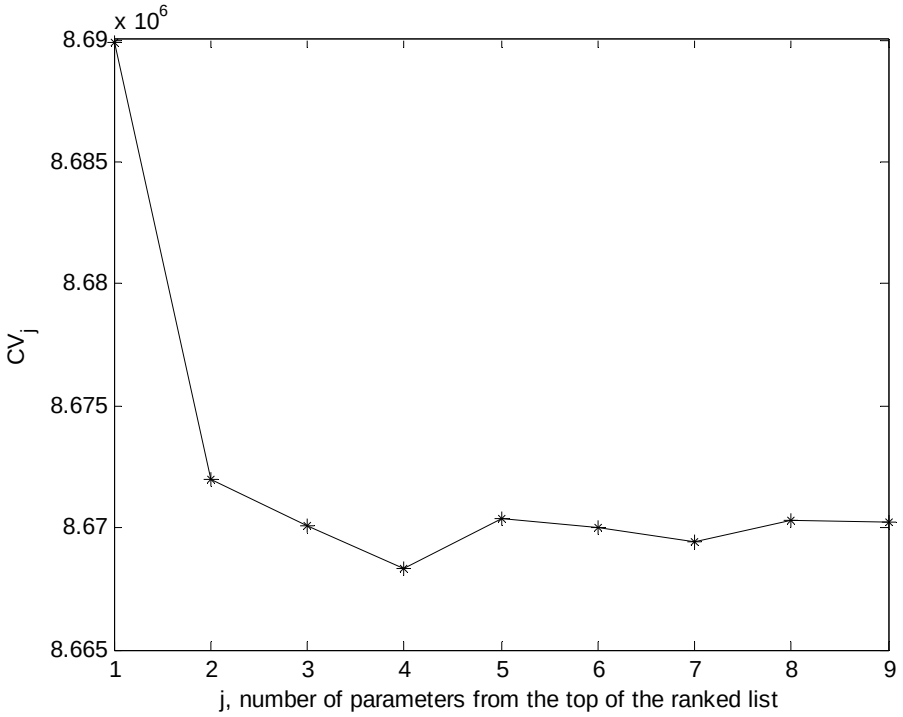

Figure 5.4: CV_j values versus number of parameters estimated from the top of the ranked list

5.6 Conclusions

Parameter estimation in complex models of chemical processes is difficult when there are too many unknown parameters to estimate and the available data for parameter estimation are limited (e.g. number of data points is small, measurements are noisy, the range of input settings is small, and/or experimental designs are highly correlated). Difficulties can be avoided by estimating only a subset of the model parameters, and leaving other parameters at their nominal values. Fortunately, better predictions can often be obtained when some of the key model parameters are estimated, and parameters that have less influence or are already well known remain fixed at their initial values. Modellers need simple and reliable tools to determine which parameters they

should estimate, so that they can obtain the best possible predictions in these less-than-ideal situations. To this end, a mean-squared error (MSE) based model-selection criterion is proposed for selecting the optimal number of parameters to estimate from the ranked parameter lists obtained using estimability analysis. A set of nested simplified models is formulated by including additional parameters from the ranked list. The MSE-based criterion, which is computed using the weighted sum of squared residuals obtained from parameter estimation, has a minimum value when the optimal number of parameters is estimated. Compared to other methods available in the literature, this approach is computationally inexpensive, with no optimization required beyond the parameter estimate step. The effectiveness is demonstrated using a nonlinear multivariate dynamic model formulated by the Dow Chemical Company. Results are consistent with those obtained using cross-validation.

One limitation of the proposed method is that the results depend on initial parameter values and on scaling factors that reflect the modeller's prior knowledge of uncertainties in initial parameter guesses and in measured response variables. Users of the proposed parameter selection methodology may have concerns about the robustness of the resulting optimal parameter set to these required initial assumptions. In future, a re-sampling based approach, similar to that used by Weijers and Vanrolleghem (1997), should be assessed to test the sensitivity of the selected parameters to the modeller's initial assumptions.

5.7 Nomenclature

d	number of response variables
f, g	nonlinear model, deterministic response
k	reaction rate constants
m	number of state variables
n	number of measurements for each response variable
p	number of unknown parameters

r	number of experimental runs
s	uncertainty
t	time
u	input trajectories
x	state variables
y	response variables
E	activation energy
J	objective function in parameter estimation
N	total number of data points for parameter estimation
R	1) residual matrix in parameter ranking algorithm 2) ideal gas constant
R_C	critical ratio
R_{CC}	corrected critical ratio
T	temperature
T_0	reference temperature
X	column selected in Z matrix in the parameter ranking algorithm
Z	sensitivity matrix

Greek Symbols

θ	unknown parameters
ϵ	stochastic component
σ	uncertainty in measured response variable

Superscripts

^{0}	initial values
$^{-1}$	matrix inverse
^{T}	matrix transcript
$\widehat{\ }$	estimates

Subscripts

$_{ilk}$	the l^{th} measurement for the i^{th} response variable in the k^{th} experimental run
$_i$	index for response variables
$_j$	index for parameters
$_k$	index for experimental runs
$_l$	index for measurements
$_K$	Kubokawa estimate

Abbreviations

max	maximum
MSC	Model-Selection Criteria

MSE	Mean-Squared-Error
SM	Simplified Model

5.8 Acknowledgements

The authors would like to thank Cybernetica, DuPont, Hatch, Matrikon, SAS, MITACS (Mathematics of Information Technology and Complex Systems) and NSERC (TJH) for financial support of this research.

5.9 References

Anh, D. T., M. P. Bonnet, G. Vachaud, C. V. Minh, N. Prieur, L. V. Duc and L. L. Anh, "Biochemical Modeling of the Nhue River (Hanoi, Vietnam) Practical Identifiability Analysis and Parameters Estimation," Ecol. Model., 193, 182-204 (2006).

Bagajewicz, M. J. and E. Cabrera, "Data Reconciliation in Gas Pipeline Systems," Ind. Eng. Chem. Res. 42(22), 5596-5606 (2003).

Bastogne, T., M. Thomassin and J. Masse, "Selection and Identification of Physical Parameters from Passive Observations. Application to a Winding Process," Control Eng. Pract., 15, 1051-1061 (2007).

Bates, D. M. and D. G. Watts, "Nonlinear Regression Analysis and Its Applications," John Wiley & Sons, NY, pp. 188-190 (1988).

Becerra, V. M., P. D. Roberts and g. W. Griffiths, "Applying the Extended Kalman Filter to Systems Described by Nonlinear Differential-Algebraic Equations," Control Eng. Pract., 9, 267-281 (2001).

Ben-Zvi, A., P. J. McLellan and K. B. McAuley, "Identifiability of Linear Time-Invariant Differential-Algebraic Systems. I. The Generalized Markov Parameter Approach," Ind. Eng. Chem. Res., 42, 6607-6618 (2003).

Ben-Zvi, A., K. McAuley and J. McLellan, "Identifiability Study of a Liquid-Liquid Phase-Transfer Catalyzed Reaction System," AIChE J., 50(10), 2493-2501 (2004a).

Ben-Zvi, P. J. McLellan and K. B. McAuley, "Identifiability of Linear Time-Invariant Differential-Algebraic Systems. II. The Differential-Algebraic Approach," Ind. Eng. Chem. Eng., 43, 1251-1259 (2004b).

Ben-Zvi, A., P. J. McLellan and K. B. McAuley, "Identifiability of Non-Linear Differential Algebraic Systems via a Linearization Appraoch," Can. J. Chem. Eng., 84, 590-596 (2006).

Bielger, L. T., J. J. Damiano and G. E. Blau, "Nonlinear Parameter Estimation – A Case Study Comparison," AIChE J., 32(1), 29-45 (1986).

Brun, R., M. Kuhni, H. Siegrist, W. Gujer and P. Reichert, "Practical Identifiability of ASM2d Parameters – Systematic Selection and Tuning of Parameter Subsets," Water Res., 36, 4113-4127 (2002).

Chang, S., T. D. Waite and A. G. Fane, "A Simplified Model for Trace Organics Removal by Continuous Flow PAC Adsorption/Submerged Membrane Processes," J. Membrane Sci. 253(1-2), 81-87 (2005).

Chu, Y. and J. Hahn, "Parameter Set Selection for Estimation of Nonlinear Dynamic Systems," AIChE J., 53(11), 2858-2870 (2007).

Chu, Y. and J. Hahn, "Parameter Set Selection via Clustering of Parameters into Pairwise Indistinguishable Groups of Parameters," Ind. Eng. Chem. Res., 48, 6000-6009 (2009).

Degenring, D., C. Froemel, G. Dikta and R. Takors, "Sensitivity Analysis for the Reduction of Complex Metabolism Models," J. Process Contr., 14, 729-745 (2004).

Dotsch, H. G. M. and P. M. J. Van Den Hof, "Test for Local Structural Identifiability of High-order Non-linearly Parametrized State Space Models," Automatica, 32(6), 875-883 (1996).

Eisenfeld, J., "A Simple Solution to the Compartmental Structural-Identifiability Problem," Math. Biosci., 79, 209-220 (1986).

Gadkar, K. G., J. Varner and F. J. Doyle III, "Model Identification of Signal Transduction Networks from Data Using a State Regulator Problem," Systems Biol., 2(1), 17-30 (2005).

Grewal, M. S. and K. Glover, "Identifiability of Linear and Nonlinear Dynamical Systems," IEEE T. Automat. Contr., 21(6), 833-836 (1976).

Guay M. and D. D. McLean, "Optimization and Sensitivity Analysis for Multiresponse Parameter Estimation in Systems of Ordinary Differential Equations," Comput. Chem. Eng. 19(12), 1271-1285 (1995).

Hiskens, I. A., "Nonlinear Dynamic Model Evaluation from Disturbance Measurements," IEEE T. Power Syst., 16(4), 702-710 (2001).

Hocking, R. R. "Analysis and Selection of Variables in Linear Regression," Biometrics 32(1), 1-49 (1976).

IPCC, IPCC expert meetings on good practice guidance and uncertainty management in national greenhouse gas inventories. Background paper. http://www.ipcc-nggip.iges.or.jp/public/gp/gpg-bgp.htm (1999).

IPCC, Good practice guidance and uncertainty management in national greenhouse gas inventories. http://www.ipcc-nggip.iges.or.jp/public/gp/gpgaum.htm (2000).

Jayasankar, B. R., A. Ben-Zvi and B. Huang, "Identifiability and Estimability Study for a Dynamic Solid Oxide Fuel Cell Model," Comput. Chem. Eng., 33, 484-492 (2009).

Koeva, V. I., S. Daneshvar, R. J. Senden, A. H. M. Imam, L. J. Schreiner and K. B. McAuley, "Mathematical modeling of PAG and NIPAM-based Polymer Gel Dosimeters Contaminated by Oxygen and Inhibitor," Macromol. Theor. Simul., accepted (2009).

Kou B., K. B. McAuley, C. C. Hsu, D. W. Bacon and K. Z. Yao, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene Homopolymerization with Supported Metallocene Catalyst," Ind. Eng. Chem. Res. 44, 2428-2442 (2005a).

Kou B., K. B. McAuley, J. C. C. Hsu and D. W. Bacon, "Mathematical Model and Parameter Estimation for Gas-Phase Ethylene/Hexene Copolymerization with Metallocene Catalyst," Macromol. Mater. Eng., 290, 537-557 (2005b).

Li, R., M. A. Henson and M. J. Kurtz, "Selection of Model Parameters for Off-line Parameter Estimation," IEEE T. Contr. Syst. T., 12(3), 402-412 (2004).

Littlejohns, J. V., K. B. McAuley and A. J. Douglas, "Model for a Solid-Liquid Airlift Two-Phase Partitioning Bioscrubber for the Treatment of BTEX," J. Chem. Technol. Biot., accepted (2009a).

Littlejohns, J. V., K. B. McAuley and A. J. Douglas, "Model for a Solid-Liquid Stirred Tank Two-Phase Partitioning Bioscrubber for the Treatment of BTEX," J. Hazard. Mater., accepted (2009b).

Ljung, L. and T. Glad, "On Global Identifiability for Arbitrary Model Parametrizations," Automatica, 30(2), 265-276 (1994).

Lund, B. F. and B. A. Foss, "Parameter Ranking by Orthogonalization – Applied to Nonlinear Mechanistic Models," Automatica, 44(1), 278-281 (2008).

Lv, P., J. Chang, T. Wang, C. Wu and N. Tsubaki, "A Kinetic Study on Biomass Fast Catalytic Pyrolysis," Energ. Fuel. 18(6), 1865-1869 (2004).

Maria, G. “A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems,” Chem. Biochem. Eng. Q. 18(3), 195-222 (2004).

Maria, G. “Application of Lumping Analysis in Modeling the Living Systems - a Trade-off between Simplicity and Model Quality,” Chem. Biochem. Eng. Q. 20(4), 353-373 (2006).

Mchaweh, A., A. Alsaygh, K. Nasrifar and M. Moshfeghian, “A Simplified Method for Calculating Saturated Liquid Densities,” Fluid Phase Equilib. 224(2), 157-167 (2004).

Nowak, U. and L. Weimann, “Numerical Solution of Nonlinear (NL) Least Squares (S) Problems with Nonlinear Constraints (CON),” Matlab™ program, Zuse Institute Berlin, http://www.zib.de/Numerik/numsoft/CodeLib/nonlin.en.html, (2001).

Perregaard, J., “Model Simplification and Reduction for Simulation and Optimization of Chemical Processes,”. Comput. Chem. Eng. 17(5-6), 465-483 (1993).

Puskas, J. E., S. Shaikh, K. Z. Yao, K. B. McAuley and G. Kaszas, “Kinetic Simulation of Living Carbocationic Polymerizations. II. Simulation of Living Isobutylene Polymerization Using a Mechanistic Model,” Eur. Polym. J., 41, 1-14 (2005).

Rao, P. “Some Notes on Misspecification in Multiple Regressions,” Am. Stat. 25(5), 37-39 (1971).

Raue, A., C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmuller and J. Timmer, “Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood,” Systems Biol., 25(15), 1923-1929 (2009).

Romdhane, M. and C. Tizaoui, “The Kinetic Modeling of a Steam Distillation Unit for the Extraction of Aniseed (Pimpinelia Anisum) Essential Oil,” J. Chem. Technol. Biot. 80(7), 759-766 (2005).

Saltelli, A., K. Chan and E. M. Scott, “Sensitivity Analysis,” John Wiley & Sons, NY, pp. 15-50 (2000).

Sandink, C. A., K. B. McAuley and P. J. McLellan, “Selection of Parameters for Updating in On-line Models,” Ind. Eng. Chem. Res., 40, 3936-3950 (2001).

Seber, G. A. F. and C. J. Wild, “Nonlinear Regression,” John Wiley & Sons, NY, pp. 529-531 (2003).

Sin, G. and P. A. Vanrolleghem, “Extensions to Modeling Aerobic Carbon Degradation using Combined Respirometric-Titrimetric Measurements in View of Activated Sludge Model Calibration,” Water Res., 41, 3345-3358 (2007).

Stone, M., "Cross-Validation Choice and Assessment of Statistical Predictions," J. Roy. Stat. Soc. B., 36(2), 111-147 (1974).

Stortelder W., "Parameter Estimation in Nonlinear Dynamical Systems," Ph.D. thesis at Centrum Wiskunde and Informatica, pp. 132-142 (1997).

Sulieman, H., I. Kucuk and P. J. McLellan, "Parametric Sensitivity: A Case Study Comparison," Comput. Stat. Data Anal., 53(7), 2640-2652 (2009).

Sun, C. and J. Hahn, "Parameter Reduction for Stable Dynamical Systems based on Hankel Singular Values and Sensitivity Analysis," Chem. Eng. Sci., 61, 5393-5403 (2006).

Thompson D. E., K. B. McAuley and P. J. McLellan, "Parameter Estimation in a Simplified MWD Model for HDPE Produced by a Ziegler-Natta Catalyst," Macromol. React. Eng. 3, 160-177 (2009).

Tjoa, I. B. and L. T. Biegler, "Simultaneous Solution and Optimization Strategies for Parameter Estimation of Differential-Algebraic Equation Systems," Ind. Eng. Chem. Res., 30, 376-385 (1991).

Vajda, S., H. Rabitz, E. Walter and Y. Lecourtier, "Qualitative and Quantitative Identifiability Analysis of Nonlinear Chemical Kinetic Models," Chem. Eng. Commun., 83, 191-219 (1989).

Velez-Reyes, M. and G. C. Verghese, "Subset Selection in Identification, and Application to Speed and Parameter Estimation for Induction Machines," Proc. 4th IEEE Conf. Control Appl., 991-997 (1995).

Walter, E. and L. Pronzato, "On the Identifiability and Distinguishability of Nonlinear Parametric Models," Math. Comput. Simulat. 42, 125-134 (1996).

Wang, F. Y., Z. H. Zhu, P. Massarotto and V. Rudolph, "A Simplified Dynamic Model for Accelerated Methane Residual Recovery from Coals," Chem. Eng. Sci. 62(12), 3268-3275 (2007).

Weijers, S. R. and P. A. Vanrolleghem, "A Procedure for Selecting Best Identifiable Parameters in Calibrating Activated Sludge Model No. 1 to Full-Scale Plant Data," Water Sci. Technol., 26(5), 69-79 (1997).

Wu, S., T. J. Harris and K. B. McAuley, "The Use of Simplified or Misspecified Models: Linear Case," Can. J. Chem. Eng. 85, 386-398 (2007).

Yao, K. Z., B. M. Shaw, K. B. McAuley and D. W. Bacon, "Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design," Polym. React. Eng., 11(3), 563-588 (2003).

Yoshida, H., Y. Takahashi and M. Terashima, "A Simplified Reaction Model for Production of Oil, Amino Acids, and Organic Acids from Fish Meat by Hydrolysis under Sub-Critical and Supercritical Conditions," J. Chem. Eng. JPN. 36(4), 441-448 (2003).

Yue H., M. Brown, J. Knowles, H. Wang, D. S. Broomhead and D. B. Kell, "Insights into the Behaviour of Systems Biology Models from Dynamic Sensitivity and Identifiability Analysis: A Case Study of an NF-κB Signalling Pathway," Mol. Biosyst., 2, 640-649 (2006).

Chapter 6

Conclusions, Contributions, and Recommendations

6.1 Conclusions and Contributions

This research develops practical and easy-to-use strategies to help modellers select the best simplified model (SM) from a set of candidate models, and to help modellers determine the optimal subset of parameters that should be estimated in complex models. Complex models, which are assumed to be able to adequately describe the behaviour of the underlying process, are referred to as extended models (EMs) in this thesis. The strategies that are developed enable modellers to obtain the best possible model predictions based on the available data and their knowledge of the underlying physics and chemistry. Chapter 2 summarizes results in the statistics literature concerned with using SMs and estimating parameters in SMs. A confidence-interval based approach is developed to assess uncertainties associated with whether a SM or the EM will provide lower mean-squared error (MSE) for model predictions made at the design points. In Chapter 3, MSE is used to compare and analyze nine model-selection criteria (MSC). This analysis shows that the relative probabilities of the various criteria for preferring SMs rather than the correctly-structured EM often appear in a fixed order that is independent of the model form and the data available for parameter estimation. A new MSE-based MSC is developed in Chapter 4, which enables selection of the best model from a group of candidate models including several SMs and the EM. This criterion is demonstrated to be effective for selecting nonlinear multivariate models. In Chapter 5, the proposed MSE-based criterion is used to help modellers determine the optimal subset of parameters to estimate in their complex models. This

methodology takes into account the uncertainties in the initial parameter values and in measured response variables, as well as the experimental settings used to collect the data.

The following specific conclusions and contributions can be drawn from this research work:

1) A critical ratio, R_C, was developed to help modellers determine whether a SM or the EM can give model predictions with lower MSE. When $R_C < 1$, the SM is expected to give better predictions (in the sense of mean-squared prediction error) than the corresponding EM. It was shown that R_C is small (and the SM tends to be preferred) when there are high noise levels, strong correlation in the design matrix, a small range of input-variables settings, or small number of experiments. All of these situations are unfavourable for obtaining precise parameter estimates. From the example studied, it is apparent that confidence intervals on R_C are broad when the data for parameter estimation are limited. As a result, decisions regarding whether the SM or the EM should be used can be very uncertain.

2) The critical ratio R_C and its associated likelihood ratio statistic estimator, $\hat{R}_C$, provide a convenient connection between nine commonly-used MSC and their sampling properties. The following preferential orderings were established: $AIC_U > AIC_C > C_p > AIC$, $BIC > AIC$ $(n > 7)$, $FPE_U > FPE$, and $S_p > C_p > FPE > R^2_{adj}$, where AIC is the Akaike Information Criterion, AIC_C and AIC_U are two corrected versions of AIC, BIC is the Bayesian Information Criterion, FPE is the Final Prediction Error, FPE_U is the corrected version of FPE, S_p is a MSE-based criterion, C_p is Mallow's C_p criterion, and R^2_{adj} is the adjusted coefficient of determination. The greater-than sign ">" in the above orderings indicates a larger propensity of selecting simplified models. Theoretical analysis and Monte Carlo simulations confirmed that MSC with strong tendencies to

guard against overfitting (e.g., AIC_U) have a high tendency to select simplified models when data are limited, and MSC with propensities for overfitting (e.g. R^2_{adj}) tend select SMs with a larger number of parameters. For the particular examples studied, BIC and FPE_U did a good job of selecting SMs with low MSE.

3) A new MSE-based MSC, $\hat{R}_{CC}$, was developed using linear statistical models. This criterion accounts for bias due to imperfect model structure and for variance in model predictions arising from noisy data. Using this criterion, the modeller can select the best model, with lowest MSE for model predictions, from a group of candidate models including several SMs and the EM.
4) Since most models that appear in chemical engineering applications are nonlinear in the parameters and have multiple response variables, the proposed criterion was extended to help modellers in selection of simplified nonlinear multi-response models. Use of the proposed criterion requires that $\hat{R}_C$ follows a noncentral F or χ^2 distribution. The validity of this assumption was verified using two examples via Monte Carlo simulations.
5) A detailed review was conducted of the engineering and scientific literature concerned with parameter subset selection in complex models. It was shown that the proposed MSE-based criterion can be applied to help modellers determine the optimal subset of parameters to estimate in their complex models, without expensive computations or arbitrary cut-off values used in other techniques. The effectiveness of the methodology was demonstrated using a DAE model from Dow Chemical Company. The parameter subset selected using $\hat{R}_{CC}$ was the same as that obtained via computationally-intensive brute- force cross-validation method.

6.2 Recommendations for Future Work

Based on the research results presented in this thesis, the following research is recommended:

1) The confidence-interval based approach developed in Chapter 2 can be extended to study the uncertainties associated with selecting alternative SMs, which are both nested within an EM. The modeller could then assess the probability that the simpler SM, rather than the SM with more parameters, will give better predictions. This analysis could also be applied to parameter subset selection to assess the relative probability of improvement in model predictions by including an additional parameter from the ranked estimability list.
2) In many modelling simulations, model predictions are desired at new input settings Z, rather than at the settings X used for data collection. The best SM for making predictions may depend on the user-selected operating conditions in the Z matrix. By studying the mapping from X to Z, the MSE-based criterion developed in Chapter 4 can be extended to account for the operating range of interest where accurate predictions are desired, which is specified in the Z matrix. This extension will be relatively straightforward.
3) The parameter ranking from the estimability analysis algorithm of Yao et al. (2003) is conditioned on the initial parameter guesses, and on the uncertainty values for parameters and measured response variables. As a result, the parameter ranking list and the optimal set of parameters selected by the MSE-based MSC can change if different initial parameter values and scaling factors are considered. In the future, a re-sampling based approach, similar to the method used by Weijers and Vanrolleghem (1997), should be studied to test the sensitivity of the selected optimal parameter set to the modeller's initial assumptions.

4) The current research focuses on MSE as the measure of model goodness; models with small MSE are preferred to those with larger MSE. In practice, due to the many advantages of using models with simpler structure or smaller number of unknown parameters, a modeller may decide to use a SM even if it gives slightly larger MSE than a more complex model. The MSE-based criterion can be extended to help modellers make decision in these situations, taking into account the modeller's willingness to accept worse predictions in the interest of selecting a simple model.

5) The current analysis focuses on parameter estimation problems in complex models when only limited data are available. Inevitably, this research is related to the literature concerned with design of experiments, where data are collected at well-designed input settings to provide maximum information, so that the best possible parameter estimates and model predictions can be obtained. The use of MSE in this research and associate key results may be useful to help modellers determine how many additional experimental runs should be conducted in a sequential experimental design, so that a desired MSE in model predictions can be obtained at minimum experimental cost.

6.3 References

Weijers, S. R. and P. A. Vanrolleghem, "A Procedure for Selecting Best Identifiable Parameters in Calibrating Activated Sludge Model No. 1 to Full-Scale Plant Data," Water Sci. Technol. 26(5), 69-79 (1997).

Yao, K. Z., B. M. Shaw, K. B. McAuley and D. W. Bacon, "Modeling Ethylene/Butene Copolymerization with Multi-site Catalysts: Parameter Estimability and Experimental Design," Polym. React. Eng. 11(3), 563-588 (2003).